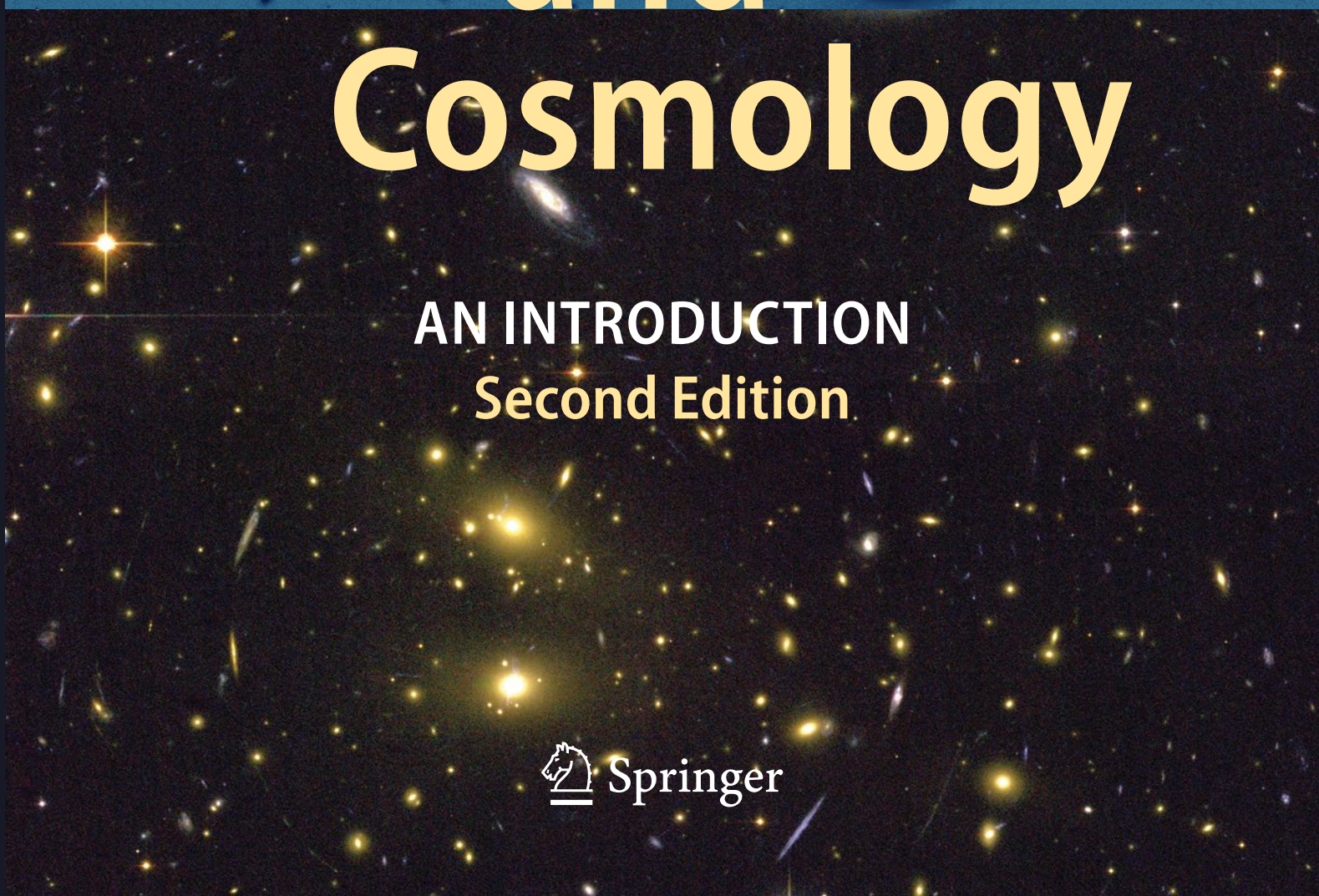
Peter Schneider

Extragalactic Astronomy and Cosmology

AN INTRODUCTION

Second Edition



Springer

Peter Schneider

Extragalactic Astronomy and Cosmology

An Introduction

Second Edition

Springer

Peter Schneider
Argelander-Institut für Astronomie
Universität Bonn
Bonn
Germany

ISBN 978-3-642-54082-0 ISBN 978-3-642-54083-7 (eBook)
DOI 10.1007/978-3-642-54083-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014946357

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

For Mónica

Preface

Amazing times! I finished the manuscript for the first edition of this book just 8 years ago—but the necessity of a new edition was urgently felt. In these years we have witnessed an enormous development in the field of extragalactic astronomy and cosmology. On the instrument side, the final servicing mission to the Hubble Space Telescope brought two new very powerful instruments to this unique observatory, the Herschel and Planck satellites were launched and conducted their very successful missions, the South Pole Telescope and the Atacama Cosmology Telescope started operation, ALMA was inaugurated and began observations, and new powerful high-resolution instruments were installed on 10-m class telescopes. Scientifically, the redshift frontier has been extended, with candidate galaxies at redshifts of ten or higher and stellar explosions seen at redshifts beyond eight, a much improved understanding of the high-redshift galaxy population has been obtained, as a consequence of which also the origin of the cosmic infrared background is now understood, and greatly improved multi-wavelength surveys carried out with the most powerful telescopes, together with new simulation techniques, have provided us with a much better understanding of the evolution of the galaxy population. The Pierre Auger observatory has shed much light on the origin of the most energetic cosmic rays, and the advances of atmospheric Cherenkov telescopes have identified dozens of active galaxies emitting at energies of teraelectron Volts.

Several blind surveys have detected galaxy clusters by their Sunyaev–Zeldovich effect, providing a new and powerful route for cluster cosmology. WMAP has finished its 9 years of surveying the microwave sky, and confirmed two of the predictions of inflation—the spatial flatness of our Universe and the finite tilt of the initial power spectrum. The first cosmological results from Planck were stunning, including an all-sky map of the gravitational potential which is responsible for lensing the cosmic microwave background. The use of baryonic acoustic oscillations as a standard rod to measure the geometry of our Universe has by now been firmly established. Two Nobel prizes in physics, given to cosmologists in 2006 and 2011 for studies of the cosmic microwave background and for the discovery of the accelerated expansion of the Universe using Type Ia supernovae, highlight the impact of this science in the broader physics context.

In this second edition, I have tried to account for these new developments, by updating and (in some cases, substantially) expanding many sections. New material has been added, including a separate chapter on galaxy evolution, as well as sections on the standard model of elementary particles and WIMPs as dark matter candidates, properties of high-redshift galaxies and the galaxy population in clusters, and several other topics. Following the suggestion of several reviewers of the first edition, problems (and solutions) have been added to most chapters. However, I have tried to preserve the style and level of the original book, aiming at a text which combines the physical exploration of cosmic objects with the fascination of astronomical and cosmological research.

I thank Frank Bertoldi, Thomas Reiprich, and Mónica Valencia for carefully reading selected chapters and their numerous helpful suggestions, as well as several colleagues who mailed comments to the first edition. Norbert Wermes provided very useful comments on the particle physics section. I would like to thank Sandra Unruh for her invaluable help in preparing this edition, including numerous comments on draft versions and her efforts to attain the right

to reproduce the many new figures from colleagues all over the world. The collaboration with Ramon Khanna of Springer-Verlag continued to be very constructive.

This book could not have been realized without the many expert colleages from around the world who agreed that their original figures be reproduced here. I thank them sincerely for that and hope that I have represented their original work in a fair way.

I very much appreciate the patience and understanding of my colleagues, in particular my students, for my highly reduced availability and level of activity on other issues during the final months of preparing the manuscript. Finally, I very much thank my wife Mónica for her love, her encouragement, and her support.

Bonn, Germany
January 2014

Peter Schneider

From the first edition

This book began as a series of lecture notes for an introductory astronomy course I have been teaching at the University of Bonn since 2001. This annual lecture course is aimed at students in the first phase of their studies. Most are enrolled in physics degrees and choose astronomy as one of their subjects. This series of lectures forms the second part of the introductory course, and since the majority of students have previously attended the first part, I therefore assume that they have acquired a basic knowledge of astronomical nomenclature and conventions, as well as on the basic properties of stars. Thus, in this part of the course, I concentrate mainly on extragalactic astronomy and cosmology, beginning with a discussion of our Milky Way as a typical (spiral) galaxy. To extend the potential readership of this book to a larger audience, the basics of astronomy and relevant facts about radiation fields and stars are summarized in the appendix.

The goal of the lecture course, and thus also of this book, is to confront physics students with astronomy early in their studies. Since their knowledge of physics is limited in their first year, many aspects of the material covered here need to be explained with simplified arguments. However, it is surprising to what extent modern extragalactic astronomy can be treated with such arguments. All the material in this book is covered in the lecture course, though not all details written up here. I believe that only by covering this wide range of topics can the students be guided to the forefront of our present astrophysical knowledge. Hence, they learn a lot about issues which are currently unsettled and under intense discussion. It is also this aspect which I consider of great importance for the role of astronomy in the framework of a physics program, since in most other subdisciplines of physics the limits of our current knowledge are approached only at a later stage in the education.

In particular, the topic of cosmology is usually met with interest by the students. Despite the large amount of material, most of them are able to digest and understand what they are taught, as evidenced from the oral examinations following this course—and this is not small-number statistics: my colleague Klaas de Boer and I together grade about 100 oral examinations per year, covering both parts of the introductory course. Some critical comments coming from students concern the extent of the material as well as its level. However, I do not see a rational reason why the level of an astronomy lecture should be lower than that of one in physics or mathematics.

Why did I turn this into a book? When preparing the concept for my lecture course, I soon noticed that there is no book which I can (or want to) follow. In particular, there are only a few astronomy textbooks in German, and they do not treat extragalactic astronomy and cosmology nearly to the extent and depth as I wanted for this course. Also, the choice of books on these topics in English is fairly limited—whereas a number of excellent introductory textbooks exist, most shy away from technical treatments of issues. However, many aspects can be explained

better if a technical argument is also given. Thus I hope that this text presents a field of modern astrophysics at a level suitable for the aforementioned group of people. A further goal is to cover extragalactic astronomy to a level such that the reader should feel comfortable turning to more professional literature.

When being introduced to astronomy, students face two different problems simultaneously. On the one hand, they should learn to understand astrophysical arguments—such as those leading to the conclusion that the central engine in AGNs is a black hole. On the other hand, they are confronted with a multitude of new terms, concepts and classifications, many of which can only be considered as historical burdens. Examples here are the classification of supernovae which, although based on observational criteria, do not agree with our current understanding of the supernova phenomenon, and the classification of the various types of AGN. In the lecture, I have tried to separate these two issues, clearly indicating when facts are presented where the students should 'just take note', or when astrophysical connections are uncovered which help to understand the properties of cosmic objects. The latter aspects are discussed in considerably more detail. I hope this distinction can still be clearly seen in this written version.

The order of the material in the course and in this book accounts for the fact that students in their first year of physics studies have a steeply rising learning curve; hence, I have tried to order the material partly according to its difficulty. For example, homogeneous world models are described first, whereas only later are the processes of structure formation discussed, motivated in the meantime by the treatment of galaxy clusters.

The topic and size of this book imply the necessity of a selection of topics. I want to apologize here to all of those colleagues whose favorite subject is not covered at the depth that they feel it deserves. I also took the freedom to elaborate on my own research topic—gravitational lensing—somewhat disproportionately. If it requires a justification: the basic equations of gravitational lensing are sufficiently simple that they and their consequences can be explained at an early stage in the astronomy education.

Many students are not only interested in the physical aspects of astronomy, they are also passionate observational astronomers. Many of them have been active in astronomy for years and are fascinated by phenomena occurring beyond the Earth. I have tried to provide a glimpse of this fascination at some points in the lecture course, for instance through some historical details, by discussing specific observations or instruments, or by highlighting some of the great achievements of modern cosmology. At such points, the text may deviate from the more traditional 'scholarly' style.

Producing the lecture notes, and their extension to a textbook, would have been impossible without the active help of several students and colleagues, whom I want to thank here. Jan Hartlap, Elisabeth Krause, and Anja von der Linden made numerous suggestions for improving the text, produced graphics or searched for figures, and TEXed tables—deep thanks go to them. Oliver Czoske, Thomas Erben, and Patrick Simon read the whole German version of the text in detail and made numerous constructive comments which led to a clear improvement of the text. Klaas de Boer and Thomas Reiprich read and commented on parts of this text. Searching for the sources of the figures, Leonardo Castaneda, Martin Kilbinger, Jasmin Pierloz, and Peter Watts provided valuable help. A first version of the English translation of the book was produced by Ole Markgraf, and I thank him for this heroic task. Furthermore, Kathleen Schrüfer, Catherine Vlahakis, and Peter Watts read the English version and made zillions of suggestions and corrections—I am very grateful to their invaluable help. Finally, I thank all my colleagues and students who provided encouragement and support for finishing this book.

The collaboration with Springer-Verlag was very fruitful. Thanks to Wolf Beiglböck and Ramon Khanna for their encouragement and constructive collaboration. Bea Laier offered to contact authors and publishers to get the copyrights for reproducing figures—without her invaluable help, the publication of the book would have been delayed substantially. The interaction with LE-TEX, where the book was produced, and in particular with Uwe Matrisch, was constructive as well.

Furthermore, I thank all those colleagues who granted permission to reproduce their figures here, as well as the public relations departments of astronomical organizations and institutes who, through their excellent work in communicating astronomical knowledge to the general public, play an invaluable role in our profession. In addition, they provide a rich source of pictorial material of which I made ample use for this book. Representative of those, I would like to mention the European Southern Observatory (ESO), the Space Telescope Science Institute (STScI), the NASA/SAO/CXC archive for Chandra data, and the Legacy Archive for Microwave Background Data Analysis (LAMBDA).

Contents

Introduction and overview

1

1.1 Introduction

The Milky Way, the galaxy in which we live, is but one of many galaxies. As a matter of fact, the Milky Way, also called the Galaxy, is a fairly average representative of the class of spiral galaxies. Two other examples of spiral galaxies are shown in Figs. 1.1 and 1.2, one of which we are viewing from above (face-on), the other from the side (edge-on). These are all stellar systems in which the majority of stars are confined to a relatively thin disk. In our own Galaxy, this disk can be seen as the band of stars stretched across the night sky, which led to it being named the Milky Way. Besides such disk galaxies, there is a second major class of luminous stellar systems, the elliptical galaxies. Their properties differ in many respects from those of the spirals.

It was less than a 100 years ago that astronomers first realized that objects exist outside our Milky Way and that our world is significantly larger than the size of the Milky Way. In fact, galaxies are mere islands in the Universe: the diameter of our Galaxy[1] (and other galaxies) is much smaller than the average separation between luminous galaxies. The discovery of the existence of other stellar systems and their variety of morphologies raised the question of the origin and evolution of these galaxies. Is there anything between the galaxies, or is it just empty space? Are there any other cosmic bodies besides galaxies? Questions like these motivated us to explore the Universe as a whole and its evolution. Is our Universe finite or infinite? Does it change over time? Does it have a beginning and an end? Mankind has long been fascinated by these questions about the origin and the history of our world. But for only a few decades have we been able to approach these questions in an empirical manner. As we shall discuss in this book, many of the questions have now been answered. However, each answer raises yet more questions, as we aim towards an ever increasing understanding of the physics of the Universe.

Fig. 1.1 The spiral galaxy NGC 1232 may resemble our Milky Way if it would be observed from 'above' (face-on). This image, observed with the VLT, has a size of $6'.8 \times 6'.8$, corresponding to a linear size of 60 kpc at its distance of 30 Mpc. If this was our Galaxy, our Sun would be located at a distance of 8.0 kpc from the center, orbiting around it at a speed of $\sim 220\,\mathrm{km/s}$. A full revolution would take us about 230×10^6 yr. The bright knots seen along the spiral arms of this galaxy are clusters of newly-formed stars, similar to bright young star clusters in our Milky Way. The different, more reddish, color of the inner part of this galaxy indicates that the average age of the stars there is higher than in the outer parts. The small galaxy at the lower left edge of the image is a companion galaxy that is distorted by the gravitational tidal forces caused by the spiral galaxy. Credit: European Southern Observatory

The stars in our Galaxy have very different ages. The oldest stars are about 12 billion years old, whereas in some regions stars are still being born today: for instance in the well-known Orion nebula. Obviously, the stellar content

[1] We shall use the terms 'Milky Way' and 'Galaxy' synonymously throughout.

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_1,

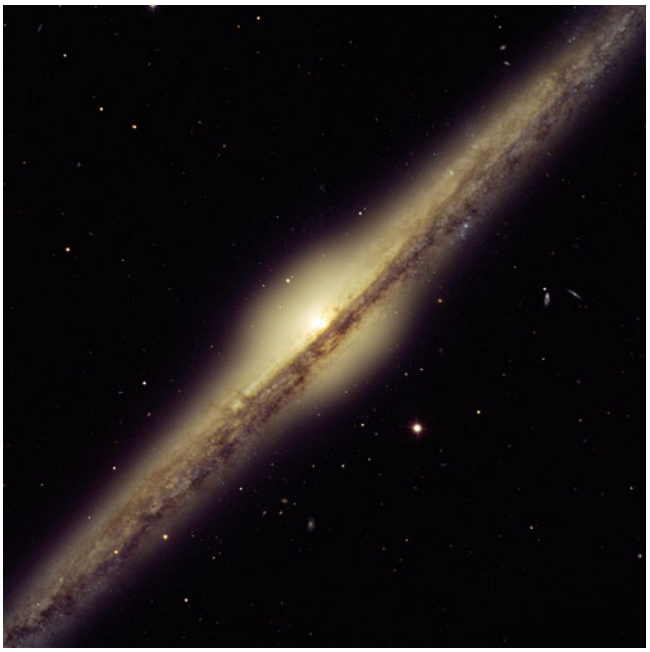


Fig. 1.2 We see the spiral galaxy NGC 4565 from the side (edge-on); an observer looking at the Milky Way from a direction which lies in the plane of the stellar disk ('from the side') may have a view like this. The disk is clearly visible, with its central region partly obscured by a layer of dust. One also sees the central bulge of this galaxy. As will be discussed at length later on, spiral galaxies like this one are surrounded by a halo of matter which is observed only through its gravitational action, e.g., by affecting the velocity of stars and gas rotating around the center of the galaxy. Credit: European Southern Observatory

of our Galaxy has changed over time. To understand the formation and evolution of the Galaxy, a view of its (and thus our own) past would be useful. Unfortunately, this is physically impossible. However, due to the finite speed of light, we see objects at large distances in an earlier state, as they were in the past. One can now try to identify and analyze such distant galaxies, which may have been the progenitors of galaxies like our own Galaxy, in this way reconstructing the main aspects of the history of the Milky Way. We will never know the exact initial conditions that led to the evolution of the Milky Way, but we may be able to find some characteristic conditions. Emerging from such initial states, cosmic evolution should produce galaxies similar to our own, which we would then be able to observe from the outside. On the other hand, only within our own Galaxy can we study the physics of galaxy evolution in situ.

We are currently witnessing an epoch of tremendous discoveries in astronomy. The technical capabilities in observation and data reduction are currently evolving at an enormous pace. Two examples taken from ground-based optical astronomy should serve to illustrate this.

In 1993 the first 10-m class telescope, the Keck telescope, was commissioned, the first increase in light-collecting power of optical telescopes since the completion of the 5-m mirror on Mt. Palomar in 1948. Currently, 13 telescopes with diameter above 8 m are in use, and planning for telescopes with 30 m diameter or more has begun. In recent years, our capabilities to find very distant, and thus very dim, objects and to examine them in detail have improved immensely thanks to the capability of these large optical telescopes.

A second example is the technical evolution and size of optical detectors. Since the introduction of CCDs (charge-coupled devices) in astronomical observations at the end of the 1970s, which replaced photographic plates as optical detectors, the sensitivity, accuracy, and data rate of optical observations have increased enormously. At the end of the 1980s, a camera with 1000×1000 pixels (*pic*ture *el*ement*s*) was considered a wide-field instrument. In 2003 a camera called Megacam began operating; it has $(18\,000)^2$ pixels and images a square degree of the sky at a sampling rate of $0\farcs2$ in a single exposure. Such a camera produces roughly 100 GB of data every night, the reduction of which requires fast computers and vast storage capacities. The largest astronomical CCD camera currently is that of the PanSTARRS-1 telescope, with more than 1.4 billion pixels, covering about $6\,\mathrm{deg}^2$ on the sky. But it is not only optical astronomy that is in a phase of major development; there has also been huge progress in instrumentation in other wavebands, allowing us a multi-wavelength view of the Universe (Fig. 1.3). Space-based observing platforms are playing a crucial role in this. We will consider this topic in Sect. 1.3.

These technical advances have led to a vast increase in knowledge and insight in astronomy, especially in extragalactic astronomy and cosmology. Large telescopes and sensitive instruments have opened up a window to the distant Universe. Since any observation of distant objects is inevitably also a view into the past, due to the finite speed of light, studying objects in the early Universe has become possible. Today, we can study galaxies which emitted the light we observe at a time when the Universe was less than 10 % of its current age; these galaxies are therefore in a very early evolutionary stage. We are thus able to observe the evolution of galaxies throughout the past history of the Universe. We have the opportunity to study the history of galaxies and thus that of our own Milky Way. We can examine at which epoch most of the stars that we observe today in the local Universe have formed because the history of star formation can be traced back to early epochs. In fact, it was found that star formation is largely hidden from our eyes and only observable with space-based telescopes operating in the far-infrared waveband.

One of the most fascinating discoveries of recent years is that most galaxies harbor a black hole in their center, with a characteristic mass of millions or even billions of Solar masses—so-called supermassive black holes (see Fig. 1.4). Although as soon as the first quasars were found in 1963 it was proposed that only processes around a supermassive

Fig. 1.3 This image of the galaxy M82 illustrates very clearly that any given waveband provides a rather restricted—and biased—view of cosmic objects. Shown is a composite image, obtained from three different telescopes. *Blue* color shows the X-ray radiation of this galaxy, as recorded by the X-ray satellite Chandra. The infrared light is shown in *red*, and was observed with the Spitzer Space Telescope. The optical light from M82 was recorded with the Hubble Space Telescope and is shown in *yellow-green*. Finally, line emission from hydrogen gas is displayed in *orange*. The distributions of radiation from different wavelengths is obviously very different; only the joint set of observations can provide us with an understanding of this galaxy. In fact, M82 is a rather special object, a so-called starburst galaxy, named because this galaxy forms new stars at a rate much higher than this happens in the Milky Way and other 'normal' spiral galaxies. The stars of the galaxy are distributed in a disk, as seen from the optical light, and most of the newly formed stars are located close to the center. The most massive of the stars explode in a supernova; these gigantic explosions can heat, and drive substantial amounts of gas and dust out of the galactic plane. The hot gas radiates X-rays and is clearly seen on both sides of the stellar disk, as well as the dust which emits in the infrared light. The image size is 7′.9, corresponding 8.5 kpc for distance of 3.7 Mpc. Credit: X-ray: NASA/CXC/JHU/D.Strickland; IR: NASA/JPL-Caltech/C. Engelbracht (University of Arizona); optical: NASA, ESA, and The Hubble Heritage Team

black hole would be able to produce the huge amount of energy emitted by these ultra-luminous objects, the idea that such black holes exist in normal galaxies is fairly recent. Even more surprising was the finding that the black hole mass is closely related to the other properties of its parent galaxy, thus providing a clear indication that the evolution of supermassive black holes is closely linked to that of their host galaxies.

Detailed studies of individual galaxies and of associations of galaxies, which are called galaxy groups or clusters of galaxies (see Fig. 1.5), led to the surprising result that these objects contain considerably more mass than is visible in the form of stars and gas. Analyses of the dynamics of galaxies and clusters show that only 10–20 % of their mass consists of stars, gas and dust that we are able to observe in emission or absorption. The largest fraction of their mass, however, is invisible. Hence, this hidden mass is called *dark matter*. We know of its presence only through its gravitational effects. The dominance of dark matter in galaxies and galaxy clusters was established in recent years from observations with radio, optical and X-ray telescopes, and it was also confirmed and quantified by other methods. However, we do not know what this dark matter consists of; the unambiguous evidence for its existence is called the 'dark matter problem'.

The nature of dark matter is one of the central questions not only in astrophysics but also poses a challenge to fundamental physics, unless the 'dark matter problem' has an astronomical solution. Does dark matter consist of non-luminous celestial bodies, for instance burned-out stars? Or is it a new kind of matter? Have astronomers indirectly proven the existence of a new elementary particle which has thus far escaped detection in terrestrial laboratories? If dark matter indeed consists of a new kind of elementary particle, which is the common presumption today, it should exist in

Fig. 1.4 The radio galaxy Hercules A, an elliptical galaxy seen at the center of this image. Superposed on this optical image is an image taken at radio wavelength, which shows a very extended source indeed. Two streams of ionized matter, so-called jets, are ejected on opposite sides of the galaxy, which terminate in two extended regions, the radio lobes. The energy of the jets is produced by a supermassive black hole with a mass of $M_\bullet \sim 2.5 \times 10^9\, M_\odot$. Credit: NASA, ESA, S. Baum and C. O'Dea (RIT), R. Perley and W. Cotton (NRAO/AUI/NSF), and the Hubble Heritage Team (STScI/AURA)

the Milky Way as well, in our immediate vicinity. Therefore, experiments which try to directly detect the constituents of dark matter with highly sensitive and sophisticated detectors have been set up in underground laboratories. Physicists and astronomers are eagerly waiting for results from the Large Hadron Collider (LHC), a particle accelerator at the European CERN research center which started regular operation in 2009, which produces particles at significantly higher energies than accessible before, and which in the first few years of operation already achieved a breakthrough with the discovery of the so-called Higgs particle. The hope is to find hints for new physics beyond the current Standard Model of particle physics, guiding us to extended models of particle physics which can accommodate an elementary particle that could serve as a constituent of dark matter.

Without doubt, the most important development in recent years is the establishment of a standard model of cosmology, i.e., the science of the Universe as a whole. The Universe is known to expand and it has a finite age; we now believe that we know its age with an uncertainty of as little as a few percent—it is $t_0 = 13.8\,\mathrm{Gyr}$. The Universe has evolved from a very dense and very hot state, the Big Bang, expanding and cooling over time. Even today, echoes of the Big Bang can be observed, for example in the form of the cosmic microwave background radiation. Accurate observations of

Fig. 1.5 The cluster of galaxies MACS J1206.2−0847, as seen in a multi-color image taken by the Hubble Space Telescope. The elliptical galaxy at the center of the image is the central galaxy of this massive galaxy cluster; many of the member galaxies of this clusters can be seen. They come in different shapes and colors, some being more reddish, which indicates stellar populations of large age, some being much bluer due to their ongoing star formation. In addition, this image shows some objects with rather peculiar shape. These are images of galaxies located behind the cluster whose observed shape is deformed by gravitational light deflection caused by the deep gravitational potential of the cluster. This image distortion can be used to determine the mass of this cluster, clearly showing that it contains far more mass than is seen in the visible cluster components. Credit: NASA, ESA, M. Postman (STScI), the CLASH Team, and the Hubble Heritage Team (STScI/AURA)

this background radiation, emitted some 380 000 years after the Big Bang, i.e., at a time $\approx 2.7 \times 10^{-5} t_0$, have made an important contribution to what we know today about the composition of the Universe. However, these results raise more questions than they answer: only $\sim 4\,\%$ of the energy content of the Universe can be accounted for by matter which is well-known from other fields of physics, the *baryonic matter* that consists mainly of atomic nuclei and electrons. About 25 % of the Universe consists of dark matter, as we already discussed in the context of galaxies and galaxy clusters. Recent observational results have shown that the mean density of dark matter dominates over that of baryonic matter also on cosmic scales.

Even more surprising than the existence of dark matter is the discovery that about 70 % of the Universe consists of something that today is called vacuum energy, or dark energy, and that is closely related to the cosmological constant introduced by Albert Einstein. The fact that various names do exist for it by no means implies that we have any idea what this dark energy is. It reveals its existence exclusively in its effect on the cosmic expansion, and it even dominates the expansion dynamics at the current epoch. Any efforts to estimate the density of dark energy from fundamental physics have failed hopelessly up to now. An estimate of the vacuum energy density using quantum mechanics results in a value that is roughly *120 orders of magnitude* larger than the value derived from cosmology. For the foreseeable future observational cosmology will be the only empirical probe for dark energy, and an understanding of its physical nature probably has to wait for quite a number of years. The existence of dark energy may well pose the greatest challenge to fundamental physics today.

In this book we will present a discussion of the extragalactic objects found in astronomy, but we will start with describing the Milky Way which, being a typical spiral galaxy, is considered a prototype of this class of stellar systems. The other central topic in this book is a presentation of modern astrophysical cosmology, which has experienced tremendous advances in recent years. Methods and results will be discussed in parallel. Besides providing an impression of the fascination that arises from astronomical observations and cosmological insights, astronomical methods and physical considerations will be our prime focus. We will start in the next section with a concise overview of the fields of extragalactic astronomy and cosmology. This is, on the one hand, intended to whet the reader's appetite and curiosity, and on the other hand to introduce some facts and technical terms that will be needed in what follows but which are discussed in detail only later in the book. In Sect. 1.3 we will describe some of the most important telescopes used in extragalactic astronomy today, and some of the most useful astronomical surveys having a broad range of applications are mentioned in Sect. 1.4.

1.2 Overview

1.2.1 Our Milky Way as a galaxy

The Milky Way is the only galaxy which we are able to examine in great detail. We can resolve individual stars and analyze them spectroscopically. We can perform detailed studies of the interstellar medium (ISM), such as the properties of molecular clouds and star forming regions. We can quantitatively examine extinction and reddening by dust. Furthermore, we can observe the local dynamics of stars and gas clouds as well as the properties of satellite galaxies (such the Magellanic Clouds). Finally, the Galactic center at a distance of only 8 kpc gives us the unique opportunity to examine the central region of a galaxy at very high resolution.[2] Only through a detailed understanding of our own Galaxy can we hope to understand the properties of other galaxies. Of course, we implicitly assume that the physical processes taking place in other galaxies obey the same laws of physics that apply to us. If this were not the case, we would barely have a chance to understand the physics of other objects in the Universe, let alone the Universe as a whole. We will return to this point shortly.

We will first discuss the properties of our own Galaxy. One of the main problems here, and in astronomy in general, is the determination of the distance to an object. Thus we will start by considering this topic. From the analysis of the distribution of stars and gas in the Milky Way we will then derive its structure. It is found that the Galaxy consists of several distinct components:

- a thin disk of stars and gas with a radius of about 20 kpc and a scale height of about 300 pc, which also hosts the Sun;
- a $\sim$ 1 kpc thick disk, which contains a different, older stellar population compared to the thin disk;
- a central bulge, as is also found in other spiral galaxies;
- and a nearly spherical halo which contains most of the globular clusters, some old stars, and gas with different densities and temperatures.

Figure 1.6 shows a schematic view of our Milky Way and its various components. For a better visual impression, Figs. 1.1 and 1.2 show two spiral galaxies, the former viewed from 'above' (face-on) and the latter from the 'side' (edge-on). In the former case, the spiral structure, from which this kind of galaxy derives its name, is clearly visible. The bright knots in the spiral arms are regions where young, luminous stars have recently formed. The image shows an obvious color

[2] 1 parsec (1 pc) is the common unit of distance in astronomy, with $1\,\mathrm{pc} = 3.086 \times 10^{18}\,\mathrm{cm}$. Also used are $1\,\mathrm{kpc} = 10^3\,\mathrm{pc}$, $1\,\mathrm{Mpc} = 10^6\,\mathrm{pc}$, $1\,\mathrm{Gpc} = 10^9\,\mathrm{pc}$. Other commonly used units and constants are listed in Appendix C.

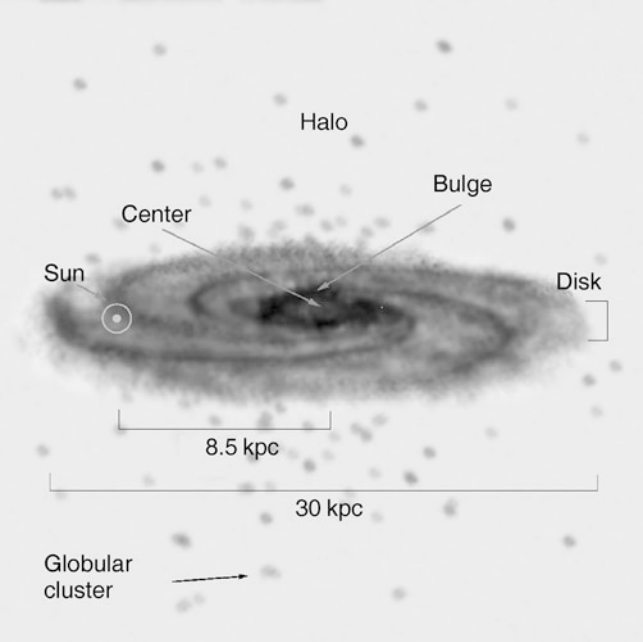


Fig. 1.6 Schematic structure of the Milky Way consisting of the disk, the central bulge with the Galactic center, and the spherical halo in which most of the globular clusters are located. The Sun orbits around the Galactic center at a distance of about 8 kpc

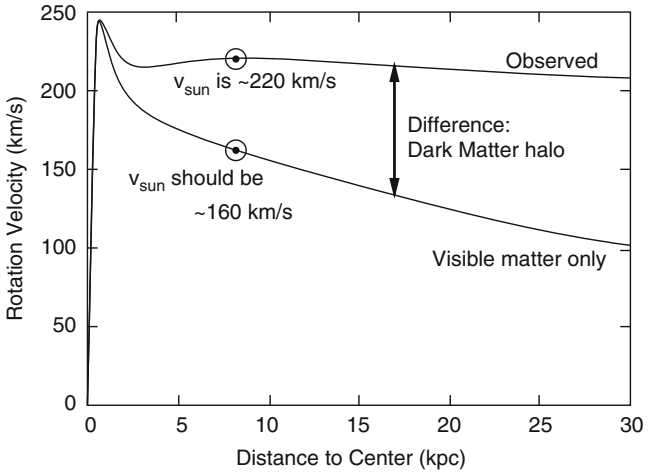


Fig. 1.7 The *upper curve* is the observed rotation curve $V(R)$ of our Galaxy, i.e., the rotational velocity of stars and gas around the Galactic center as a function of their galacto-centric distance. The *lower curve* is the rotation curve that we would predict based solely on the observed stellar mass of the Galaxy. The difference between these two curves is ascribed to the presence of dark matter, in which the Milky Way disk is embedded. This image is adapted from Nick Strobel's webpage at www.astronomynotes.com

gradient: the galaxy is redder in the center and bluest in the spiral arms—while star formation is currently taking place in the spiral arms, we find mainly old stars towards the center, especially in the bulge.

The Galactic disk rotates, with rotational velocity $V(R)$ depending on the distance R from the center. We can estimate the mass of the Galaxy from the distribution of the stellar light and the mean mass-to-light ratio of the stellar population, since gas and dust represent less than $\sim$ 10 % of the mass of the stars. From this mass estimate we can predict the rotational velocity as a function of radius simply from Newtonian mechanics. However, the observed rotational velocity of the Sun around the Galactic center is significantly higher than would be expected from the observed mass distribution. If $M(R_0)$ is the mass inside a sphere around the Galactic center with radius $R_0 \approx 8$ kpc, then the rotational velocity from Newtonian mechanics[3] is

$$V_0 = \sqrt{\frac{G\,M(R_0)}{R_0}}\,. \tag{1.1}$$

From the visible matter in stars we would expect a rotational velocity of $\sim$ 160 km/s, but we observe $V_0 \sim 220$ km/s (see Fig. 1.7). This discrepancy, and the shape of the rotation curve $V(R)$ for larger distances R from the Galactic center, indicates that our Galaxy contains significantly more mass than is visible in the form of stars.[4] This additional mass is called *dark matter*. Its physical nature is still unknown. The main candidates are weakly interacting elementary particles like those postulated by some elementary particle theories, but they have yet not been detected in the laboratory. Macroscopic objects (i.e., celestial bodies) are also in principle viable candidates if they emit very little light. We will discuss experiments which allow us to identify such macroscopic objects and come to the conclusion that the solution of the dark matter problem probably can not be found in astronomy, but rather most likely in particle physics.

The stars in the various components of our Galaxy have different properties regarding their age and their chemical composition. By interpreting this fact one can infer some aspects of the evolution of the Galaxy. The relatively young age of the stars in the thin disk, compared to that of the older population in the bulge, suggests different phases in the formation and evolution of the Milky Way. Indeed, our Galaxy is a highly dynamic object that is still changing today. We see cold gas falling into the Galactic disk and hot gas outflowing. Currently the small neighboring Sagittarius dwarf galaxy is being torn apart in the tidal gravitational field of the Milky

[3] We use standard notation: G is the Newtonian gravitational constant, c the speed of light.

[4] Strictly speaking, (1.1) is valid only for a spherically symmetric mass distribution. However, the rotational velocity for an oblate density distribution does not differ much, so we can use this relation as an approximation.

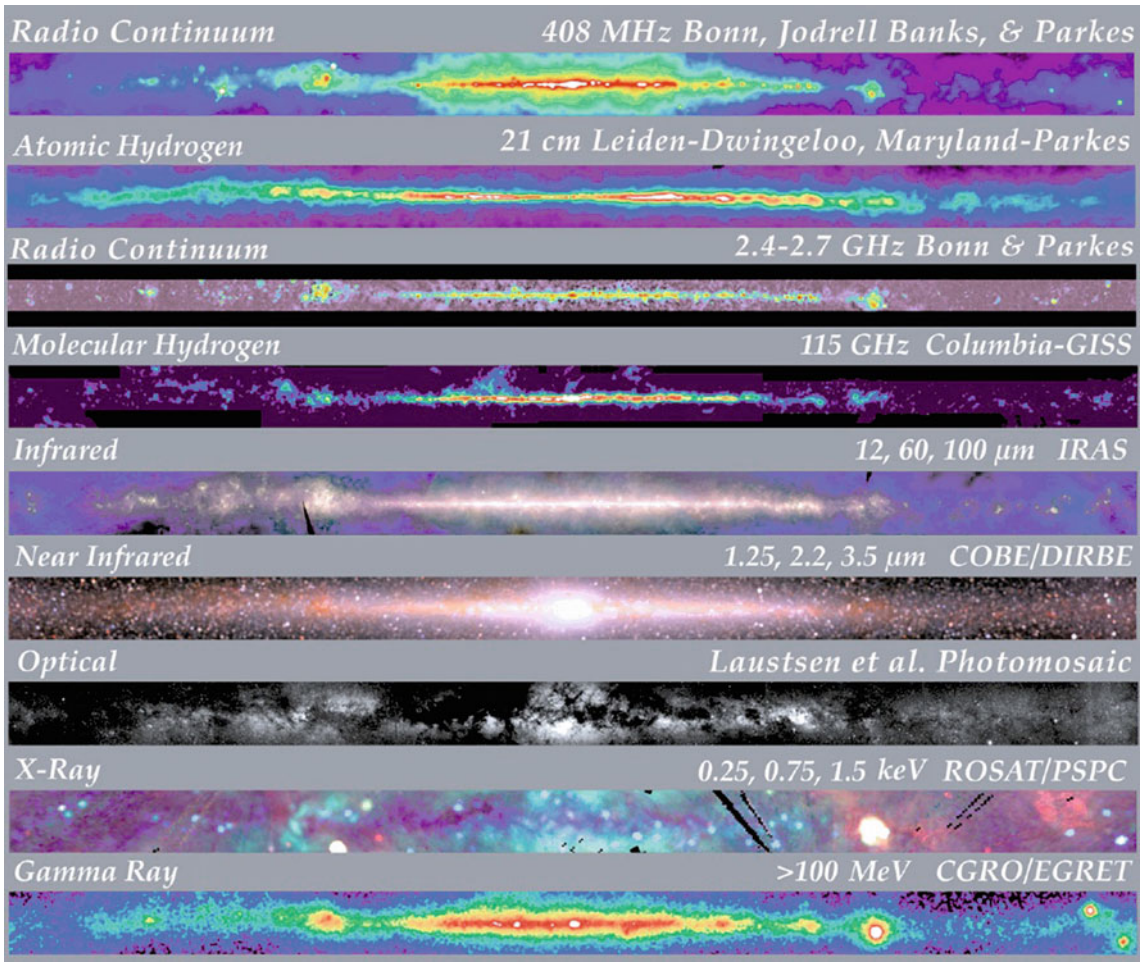


Fig. 1.8 The Galactic disk observed in nine different wavebands. Its appearance differs strongly in the various images; for example, the distribution of atomic hydrogen and of molecular gas is much more concentrated towards the Galactic plane than the distribution of stars observed in the near-infrared, the latter clearly showing the presence of a central bulge. The absorption by dust at optical wavelengths is also clearly visible and can be compared to that in Fig. 1.2. Credit: NASA's Goddard Space Flight Center

Way and will merge with it in the (cosmologically speaking) near future.

One cannot see far through the disk of the Galaxy at optical wavelengths due to extinction by dust. Therefore, the immediate vicinity of the Galactic center can be examined only in other wavebands, especially the infrared (IR) and the radio parts of the electromagnetic spectrum (see also Fig. 1.8). The Galactic center is a highly complex region but we have been able to study it in recent years thanks to various substantial improvements in IR observations regarding sensitivity and angular resolution. Proper motions, i.e., changes of the positions on the sky with time, of bright stars close to the center have been observed. They enable us to determine the mass M inside a volume of radius $\sim 0.1\,\mathrm{pc}$ to be $M(0.1\,\mathrm{pc}) \sim 4 \times 10^6 M_\odot$. Although the data do not allow us to make a totally unambiguous interpretation of this mass concentration there is no plausible alternative to the conclusion that the center of the Milky Way harbors a supermassive black hole (SMBH) of roughly this mass. And yet this SMBH is far less massive than the ones that have been found in many other galaxies.

Unfortunately, we are unable to look at our Galaxy from the outside. This view from the inside renders it difficult to observe the global properties of the Milky Way. The structure and geometry of the Galaxy, e.g., its spiral arms, are hard to identify from our location. In addition, the extinction by dust hides large parts of the Galaxy from our view (see Fig. 1.9), so that the global parameters of the Milky Way (like its total luminosity) are difficult to measure. These parameters are estimated much better from outside, i.e., in other similar spiral galaxies. In order to understand the large-scale properties of our Galaxy, a comparison with similar galaxies which we can examine in their entirety is extremely helpful. Only by combining the study of the Milky Way with that of other galaxies can we hope to fully understand the physical nature of galaxies and their evolution.

Fig. 1.9 The galaxy Dwingeloo 1 is only five times more distant than our closest large neighboring galaxy, Andromeda, yet it was not discovered until the 1990s because it hides behind the Galactic center. The absorption in this direction and numerous bright stars prevented it from being discovered earlier. The figure shows an image observed with the Isaac Newton Telescope in the V, R, and I bands. Credit: S. Hughes & S. Maddox; Isaac Newton Telescope

Fig. 1.10 NGC 2997 is a typical spiral galaxy, with its disk inclined by about 45° with respect to the line-of-sight. Like most spiral galaxies it has two spiral arms; they are significantly bluer than other parts of the galaxy. This is caused by ongoing star formation in these regions so that young, hot and thus blue stars are present in the arms, whereas the center of the galaxy, especially the bulge, consists mainly of old stars. Credit: M. Altmann, Sternwarte der Universität Bonn

1.2.2 The world of galaxies

Next we will discuss the properties of other galaxies. The two main types of galaxies are spirals (like the Milky Way, see also Fig. 1.10) and elliptical galaxies (Fig. 1.11). Besides these, there are additional classes such as irregular and dwarf galaxies, active galaxies, and starburst galaxies, where the latter have a very high star-formation rate in comparison to normal galaxies. These classes differ not only in their morphology, which forms the basis for their classification, but also in their physical properties such as color (indicating a different stellar content), internal reddening (depending on their dust content), amount of interstellar gas, star-formation rate, etc. Galaxies of different morphologies have evolved in different ways.

Fig. 1.11 M87 is a very luminous elliptical galaxy in the center of the Virgo cluster, at a distance of about 18 Mpc. The diameter of the visible part of this galaxy is about 40 kpc; it is significantly more massive than the Milky Way ($M > 3 \times 10^{12} M_{\odot}$). We will frequently refer to this galaxy: it is not only an excellent example of a central cluster galaxy but also a representative of the family of 'active galaxies'. It is a strong radio emitter (radio astronomers also know it as Virgo A), and it has an optical jet in its center. Credit: S. Frey & J.E. Gunn, Princeton University

Spiral galaxies are stellar systems in which active star formation is still taking place today, whereas elliptical galaxies consist mainly of old stars—their star formation was terminated a long time ago. The S0 galaxies, an intermediate type, show a disk similar to that of spiral galaxies but like ellipticals they consist mainly of old stars, i.e., stars of low mass and low temperature. Ellipticals and S0 galaxies together are often called *early-type galaxies*, whereas spirals are termed *late-type galaxies*. These names do not imply any interpretation but exist only for historical reasons.

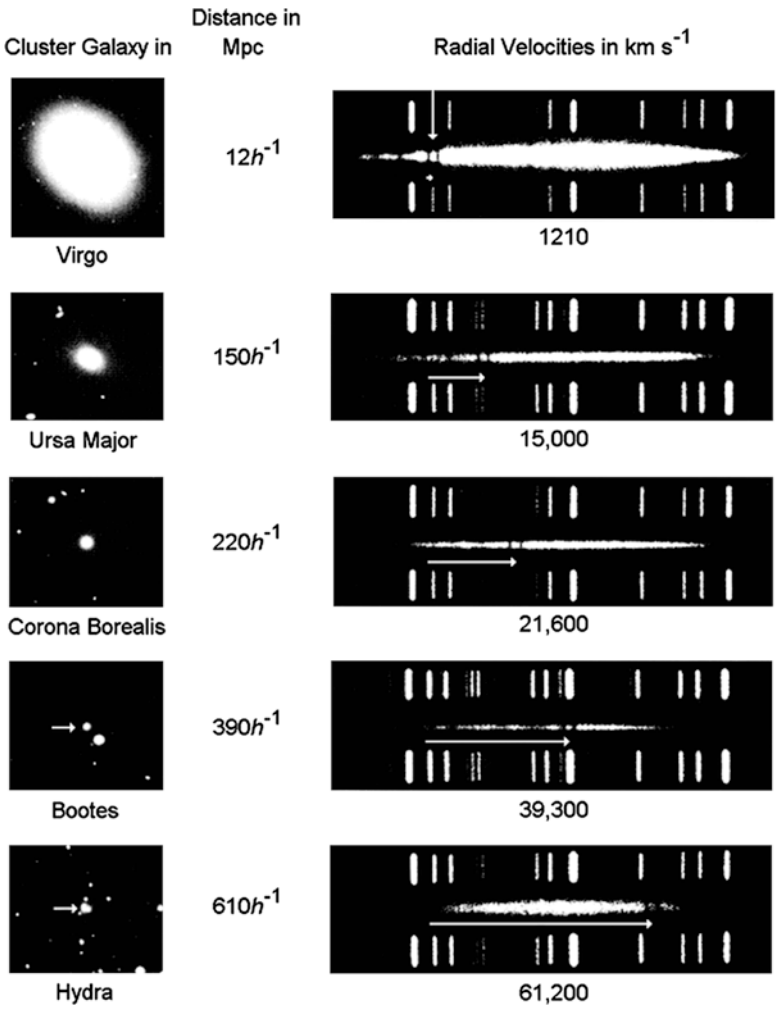


Fig. 1.12 The spectra of galaxies show characteristic spectral lines, e.g., the H+K lines of calcium. These lines, however, do not appear at the wavelengths measured in the laboratory but are in general shifted towards longer wavelengths. This is shown here for a set of sample galaxies, with distance increasing from top to bottom. The shift in the lines, interpreted as being due to the Doppler effect, allows us to determine the relative radial velocity—the larger it is, the more distant the galaxy is. The discrete lines above and below the spectra are for calibration purposes only. Credit: Hale Observatories; J. Silk, The Big Bang, 2nd Ed.

The disks of spiral galaxies rotate differentially. As for the Milky Way, one can determine the mass from the rotational velocity using the Kepler law (1.1). One finds that, contrary to the expectation from the distribution of light, the rotation curve does not decline at larger distances from the center. *Like our own Galaxy, spiral galaxies contain a large amount of dark matter; the visible matter is embedded in a halo of dark matter.* We can only get rough estimates of the extent of this halo, but there are strong indications that it is substantially larger than the extent of the visual matter. For instance, the rotation curve is flat up to the largest radii where one still finds gas to measure the velocity. Studying dark matter in elliptical galaxies is more complicated, but the existence of dark halos has also been proven for ellipticals.

The Hertzsprung–Russell diagram of stars, or their color-magnitude diagram (see Appendix B), has turned out to be the most important diagram in stellar astrophysics. The fact that most stars are aligned along a one-dimensional sequence, the main sequence, led to the conclusion that, for main sequence stars, the luminosity and the surface temperature are not independent parameters. Instead, the properties of such stars are in principle characterized by only a single parameter: the stellar mass. We will also see that the various properties of galaxies are not independent parameters. Rather, dynamical properties (such as the rotational velocity of spirals) are closely related to the luminosity. These scaling relations are of similar importance to the study of galaxies as the Hertzsprung–Russell diagram is for stars. In addition, they turn out to be very convenient tools for the determination of galaxy distances.

Like our Milky Way, other galaxies also seem to harbor a SMBH in their center. We obtained the astonishing result that the mass of such a SMBH is closely related to the velocity distribution of stars in elliptical galaxies or in the bulge of spirals. The physical reason for this close correlation is as yet not known in detail, but it strongly suggests a joint evolution of galaxies and their SMBHs.

1.2.3 The Hubble expansion of the Universe

The radial velocity of galaxies, measured by means of the Doppler shift of spectral lines (Fig. 1.12), is positive for nearly all galaxies, i.e., they appear to be moving away

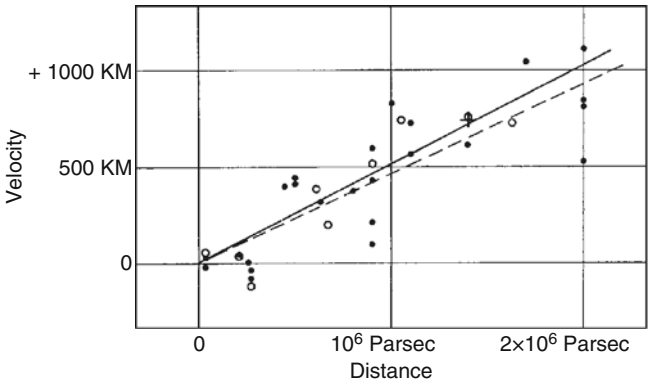


Fig. 1.13 The original 1929 version of the Hubble diagram shows the radial velocity of galaxies as a function of their distance. The reader may notice that the velocity axis is labeled with erroneous units—of course they should read km/s. While the radial (escape) velocity is easily measured by means of the Doppler shift in spectral lines, an accurate determination of distances is much more difficult; we will discuss methods of distance determination for galaxies in Sect. 3.9. Hubble has underestimated the distances considerably, resulting in too high a value for the Hubble constant. Only very few and very close galaxies show a blueshift, i.e., they move towards us; one of these is Andromeda (=M31). Adapted from: E. Hubble 1929, *A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae*, Proc. Nat. Academy Sciences 15, No. 3, March 15, 1929, Fig. 1

from us. In 1928, Edwin Hubble discovered that this escape velocity v increases with the distance of the galaxy. He identified a linear relation between the radial velocity v and the distance D of galaxies, called the Hubble law,

$$\boxed{v = H_0 D} \,, \tag{1.2}$$

where H_0 is a constant. If we plot the radial velocity of galaxies against their distance, as is done in the Hubble diagram of Fig. 1.13, the resulting points are approximated by a straight line, with the slope being determined by the constant of proportionality, H_0, which is called the *Hubble constant*. The fact that all galaxies seem to move away from us with a velocity which increases linearly with their distance is interpreted such that the Universe is expanding. We will see later that this *Hubble expansion* of the Universe is a natural property of cosmological world models.

For a long time, the value of H_0 was uncertain by almost a factor of two. However, in recent years the uncertainty was reduced to about 5 %, yielding

$$H_0 = (71 \pm 4)\,\mathrm{km\,s^{-1}Mpc^{-1}} \,, \tag{1.3}$$

obtained from several different methods which will be discussed later. The error margins vary for the different methods. The main problem in determining H_0 is in measuring the absolute distance of galaxies (as will be discussed in Sect. 3.9), whereas Doppler shifts are easily measurable. If one assumes (1.2) to be valid, the radial velocity of a galaxy is a measure of its distance. One defines the *redshift*, z, of an object from the wavelength shift in spectral lines,

$$z := \frac{\lambda_{\mathrm{obs}} - \lambda_0}{\lambda_0} \,, \quad \lambda_{\mathrm{obs}} = (1+z)\lambda_0 \,, \tag{1.4}$$

with λ_0 denoting the wavelength of a spectral transition in the rest-frame of the emitter and λ_{obs} the observed wavelength. For instance, the Lyman-α transition, i.e., the transition from the first excited level to the ground state in the hydrogen atom is at $\lambda_0 = 1216\,\text{Å}$. For small redshifts,

$$v \approx zc \,, \tag{1.5}$$

whereas this relation has to be modified for large redshifts, together with the interpretation of the redshift itself.[5] Combining (1.2) and (1.5), we obtain

$$D \approx \frac{zc}{H_0} \approx 3000\, z\, h^{-1}\mathrm{Mpc} \,, \tag{1.6}$$

where the uncertainty in determining H_0 is parametrized by the scaled Hubble constant h, defined as

$$\boxed{H_0 = h\,100\,\mathrm{km\,s^{-1}Mpc^{-1}}} \,. \tag{1.7}$$

Distance determinations based on redshift therefore always contain a factor of h^{-1}, as seen in (1.6). With the recent determination of the Hubble constant (1.3), we have $h = 0.71 \pm 0.04$. It needs to be emphasized once more that (1.5) and (1.6) are valid only for $z \ll 1$; the generalization for larger redshifts will be discussed in Sect. 4.3. Nevertheless, z is also a measure of distance for large redshifts.

1.2.4 Active galaxies and starburst galaxies

A special class of galaxies are the so-called active galaxies which have a very strong energy source in their center (active galactic nucleus, AGN). The best-known representatives of these AGNs are the quasars, objects typically at high redshift and with quite exotic properties. Their spectrum

[5] What is observed is the wavelength shift of spectral lines. Depending on the context, it is interpreted either as a radial velocity of a source moving away from us—for instance, if we measure the radial velocity of stars in the Milky Way—or as a cosmological escape velocity, as is the case for the Hubble law. It is in principle impossible to distinguish between these two interpretations, because a galaxy not only takes part in the cosmic expansion but it can, in addition, have a so-called peculiar velocity. We will therefore use the words 'Doppler shift' and 'redshift', respectively, and 'radial velocity' depending on the context, but always keeping in mind that both are measured by the shift of spectral lines. Only when observing the distant Universe where the Doppler shift is fully dominated by the cosmic expansion will we exclusively call it 'redshift'.

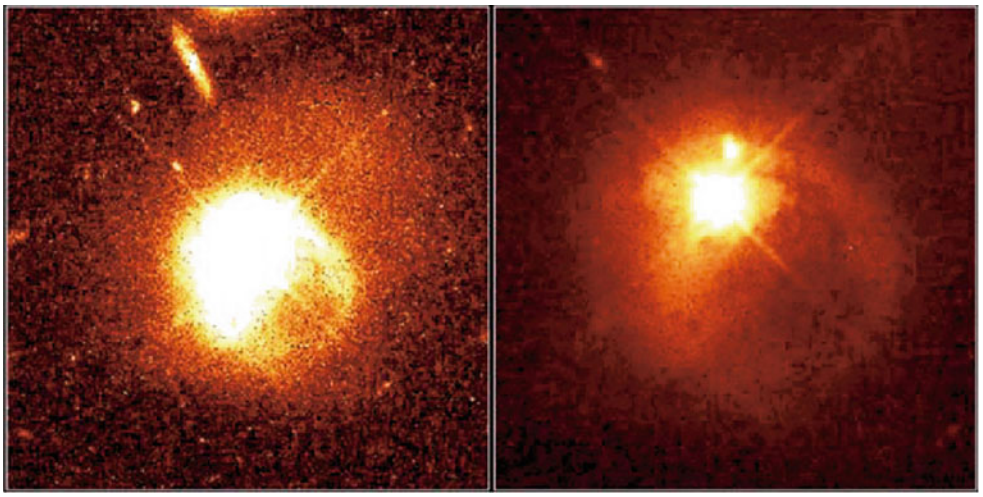


Fig. 1.14 The quasar PKS 2349 is located at the center of a galaxy, its host galaxy. The two images shown here differ only in their brightness contrast. The diffraction spikes (diffraction patterns caused by the suspension of the telescope's secondary mirror) in the middle of the object show that the center of the galaxy contains a point source, the actual quasar, which is significantly brighter than its host galaxy. The galaxy shows clear signs of distortion, visible as large and thin tidal tails. The tails are caused by a neighboring galaxy that is visible in the *right-hand image*, just above the quasar; it is about the size of the Large Magellanic Cloud. Quasar host galaxies are often distorted or in the process of merging with other galaxies. Credit: J. Bahcall (IAS, Princeton), M. Disney (University of Wales), NASA

shows strong emission lines which can be extremely broad, with a relative width of $\Delta\lambda/\lambda \sim 0.03$. The line width is caused by very high random velocities of the gas which emits these lines: if we interpret the line width as due to Doppler broadening resulting from the superposition of lines of emitting gas with a very broad velocity distribution, we obtain velocities of typically $\Delta v \sim 10\,000\,\mathrm{km/s}$. The central source in these objects is much brighter than the other parts of the galaxy, making these sources appear nearly point-like on optical images. Only with the Hubble Space Telescope (HST) did astronomers succeed in detecting structure in the optical emission for a large sample of quasars (Fig. 1.14).

Many properties of quasars resemble those of Seyfert type I galaxies, which are galaxies with a very luminous nucleus and very broad emission lines. For this reason, quasars are considered as particularly luminous members of this class. The total luminosity of quasars is extremely large, with some of them emitting more than a 1000 times the luminosity of our Galaxy. In addition, this radiation must originate from a very small spatial region whose size can be estimated, e.g., from the variability time-scale of the source. Due to these and other properties which will be discussed in Chap. 5, it is concluded that the nuclei of active galaxies must contain a supermassive black hole as the central powerhouse. The radiation is produced by matter falling towards this black hole, a process called accretion, thereby converting its gravitational potential energy into kinetic energy. If this kinetic energy is then transformed into internal energy (i.e., heat) as happens in the so-called accretion disk due to friction, it can get radiated away. This is in fact an extremely efficient process of energy production. For a given mass, the accretion onto a black hole is about ten times more efficient than the nuclear fusion of hydrogen into helium. AGNs often emit radiation across a very large portion of the electromagnetic spectrum, from radio up to X-ray and gamma radiation.

Spiral galaxies still form stars today, indeed star formation is a common phenomenon in galaxies. In addition, there are galaxies with a considerably higher star-formation rate than 'normal' spirals. These galaxies are undergoing a burst of star formation and are thus known as *starburst galaxies*. Their star-formation rates are typically between 10 and $300 M_\odot/\mathrm{yr}$, whereas our Milky Way gives birth to about $2 M_\odot/\mathrm{yr}$ of new stars. This vigorous star formation often takes place in localized regions, e.g., in the vicinity of the center of the respective galaxy. Starbursts are substantially affected, if not triggered, by disturbances in the gravitational field of the galaxy, such as those caused by galaxy interactions. Such starburst galaxies (see Fig. 1.15) can be extremely luminous in the far-infrared (FIR); they emit up to 98 % of their total luminosity in this part of the spectrum. This happens by dust emission: dust in these galaxies absorbs a large proportion of the energetic UV radiation produced by the massive stars in the star-formation region, thereby heats up, and then re-emits this energy in the form of thermal radiation in the FIR.

1.2.5 Voids, clusters of galaxies, and dark matter

The likelihood of galaxies to interact (Fig. 1.16) is enhanced by the fact that galaxies are not randomly distributed in

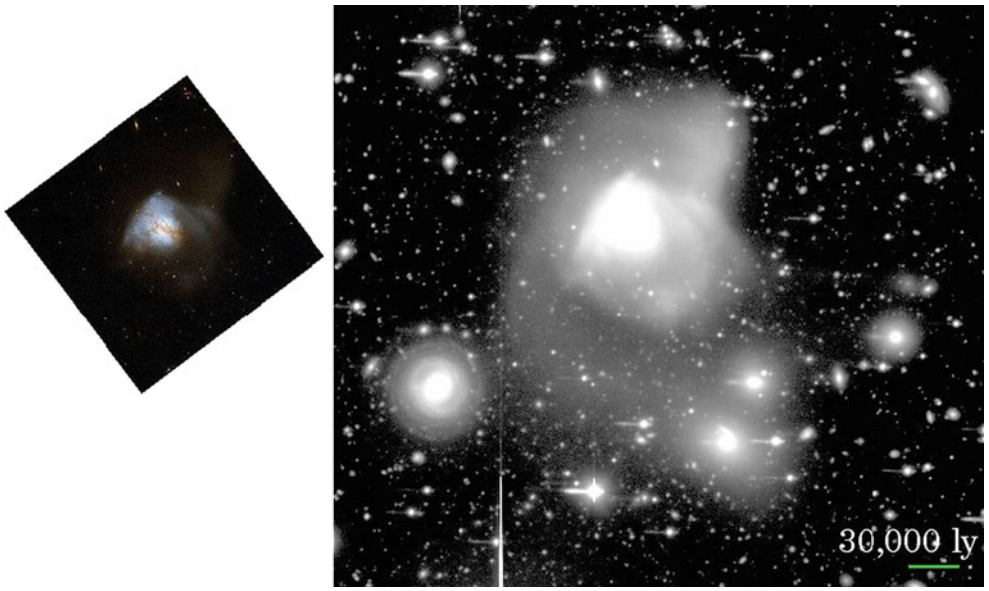


Fig. 1.15 Arp 220 is the most luminous object in the local Universe. Originally cataloged as a peculiar galaxy, the infrared satellite IRAS later discovered its enormous luminosity ($L \sim 10^{12} L_{\odot}$) in the infrared (IR). Arp 220 is the prototype of ultra-luminous infrared galaxies (ULIRGs). The *left panel* shows a near-IR image taken with the Hubble Space Telescope (HST). The *right panel* shows a spectacular image taken with the Subaru telescope on Mauna Kea; it unveils the structure of this object. With two colliding spiral galaxies in the center of Arp 220, the disturbances in the interstellar medium caused by this collision trigger a starburst. Dust in the galaxy absorbs most of the ultraviolet (UV) radiation from the young hot stars and re-emits it in the IR. Credit: *Left*: Hubble Space Telescope/NASA. *Right*: Ehime University/NAOJ. ©National Astronomical Observatory of Japan, ©NAOJ, Courtesy of the National Astronomical Observatory of Japan, and Courtesy of NAOJ

Fig. 1.16 Two spiral galaxies interacting with each other. NGC 2207 (on the *left*) and IC 2163 are not only close neighbors in projection: the strong gravitational tidal interaction they are exerting on each other is clearly visible in the pronounced tidal arms, particularly visible to the right of the right-hand galaxy. Furthermore, a bridge of stars is seen to connect these two galaxies, also due to tidal gravitational forces. This image was taken with the Hubble Space Telescope. Credit: The Hubble Heritage Project, STScI, NASA

space. The projection of galaxies on the celestial sphere, for instance, shows a distinct structure. In addition, measuring the distances of galaxies allows a determination of their three-dimensional distribution. One finds a strong correlation of the galaxy positions. There are regions in space that have a very high galaxy density, but also regions where nearly no galaxies are seen at all. The latter are called *voids*. Such voids can have diameters of up to $30h^{-1}$Mpc.

Clusters of galaxies are gravitationally bound systems of a hundred or more galaxies in a volume of diameter

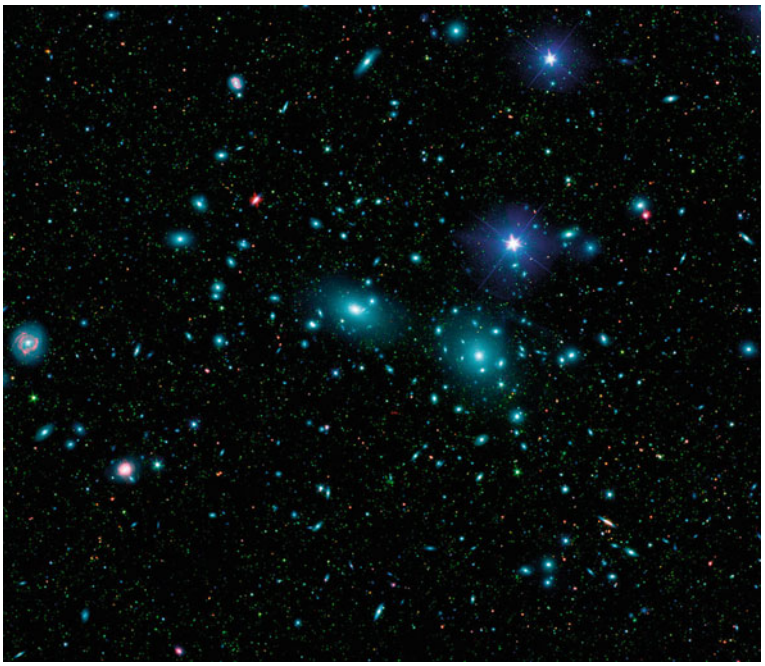


Fig. 1.17 The Coma cluster of galaxies, at a distance of roughly 90 Mpc from us, is the closest massive regular cluster of galaxies. Almost all brighter objects visible in this image of the central region of Coma are galaxies associated with the cluster—Coma contains more than a 1000 luminous galaxies. This images is a color composite made from optical data from the Sloan Digital Sky Survey (SDSS), shown in *blue*, and infrared data from the Spitzer Space Telescope, shown in *red* and *green*, for the longer and shorter wavelength, respectively. Credit: NASA/JPL-Caltech/GSFC/SDSS

$\sim 2\ h^{-1}$Mpc. Clusters predominantly contain early-type galaxies, so there is not much star formation taking place any more. Some clusters of galaxies seem to be rather circular in projection, others have a highly elliptical or irregular distribution of galaxies; some even have more than one center. The cluster of galaxies closest to us is the Virgo cluster, at a distance of $\sim$ 18 Mpc; it is a cluster with an irregular galaxy distribution. The closest regular cluster is Coma, at a distance of $\sim$ 90 Mpc.[6] Coma (Fig. 1.17) contains about 1000 luminous galaxies, of which 85 % are early-type galaxies.

In 1933, Fritz Zwicky measured the radial velocities of the galaxies in Coma and found that their distribution around the mean has a dispersion of about 1000 km/s. From the total luminosity of all its galaxies the mass of the cluster can be estimated. If the stars in the cluster galaxies have an average mass-to-light ratio (M/L) similar to that of our Sun, we would conclude $M = (M_\odot/L_\odot)L$. However, stars in early-type galaxies are on average slightly less massive than the Sun and thus have a slightly higher M/L.[7] Thus, the above mass estimate needs to be increased by a factor of ~ 10.

Zwicky then estimated the mass of the cluster by multiplying the luminosity of its member galaxies with the mass-to-light ratio. From this mass and the size of the cluster, he could then estimate the velocity that a galaxy needs to have in order to escape from the gravitational field of the cluster—the escape velocity. He found that the characteristic peculiar velocity of cluster galaxies (i.e., the velocity relative to the mean velocity) is substantially larger than this escape velocity. In this case, the galaxies of the cluster would fly apart on a time-scale of about 10^9 yr—the time it takes a galaxy to cross through the cluster once—and, consequently, the cluster would dissolve. However, since Coma seems to be a relaxed cluster, i.e., it is in equilibrium and thus its age is definitely larger than the dynamical time scale of 10^9 yr, Zwicky concluded that the Coma cluster contains significantly more mass than the sum of the masses of its galaxies. Using the virial theorem[8] he was able to estimate the mass of the cluster from the velocity distribution of the galaxies. This was the first clear indicator of the existence of dark matter.

X-ray satellites later revealed that clusters of galaxies are strong sources of X-ray radiation. They contain hot gas, with temperatures ranging from 10^7 up to 10^8 K (Fig. 1.18). This gas temperature is another measure for the depth of the cluster's potential well, since the hotter the gas is, the

[6]The distances of these two clusters are not determined from redshift measurements, but by direct methods that will be discussed in Sect. 3.9; such direct measurements are one of the most successful methods of determining the Hubble constant.

[7]In Chap. 3 we will see that for stars in spiral galaxies $M/L \sim 3M_\odot/L_\odot$ on average, while for those in elliptical galaxies a larger value of $M/L \sim 10M_\odot/L_\odot$ applies. Here and throughout this book, mass-to-light ratios are quoted in Solar units.

[8]The virial theorem in its simplest form says that, for an isolated dynamical system in a stationary state of equilibrium, the kinetic energy is just half the potential energy,

$$E_{\rm kin} = \frac{1}{2}\left|E_{\rm pot}\right| \ . \tag{1.8}$$

In particular, the system's total energy is $E_{\rm tot} = E_{\rm kin} + E_{\rm pot} = E_{\rm pot}/2 = -E_{\rm kin}$.

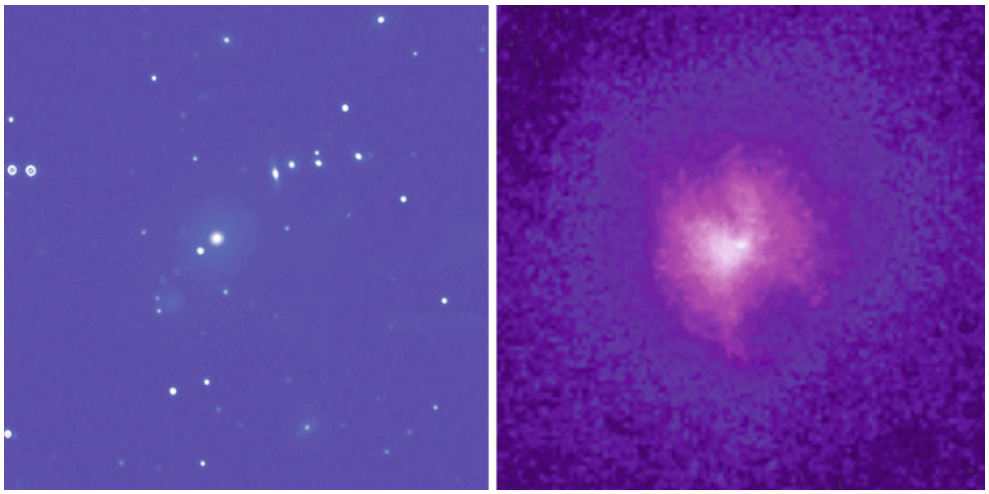


Fig. 1.18 The Hydra A cluster of galaxies. The *left-hand figure* shows an optical image, the one *on the right* an image taken with the X-ray satellite Chandra. The cluster has a redshift of $z \approx 0.054$ and is thus located at a distance of about 250 Mpc. The X-ray emission originates from gas at a temperature of 40×10^6K which fills the space between the cluster galaxies. In the center of the cluster, the gas is cooler by about 15 %. Credit: Optical: B. McNamara, La Palma; X-ray: NASA/CXC/SAO

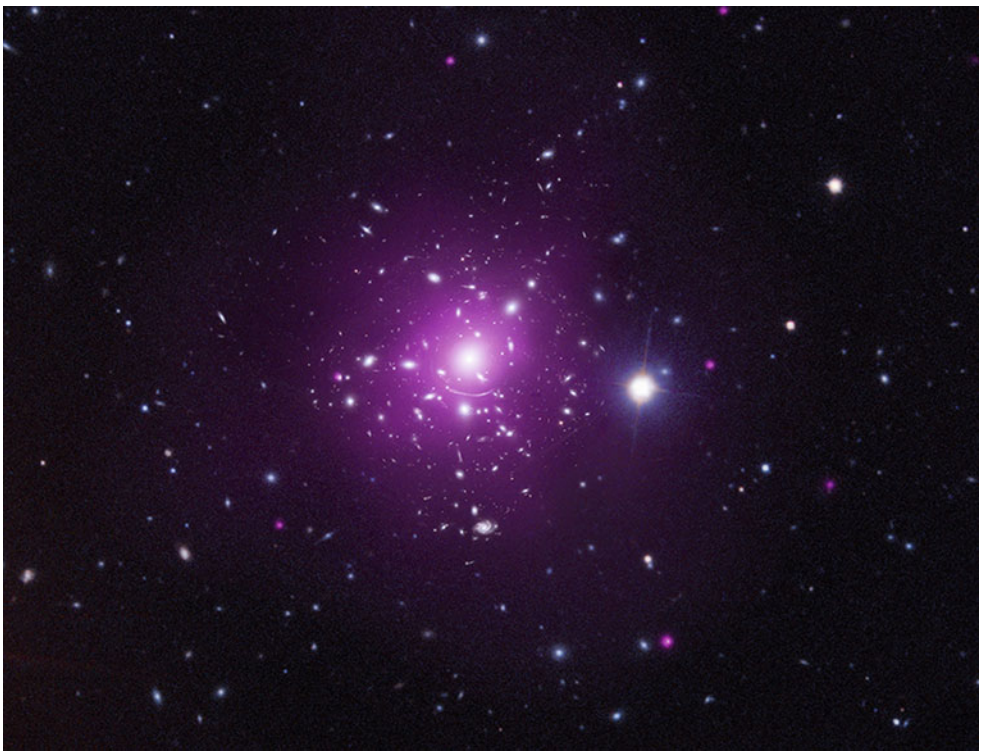


Fig. 1.19 The cluster of galaxies Abell 383, as seen in optical light, superposed by an image taken at X-ray energies (*purple*) with the Chandra satellite observatory. The space between the galaxies is filled by a hot gas, with temperature of about 50 million degrees, which emits the energetic X-ray radiation. The cluster is at a redshift of $z = 0.19$, corresponding to a distance of about 800 Mpc, and has an estimated mass of $\sim 3 \times 10^{14}\, M_\odot$. Credit: X-ray: NASA/CXC/Caltech/A. Newman et al./Tel Aviv/A. Morandi & M. Limousin; Optical: NASA/STScI, ESO/VLT, SDSS

deeper the potential well has to be to prevent the gas from escaping via evaporation. Mass estimates based on the X-ray temperature result in values that are comparable to those from the velocity dispersion of the cluster galaxies. Whereas the X-ray emitting gas provides a further mass component of ordinary, baryonic matter—in fact, the X-ray emitting gas contains more mass than the stars in the cluster galaxies—the total mass of clusters exceeds that of stars and gas by a factor of about five, thus clearly confirming the hypothesis of the existence of dark matter in clusters (Fig. 1.19). A third method for determining cluster masses, the so-called gravitational lensing effect, utilizes the fact that light is deflected in a gravitational field. The angle through which light rays are bent due to the presence of a massive object depends on the mass of that object. From observation and analysis of the gravitational lensing effect in clusters of galaxies, cluster masses are derived that are in agreement with those from the two other methods. Therefore, clusters of galaxies are a second class of cosmic objects whose mass is dominated by dark matter.

Clusters of galaxies are cosmologically young structures. Their dynamical time-scale, i.e., the time in which the mass distribution in a cluster settles into an equilibrium state, is estimated as the time it takes a member galaxy to fully cross the cluster once. With a characteristic velocity of $v \sim 1000\,\mathrm{km/s}$ and a diameter of $2R \sim 2\,\mathrm{Mpc}$ one thus finds

$$t_{\rm dyn} \sim \frac{2R}{v} \sim 2 \times 10^9\,{\rm yr}\,. \tag{1.9}$$

As we will later see, the Universe is about 14×10^9 yr old. During this time galaxies have not had a chance to cross the cluster many times. Therefore, clusters still contain, at least in principle, information about their initial state. Most clusters have not had the time to fully relax and evolve into a state of equilibrium that would be largely independent of their initial conditions. Comparing this with the time taken for the Sun to rotate around the center of the Milky Way—about 2×10^8 yr—galaxies thus have had plenty of time to reach their state of equilibrium.

Besides massive clusters of galaxies there are also galaxy groups, which sometimes contain only a few luminous galaxies. In fact, the number density of groups is far larger than that of clusters. Our Milky Way is part of such a group, the Local Group, which also contains M31 (Andromeda), a second luminous spiral galaxy besides the Milky Way, as well as some far less luminous galaxies such as the Magellanic Clouds. Some groups of galaxies are very compact, i.e., their galaxies are confined within a very small volume (Fig. 1.20). Interactions between these galaxies cause the lifetimes of many such groups to be much smaller than the age of the Universe, and the galaxies in such groups will merge.

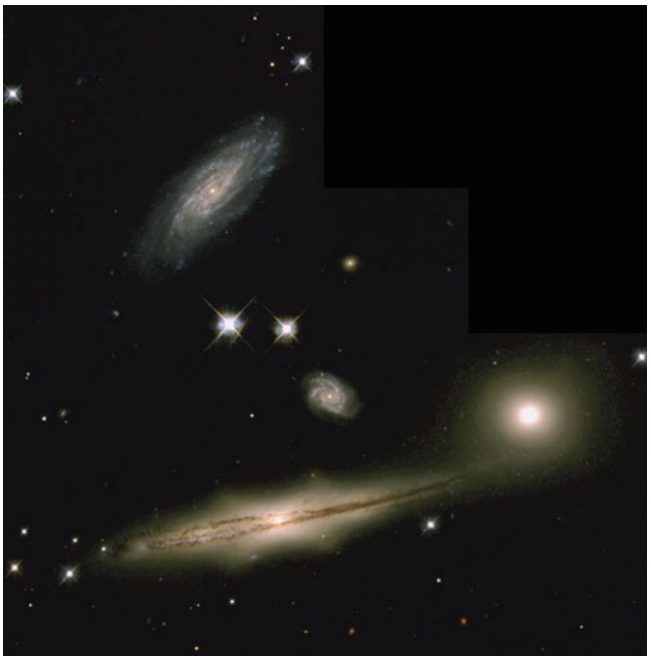


Fig. 1.20 The galaxy group HCG87 belongs to the class of so-called compact groups. In this HST image we can see three massive galaxies belonging to this group: an edge-on spiral in the *lower part* of the image, an elliptical galaxy to the *lower right*, and another spiral in the *upper part*. The small spiral in the center is a background object and therefore does not belong to the group. The two lower galaxies have an active galactic nucleus, whereas the upper spiral seems to be undergoing a phase of star formation. The galaxies in this group are so close together that in projection they appear to touch. Between the galaxies, gas streams can be detected. The galaxies are disturbing each other, which could be the cause of the nuclear activity and star formation. The galaxies are bound in a common gravitational potential and will heavily interfere and presumably merge on a cosmologically small time scale, which means in only a few orbits, with an orbit taking about 10^8 yr. Such merging processes are of utmost importance for the evolution of the galaxy population. Credit: STScI and the Hubble Heritage Project

1.2.6 World models and the thermal history of the Universe

Quasars, clusters of galaxies, and nowadays even single galaxies are also found at very high redshifts where the simple form of the Hubble law (1.2) is no longer valid. It is therefore necessary to generalize the distance-redshift relation. This requires considering world models as a whole, which are also called cosmological models. The dominant force in the Universe is gravitation. On the one hand, weak and strong interactions both have an extremely small (sub-atomic) range, and on the other hand, electromagnetic interactions do not play a role on large scales since the matter in the Universe is on average electrically neutral. Indeed, if it was not, currents would immediately flow to balance net charge densities. The accepted theory of gravitation is the theory of General Relativity (GR), formulated by Albert Einstein in 1915.

Based on the two postulates that (1) our place in the Universe is not special, and thus not distinguished from other locations and that (2) the distribution of matter around us is isotropic, at least on large scales, one can construct homogeneous and isotropic world models (so-called Friedmann–Lemaître models) that obey the laws of General Relativity. Expanding world models that contain the Hubble expansion result from this theory naturally. Essentially, these models are characterized by three parameters:

- the current expansion rate of the Universe, i.e., the Hubble constant H_0;
- the current mean matter density of the Universe $\rho_{\rm m}$, often parametrized by the dimensionless *density parameter* of matter,

$$\Omega_{\rm m} = \frac{8\pi G}{3H_0^2}\rho_{\rm m}\,; \tag{1.10}$$

- and the density of the so-called vacuum energy, described by the cosmological constant Λ or by the corresponding density parameter of the vacuum

$$\Omega_\Lambda = \frac{\Lambda}{3H_0^2}\,. \tag{1.11}$$

The cosmological constant was originally introduced by Einstein to allow stationary world models within GR. After the discovery of the Hubble expansion he is quoted to have called the introduction of Λ into his equations his

greatest blunder. In quantum mechanics Λ attains a different interpretation, related to an energy density of the vacuum.

The values of the cosmological parameters are known quite accurately today (see Chap. 8), with values of $\Omega_{\rm m} \approx 0.3$ and $\Omega_\Lambda \approx 0.7$. The discovery of a non-vanishing Ω_Λ came completely unexpectedly. To date, all attempts have failed to compute a reasonable value for Ω_Λ from quantum mechanics. By that we mean a value which has the same order-of-magnitude as the one we derive from cosmological observations. In fact, simple and plausible estimates lead to a value of Λ that is $\sim 10^{120}$ times larger than that obtained from observation, a tremendously bad estimate indeed. This huge discrepancy is probably one of the biggest challenges in fundamental physics today.

According to the Friedmann–Lemaître models, the Universe used to be smaller and hotter in the past, and it has continuously cooled down in the course of expansion. We are able to trace back the cosmic expansion under the assumption that the known laws of physics were also valid in the past. From that we get the Big Bang model of the Universe, according to which our Universe has evolved out of a very dense and very hot state, the so-called *Big Bang*. The Big Bang marks the beginning of the Universe, at least as far as physics can describe it, and is taken to be the origin of cosmic time. This world model makes a number of predictions that have been verified convincingly:

1. About 1/4 of the baryonic matter in the Universe should consist of helium which formed about 3 min after the Big Bang, while most of the rest consists of hydrogen. This is indeed the case: the mass fraction of helium in metal-poor objects, whose chemical composition has not been significantly modified by processes of stellar evolution, is about 24 %.
2. From the exact fraction of helium one can derive the number of neutrino families—the more neutrino species that exist, the larger the fraction of helium will be. From this, it was derived in 1981 that there are three kinds of neutrinos. This result was later confirmed by particle accelerator experiments.
3. Thermal radiation from the hot early phase of the Universe should still be measurable today. Predicted in 1946 by George Gamow, it was discovered by Arno Penzias and Robert Wilson in 1965. The corresponding photons have propagated freely after the Universe cooled down to about 3000 K and the plasma constituents—atomic nuclei and electrons—combined to neutral atoms, an epoch called *recombination*. As a result of cosmic expansion, this radiation has cooled down to about $T_0 \approx 2.73$ K. This microwave radiation is observed to be nearly perfectly isotropic, once we subtract the radiation which is emitted locally by the Milky Way (see Fig. 1.21). Indeed, measurements from the COBE satellite showed that the cosmic microwave background (CMB) is the most accurate blackbody spectrum ever measured.
4. Today's structures in the Universe have evolved out of very small density fluctuations in the early cosmos. The seeds of structure formation must have already been present in the very early phases of cosmic evolution. These density fluctuations should also be visible as small

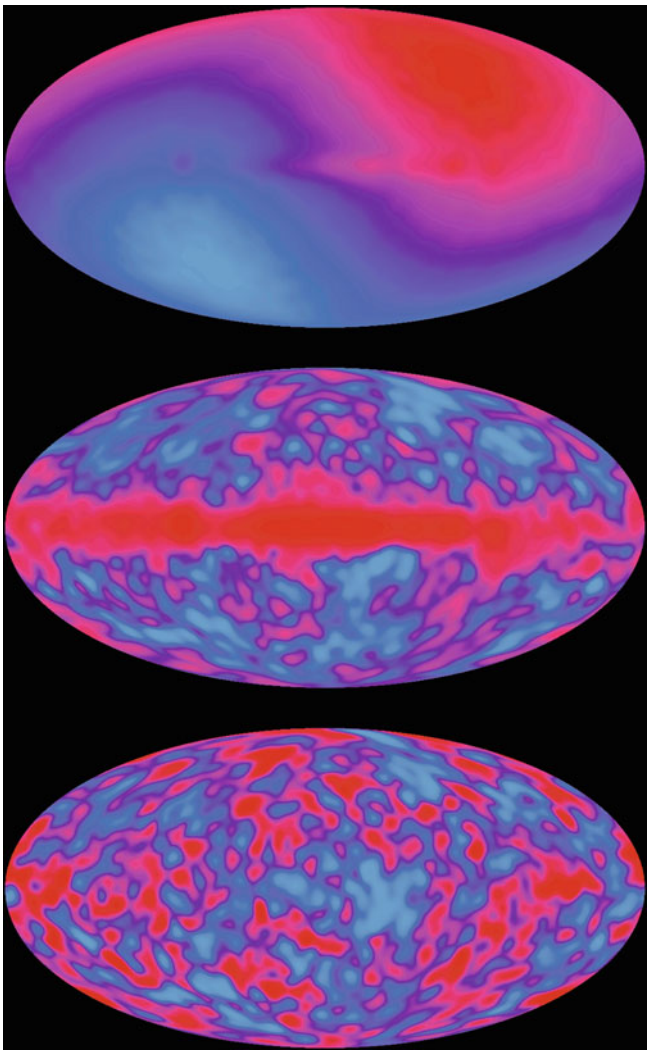


Fig. 1.21 Temperature distribution of the cosmic microwave background on the sky as measured by the COBE satellite. The *top image* shows a dipole distribution; it originates from the Earth's motion relative to the restframe of the CMB. Our Solar System moves at a speed of 369 km/s relative to that system, which leads to a dipole anisotropy with an amplitude of $\Delta T/T \sim v/c \sim 1.2 \times 10^{-3}$ due to the Doppler effect. If this dipole contribution is subtracted, we get the *map in the middle* which clearly shows the emission from the Galactic disk. Since this emission has a different spectral energy distribution (it is not a blackbody of $T \sim 3$ K), it can also be subtracted to get the temperature *map at the bottom*. These are the primordial fluctuations of the CMB, with an amplitude of about $\Delta T/T \sim 2 \times 10^{-5}$. Credit: COBE/DRM team, NASA

temperature fluctuations in the microwave background emitted about 380 000 years after the Big Bang at the epoch of recombination. In fact, COBE was the first to observe these predicted anisotropies (see Fig. 1.21). Later experiments, especially the WMAP and Planck satellites, observed the structure of the microwave background at much improved angular resolution and verified the theory of structure formation in the Universe in detail (see Sect. 8.6).

With these predictions so impressively confirmed, in this book we will exclusively consider this cosmological model; currently there is no competing model of the Universe that could explain these very basic cosmological observations in a natural way. In addition, this model does not seem to contradict any fundamental observation in cosmology. However, as the existence of a non-vanishing vacuum energy density shows, together with a matter density $\rho_{\rm m}$ that is about six times the mean baryon density in the Universe (which can be derived from the abundance of the chemical elements formed in the Big Bang), the physical nature of about 95 % of the content of our Universe is not yet understood.

Most of the CMB photons we receive today had their last physical interaction with matter when the Universe was about 3.8×10^5 yr old. Also, the most distant galaxies and quasars known today (at $z \sim 7$) are strikingly young—we see them at a time when the Universe was less than a tenth of its current age. The exact relation between the age of the Universe at the time of the light emission and the redshift depends on the cosmological parameters H_0, $\Omega_{\rm m}$, and Ω_Λ. In the special case that $\Omega_{\rm m} = 1$ and $\Omega_\Lambda = 0$, called the *Einstein–de Sitter model*, one obtains

$$t(z) = \frac{2}{3H_0} \frac{1}{(1+z)^{3/2}} \,. \tag{1.12}$$

In particular, the age of the Universe today (i.e., at $z = 0$) in this model is

$$t_0 = \frac{2}{3H_0} \approx 6.5 \times 10^9 \, h^{-1} \mathrm{yr} \,. \tag{1.13}$$

The Einstein–de Sitter (EdS) model is the simplest world model and we will sometimes use it as a reference, despite the fact the our Universe does not follow the EdS model, since $\Omega_{\rm m} < 1$ and $\Omega_\Lambda > 0$. However, due to its mathematical simplicity, it is often convenient to obtain rough estimates within this model. The mean density of the Universe in the EdS model is

$$\rho_0 = \rho_{\rm cr} \equiv \frac{3H_0^2}{8\pi G} \approx 1.9 \times 10^{-29} h^2 \, \mathrm{g\,cm^{-3}} \,, \tag{1.14}$$

hence it is really, really small.

1.2.7 Structure formation and galaxy evolution

The low amplitude of the CMB anisotropies implies that the inhomogeneities must have been very small at the epoch of recombination, whereas today's Universe features very large density fluctuations, at least on scales of clusters of galaxies. Hence, the density field of the cosmic matter must have evolved. This structure evolution occurs because of gravitational instability, in that an overdense region will expand more slowly than the mean Universe due to its self-gravity. Therefore, any relative overdensity becomes amplified in time. The growth of density fluctuations in time will then cause the formation of large-scale structures, and the gravitational instability is also responsible for the formation of galaxies and clusters. Our world model sketched above predicts the abundance of galaxy clusters as a function of redshift, which can be compared with the observed cluster counts. This comparison can then be used to determine cosmological parameters.

Another essential conclusion from the smallness of the CMB anisotropies is the existence of dark matter on cosmic scales. The major fraction of cosmic matter is dark matter. The baryonic contribution to the matter density is $\lesssim$20 % and to the total energy density $\lesssim$5 %. The energy density of the Universe is dominated by the vacuum energy.

Unfortunately, the spatial distribution of dark matter on large scales is not directly observable. We only observe galaxies or, more precisely, their stars and gas. One might expect that galaxies would be located preferentially where the dark matter density is high. However, it is by no means clear that local fluctuations of the galaxy number density are strictly proportional to the density fluctuations of dark matter. The relation between the dark and luminous matter distributions is currently only approximately understood.

Eventually, this relation has to result from a detailed understanding of galaxy formation and evolution. Locations with a high density of dark matter can support the formation of galaxies. Thus we will have to examine how galaxies form and why there are different kinds of galaxies. In other words, what decides whether a forming galaxy will become an elliptical or a spiral? This question has not been definitively answered yet, but it is supposed that ellipticals can form only by the merging of galaxies. Indeed, the standard model of the Universe predicts that small galaxies will form first; larger galaxies will be formed later through the ongoing merger of smaller ones.

The evolution of galaxies can actually be observed directly. Galaxies at high redshift (i.e., cosmologically young galaxies) are in general smaller and bluer, and the star-formation rate was significantly higher at earlier times than it is today. The change in the mean color of galaxies as

Fig. 1.22 Galaxy evolution caught in the act: The Tadpole galaxy (also called Arp 188) shows a $\sim$ 90 kpc long tail. Most likely, this spiral galaxy collided with a smaller galaxy some time in the past, which ripped part of its matter away from the main body of the galaxy. Inside the tail, clusters of newly formed stars are visible. Clearly, this galaxy will have changed from its earlier state after it becomes settled again. Credit: H. Ford (JHU), M. Clampin (STScI), G. Hartig (STScI), G. Illingworth (UCO/Lick), ACS Science Team, ESA, NASA

a function of redshift can be understood as a combination of changes in the star formation processes and an aging of the stellar population. Also, galaxy collisions which can be directly observed in the local Universe (Fig. 1.22; see also Fig. 1.16) have a strong impact on individual galaxies and need to be considered in models of galaxy evolution.

1.2.8 Cosmology as a triumph of the human mind

Cosmology, extragalactic astronomy, and astrophysics as a whole are a heroic undertaking of the human mind and a triumph of physics. To understand the Universe we apply physical laws that were found empirically under completely different circumstances. All the known laws of physics were derived 'today' and, except for General Relativity, are based on experiments on a laboratory scale or, at most, on observations in the Solar System, such as Kepler's laws which formed the foundation for the Newtonian theory of gravitation. Is there any a priori reason to assume that these laws are also valid in other regions of the Universe or at completely different times? However, this is apparently indeed the case: nuclear reactions in the early Universe seem to obey the same laws of strong interaction that are measured today in our laboratories, since otherwise the agreement of the prediction of a 25 % helium mass fraction from nuclear reactions in the first minutes of our Universe with the observed helium abundance would be a pure coincidence. Quantum mechanics, describing the wavelengths of atomic transitions, also seems to be valid at very large distances—since even the most distant objects show emission lines in their spectra with frequency ratios (which are described by the laws of quantum mechanics) identical to those in nearby objects. In fact, cosmologists can put very tight upper limits on a possible variation of the 'constants' of nature with time, such as the fine-structure constant or the electron to proton mass ratio.

By far the greatest achievement is General Relativity. It was originally formulated by Albert Einstein since his special theory of relativity did not allow him to incorporate the laws of Newtonian gravity. No empirical findings were known at that time (1915) which would not have been explained by the Newtonian theory of gravity. Nevertheless, Einstein developed a totally new theory of gravitation for purely theoretical reasons. The first success of this theory was the correct description of the gravitational deflection of light by the Sun, measured in 1919, and of the perihelion rotation of Mercury.[9] His theory permits a description of the expanding Universe, which became necessary after Hubble's discovery in 1928. Only with the help of this theory can we reconstruct the history of the Universe back into the past. Today this history seems to be well understood up to the time when the Universe was about 10^{-6} s old and had a temperature of about 10^{13} K. Particle physics models allow an extrapolation to even earlier epochs.

The cosmological predictions discussed above are based on General Relativity describing an expanding Universe, therefore providing a test of Einstein's theory. On the other hand, General Relativity also describes much smaller systems and with much stronger gravitational fields, such as neutron stars and black holes. With the discovery of a binary system consisting of two neutron stars, the binary pulsar PSR 1913+16, in the last $\sim$ 40 yr very accurate tests of General Relativity have become possible. For example, the observed perihelion rotation in this binary system and the shrinking of the binary orbit over time due to the radiation of energy by gravitational waves is extremely accurately described by General Relativity. Together, General Relativity has been successfully tested on length-scales from 10^{11}cm (the characteristic scale of the binary pulsar) to 10^{28}cm (the size of the visible Universe), that is over more than 10^{17} orders of magnitude—an impressive result indeed!

1.2.9 Astrophysics & Physics

Exploring the laws of gravity by astronomical observations in the Solar System and beyond is just one example for

[9] This was already known in 1915, but it was not clear whether it might not have any other explanation, e.g., a quadrupole moment of the mass distribution of the Sun.

the close connection between physics and astronomy. As the word 'astro*physics*' implies, the science of the Universe has become an integral part of physics. Astrophysics and cosmology not only apply the laws of physics to interpret and understand the cosmic objects like stars, galaxies and black holes, but also have led to discoveries concerning fundamental physics. A few examples should illustrate that point.

- Measuring the flux of neutrinos from the Sun, which are produced by nuclear fusion in the Solar center to generate the Solar luminosity, it was found that the observed neutrino rate is only half of what was expected. The solution of this Solar neutrino problem led to the discovery of neutrino oscillations, disclosing a fundamental property of these weakly interacting particles, requiring them to have a finite rest mass (see Sect. 4.4.6), and providing clear evidence for the incompleteness of the Standard Model of elementary particles.
- The fact that carbon has a large abundance in the Universe led Frey Hoyle in 1952 to suspect that there should be a particular excited state of this nucleus in resonance with energy levels of beryllium and helium, to enable the formation of carbon through nuclear fusion in stellar interiors. This previously unknown state, essential for the occurrence of the so-called triple-alpha process, was later on found in experiments.
- Much of the development in the field of plasma physics was driven by astrophysicists in order to understand the behavior of plasmas in cosmic objects—ranging from the Sun (e.g., the occurrence of sunspots, the Solar corona) to the interplanetary space filled with the Solar wind and its interaction with the Earth's magnetosphere, to the impact of magnetic fields on star-forming regions, to the formation of relativistic jets in active galactic nuclei.

Arguably the largest impact astrophysics has on other branches of physics today is related to the finding that the Universe is dominated by dark matter and dark energy—there is no evidence for the existence of these new forms of energy apart from astronomical observations. We will come back to that theme repeatedly in the course of this book.

However, there is one important difference between astrophysics and other branches of physics: we cannot do experiments with the subjects of interest, we cannot prepare the system is a way which allows a clean measurement under controlled external conditions, and repeat the measurement with changing conditions. We can only observe how our objects behave, how different objects of the same kind are similar—or different—in their behavior, and draw conclusions from it.

1.3 The tools of extragalactic astronomy

Extragalactic sources—galaxies, quasars, clusters of galaxies—are at large distances. This means that in general they appear to be faint even if they are intrinsically luminous. They are also seen to have a very small angular size despite their possibly large linear extent. In fact, just three extragalactic sources are visible to the naked eye: the Andromeda galaxy (M31) and the Large and Small Magellanic Clouds. Thus for extragalactic astronomy, telescopes are needed that have large apertures (photon collecting area) and a high angular resolution. This applies to all wavebands, from radio astronomy to gamma ray astronomy.

The properties of astronomical telescopes and their instruments can be judged by different criteria, and we will briefly describe the most important ones. The *sensitivity* specifies how dim a source can be and still be observable in a given integration time. The sensitivity depends on the aperture of the telescope as well as on the efficiency of the instrument and the sensitivity of the detector. The sensitivity of optical telescopes, for instance, was increased by a large factor when CCDs replaced photographic plates as detectors in the early 1980s. The sensitivity also depends on the sky background, i.e., the brightness of the sky caused by non-astronomical sources. Artificial light in inhabited regions has forced optical telescopes to retreat into more and more remote areas of the world where *light pollution* is minimized. Radio astronomers have similar problems caused by radio emission from the telecommunication infrastructure of modern civilization. The *angular resolution* of a telescope specifies down to which angular separation two sources in the sky can still be separated by the detector. For diffraction-limited observations like those made with radio telescopes or space-born telescopes, the angular resolution $\Delta\theta$ is limited by the diameter D of the telescope. For a wavelength λ one has $\Delta\theta = \lambda/D$. For optical and near-infrared observations from the ground, the angular resolution is in general limited by turbulence in the atmosphere, which explains the choice of high mountain tops as sites for optical telescopes. These atmospheric turbulences cause, due to scintillation, the smearing of the images of astronomical sources, an effect that is called *seeing*. In interferometry, where one combines radiation detected by several telescopes, the angular resolution is limited by the spatial separation of the telescopes. The *spectral resolution* of an instrument specifies its capability to separate different wavelengths. The *throughput* of a telescope/instrument system is of particular importance in large sky surveys. For instance, the efficiency of spectroscopic surveys depends, in addition to the aperture of the telescope, on the number of spectra that can be observed simultaneously. Special multiplex spectrographs have been constructed for such tasks. Likewise, the efficiency of photometric surveys depends on the telescope's diameter and the region of sky that can be observed simultaneously, i.e., the field-of-view of the camera. Finally, the efficiency of observations also depends on factors like the number of clear nights at an astronomical site, the fraction of an observing night in which actual science data is taken, the fraction of

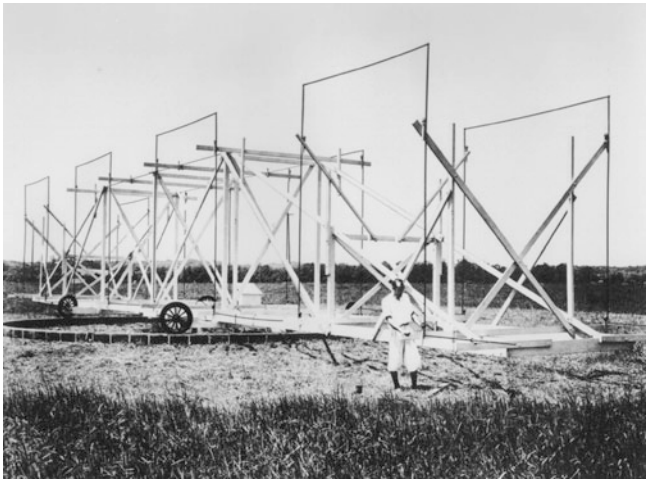


Fig. 1.23 "Jansky's Merry-Go-Round". By turning the structure in an azimuthal direction, a rough estimate of the position of radio sources could be obtained. Credit: NRAO/AUI

time an instrument cannot be used due to technical problems, the stability of the instrumental set-up (which determines the time required for calibration measurements), and many other such aspects.

In the rest of this section some telescopes will be presented that are of special relevance to extragalactic astronomy and to which we will frequently refer throughout the course of this book.

1.3.1 Radio telescopes

With the exception of optical wavelengths, the Earth's atmosphere is transparent only for very large wavelengths—radio waves. The radio window of the atmosphere is cut off towards lower frequencies, at about $\nu \sim 10\,\mathrm{MHz}$, because radiation of a wavelength larger than $\lambda \sim 30\,\mathrm{m}$ is reflected by the Earth's ionosphere and therefore cannot reach the ground. Below $\lambda \sim 5\,\mathrm{mm}$ radiation is increasingly absorbed by oxygen and water vapor in the atmosphere, and below about $\lambda \sim 0.3\,\mathrm{mm}$ ground-based observations are no longer possible.

Mankind became aware of cosmic radio radiation—in the early 1930s—only when noise in radio antennas was found that would not vanish, no matter how quiet the device was made. In order to identify the source of this noise the AT&T Bell Labs hired Karl Jansky, who constructed a movable antenna called "Jansky's Merry-Go-Round" (Fig. 1.23). After some months Jansky had identified, besides thunderstorms, one source of interference that rose and set every day. However, it did not follow the course of the Sun which was originally suspected to be the source. Rather, it followed the stars. Jansky finally discovered that the signal originated from the direction of the center of the Milky Way. He published his result in 1933, but this publication also marked the end of his career as the world's first radio astronomer.

Inspired by Jansky's discovery, Grote Reber was the first to carry out real astronomy with radio waves. When AT&T refused to employ him, he built his own radio "dish" in his garden, with a diameter of nearly 10 m. Between 1938 and 1943, Reber compiled the first sky maps in the radio domain. Besides strong radiation from the center of the Milky Way he also identified sources in Cygnus and in Cassiopeia. Through Reber's research and publications radio astronomy became an accepted field of science after World War II.

The largest single-dish radio telescope is the Arecibo telescope, shown in Fig. 1.24. Due to its enormous area, and thus high sensitivity, this telescope, among other achievements, detected the first pulsar in a binary system, which is used as an important test laboratory for General Relativity (see Sect. 7.9). Also, the first extra-solar planet, in orbit around a pulsar, was discovered with the Arecibo telescope. For extragalactic astronomy Arecibo plays an important role in measuring the redshifts and line widths of spiral galaxies, both determined from the 21 cm emission line of neutral hydrogen (see Sect. 3.4).

The Effelsberg 100 m radio telescope of the Max-Planck-Institut für Radioastronomie was, for many years, the world's largest fully steerable radio telescope, but since 2000 this title has been claimed by the new Green Bank Telescope (see Fig. 1.25) after the old one collapsed in 1988. With Effelsberg, for example, star formation regions can be investigated. Using molecular line spectroscopy, one can measure their densities and temperatures. Magnetic fields also play a role in star formation, though many details still need to be clarified. By measuring the polarized radio flux, Effelsberg has mapped the magnetic fields of numerous spiral galaxies. It is also used to map the neutral hydrogen distribution in the Galaxy, its neighborhood and galaxies in the nearby Universe, as well as for pulsar research. In addition, due to its huge collecting area Effelsberg plays an important role in interferometry at very long baselines (see below).

Because of the long wavelength, the angular resolution of even large radio telescopes is fairly low, compared to optical telescopes. For this reason, radio astronomers soon began utilizing interferometric methods, where the signals obtained by several telescopes are correlated to get an interference pattern. One can then reconstruct the structure of the source from this pattern using Fourier transformation. With this method one gets the same angular resolution (though, of course, not the same sensitivity) as one would achieve with a single telescope of a diameter corresponding to the maximum pair separation of the individual telescopes used.

Following the first interferometric measurements in England (around 1960) and the construction of the large Westerbork Synthesis Radio Telescope in the Netherlands (around 1970), at the end of the 1970s the Very Large Array (VLA)

Fig. 1.24 With a diameter of 305 m, the Arecibo telescope in Puerto Rico is the largest single-dish telescope in the world; it may also be known from the James Bond movie "GoldenEye". The disadvantage of its construction is its lack of steerability. Tracking of sources is only possible within narrow limits by moving the secondary mirror. Credit: Courtesy of the NAIC-Arecibo Observatory, a facility of the NSF

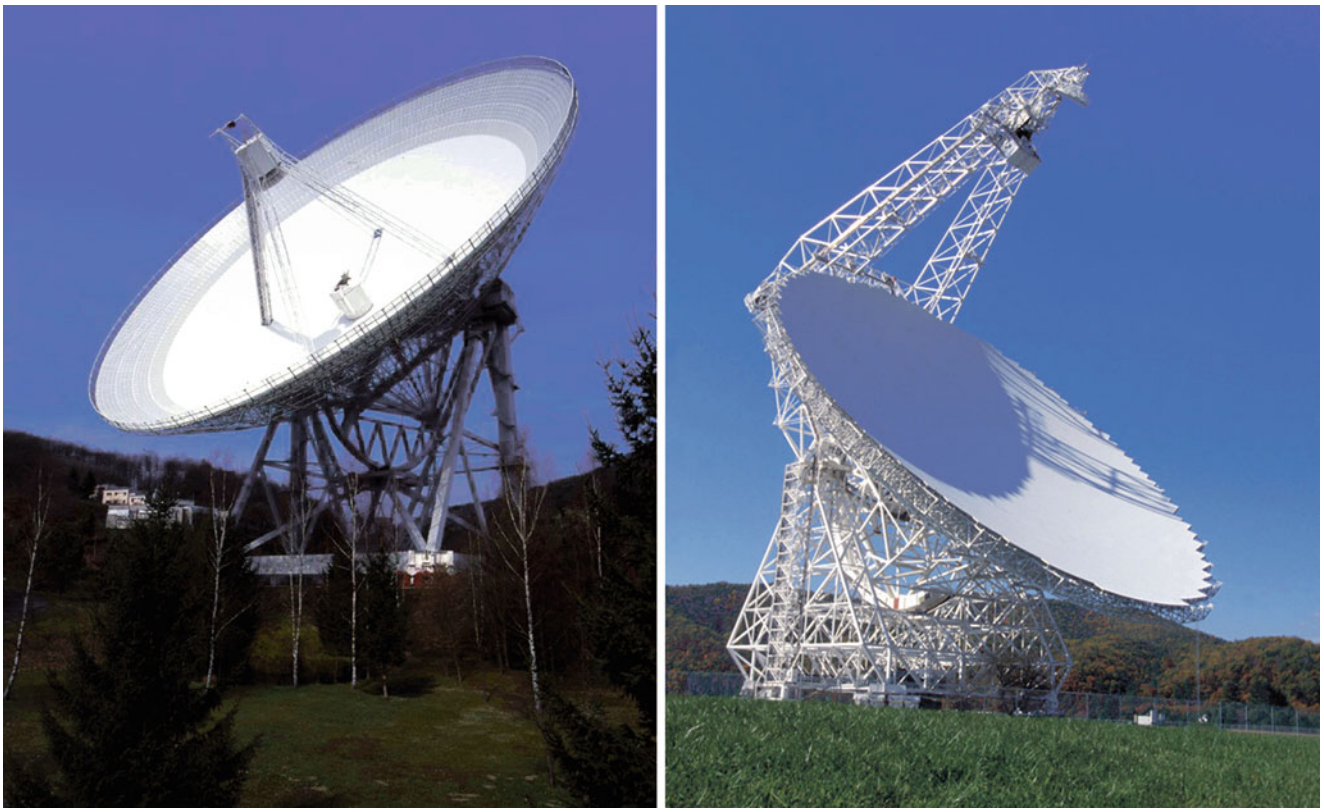


Fig. 1.25 The world's two largest fully steerable radio telescopes. *Left*: The 100 m telescope in Effelsberg. It was commissioned in 1972 and is used in the wavelength range from 3.5 mm to 35 cm. Eighteen different detector systems are necessary for this. *Right*: The Green Bank Telescope. It does not have a rotationally symmetric mirror; one axis has a diameter of 100 m and the other 110 m. Credit: *Left*: Max-Planck-Institut für Radioastronomie. *Right*: NRAO/AUI

in New Mexico (see Fig. 1.26) began operating. With the VLA one achieved an angular resolution in the radio domain comparable to that of optical telescopes at that time. For the first time, this allowed the combination of radio and optical images with the same resolution and thus the study of cosmic sources over a range of several clearly separated wavelength regimes. With the advent of the VLA radio astronomy experienced an enormous breakthrough, particularly in the study

Fig. 1.26 The Very Large Array (VLA) in New Mexico consists of 27 antennas with a diameter of 25 m each that can be moved on rails. It is used in four different configurations that vary in the separation of the telescopes; switching configurations takes about 2 weeks. Credit: NRAO/AUI

of AGNs. It became possible to examine the large extended jets of quasars and radio galaxies in detail (see Sect. 5.1.2). Other radio interferometers must also be mentioned here, such as the British MERLIN (Multi-Element Radio Linked Interferometer Network), where seven telescopes with a maximum separation of 230 km are combined.

The VLA recently underwent a major upgrade, in particular by installing new receivers. By increasing their bandwidth relative to the older ones, and installing state-of-the-art electronics, the sensitivity of the eVLA is about ten times higher than the 'old' VLA. Together with the introduction of two new frequency bands, the eVLA now covers the full frequency range from 1 to 50 GHz.

In the radio domain it is also possible to interconnect completely independent and diverse antennas to form an interferometer, since one can record the amplitude and phases of the electromagnetic radiation. For example, in Very Long Baseline Interferometry (VLBI) radio telescopes on different continents are used simultaneously. These frequently also include Effelsberg and the VLA. In 1995 a system of ten identical 25 m antennas was set up in the USA, exclusively to be used in VLBI, the Very Long Baseline Array (VLBA). Angular resolutions of better than a milliarcsecond (mas) can be achieved with VLBI. Therefore, in extragalactic astronomy VLBI is particularly used in the study of AGNs. With VLBI we have learned a great deal about the central regions of AGNs, such as the occurrence of apparent superluminal velocities in these sources. A further increase in angular resolution with VLBI will be obtained with the Russian RadioAstron mission, launched in 2011; it is a 10-m radio telescope on a highly eccentric orbit, reaching distances from the Earth up to almost 400 000 km.

Some of the radio telescopes described above are also capable of observing in the millimeter regime. For shorter wavelengths the surfaces of the antennas are typically too coarse, so that special telescopes are needed for wavelengths of 1 mm and below. The 30 m telescope on Pico Veleta (Fig. 1.27), with its exact surface shape, allows observations in the millimeter range. It is particularly used for molecular spectroscopy at these frequencies. Furthermore, important observations of high-redshift galaxies at 1.2 mm have been made with this telescope using the bolometer camera MAMBO (Max-Planck Millimeter Bolometer). Similar observations are also conducted with the SCUBA (Submillimeter Common-User Bolometer Array) camera at the James Clerk Maxwell Telescope (JCMT; Fig. 1.28) on Mauna Kea, Hawaii, which observes at wavelengths between 3 and 0.3 mm. With the SCUBA-camera, operating at 850 μm (0.85 mm), we can observe star-formation regions in distant galaxies for which the optical emission is nearly completely absorbed by dust in these sources. These dusty star-forming galaxies can be observed in the (sub-)millimeter regime of the electromagnetic spectrum even out to large redshifts, as will be discussed in Sect. 9.3.3. Recently, the SCUBA-2 camera replaced the original one; its much larger field-of-view (10^4 pixels vs. the 37 of SCUBA) enhances its survey capability by about a factor of 1000.

An even better site for (sub-)mm astronomy than Mauna Kea is the Cerro Chajnantor, a 5100 m altitude plateau in the Chilean Atacama desert, due to the smaller column of water vapor. Since 2005, the Atacama Pathfinder Experiment (APEX) operates there. It is a 12 m telescope (Fig. 1.28), equipped with several highly sensitive instruments. One of them is a bolometer array specifically designed to

Fig. 1.27 The 30 m telescope on Pico Veleta was designed for observations in the millimeter range of the spectrum. This telescope, like all millimeter telescopes, is located on a mountain to minimize the column density of water in the atmosphere. Credit: MPIfR, IRAM

Fig. 1.28 *Left*: The James Clerk Maxwell Telescope (JCMT) on Mauna Kea has a 15 m dish. It is protected by the largest single piece of Gore-Tex, which has a transmissivity of 97 % at sub-millimeter wavelengths. *Right*: The Atacama Pathfinder Experiment (APEX) 12-m sub-millimeter telescope has been in operation since 2005. It is located at $\sim$ 5000 m altitude on the Chajnantor plateau in the Atacama Desert in Chile, the same location as that of the ALMA observatory. APEX observes at wavelengths between 200 μm and 1.5 mm. Credit: *Left*: Joint Astronomy Center. *Right*: ESO/H.H. Heyer

observe the Sunyaev–Zeldovich effect in galaxy clusters (see Sect. 6.4.4).

The site also hosts the ALMA (Atacama Large Millimeter/sub-millimeter Array) observatory, one of the most ambitious projects of ground-based astronomy yet (Fig. 1.29). ALMA consists of 50 antennas with 12 m diameter each, which can be moved around to change the separation between the telescopes (up to 16 km), i.e., the baselines for interferometry. In addition, it has a compact array consisting of twelve 7-m and four 12-m antennas. ALMA will provide a giant jump in the capabilities of sub-millimeter astronomy, due to the large collecting area, the large baselines, as well as the sensitivity and bandwidth of the receivers. The construction of ALMA, which observes at wavelength longer than 300 μm, was completed in 2013, but even with the incomplete array, observations were conducted, showing the impressive capabilities of this new observatory. ALMA is the result of a global cooperation between North America, Europe and East Asia, together with the host country Chile.

To measure the tiny temperature fluctuations of the cosmic microwave background radiation one needs extremely stable observing conditions and low-noise detectors. In order to avoid the thermal radiation of the atmosphere as much as possible, balloons and satellites were constructed to operate instruments at very high altitude or in space. The American

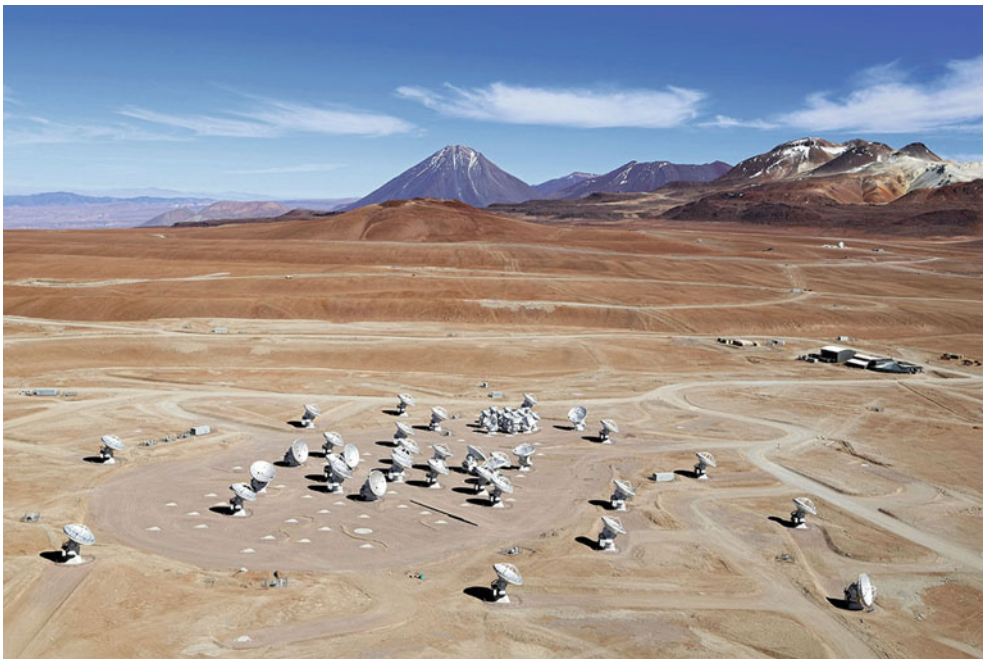


Fig. 1.29 The Atacama Large Millimeter/sub-millimeter Array (ALMA) on the Chajnantor Plateau, located at an altitude of 5000 m in the Chilean Andes. This photo was taken in December 2012, 4 months prior to the ALMA inauguration. ALMA consists of 50 12-m antennas which can be reconfigurated, to yield different baselines for interferometry, and an additional compact array of 16 antennas with diameter of 7 and 12 m. Credit: Clem & Adri Bacri-Normier (wingsforscience.com)/ESO

COBE (Cosmic Background Explorer) satellite measured the anisotropies of the CMB for the first time, at wavelengths of a few millimeters. In addition, the frequency spectrum of the CMB was precisely measured with instruments on COBE. The WMAP (Wilkinson Microwave Anisotropy Probe) satellite obtained, like COBE, a map of the full sky in the microwave regime, but at a significantly improved angular resolution and sensitivity. The first results from WMAP, published in February 2003, were an enormously important milestone for cosmology, as will be discussed in Sect. 8.6.5. WMAP observed for a total of 9 years, with a final data release at the end of 2012. The new European Planck satellite was launched in May 2009, together with the Herschel satellite. It has a much larger frequency coverage than WMAP, from 30 to 850 GHz (thus covering both sides of the peak of the CMB spectrum), a higher sensitivity, and a better angular resolution. First cosmological results from Planck were released in March 2013 (see Sect. 8.6.6). Besides observing the CMB these missions (see Fig. 1.30) are also of great importance for millimeter astronomy; these satellites not only measure the cosmic background radiation but of course also the microwave radiation of the Milky Way and of other galaxies.

Beside space, the Antarctica offers excellent observing conditions in the millimeter regime; the very low temperatures and high altitude (the Antarctic Plateau lies at an altitude of 2800 m) yields a particularly low column density of water vapor. Together with the fact that the Sun does not rise and set every day, the atmospheric conditions are also very stable. Therefore, several astronomical projects are conducted, or are being planned, in the Antarctica. One of them is the South Pole Telescope (SPT, see Fig. 1.31), a 10 m telescope with a one square degree field-of-view. Its main scientific activity is a survey of clusters of galaxies through their Sunyaev–Zeldovich effect caused by the hot intracluster gas, as well as studying the CMB fluctuations at small angular scales.

1.3.2 Infrared telescopes

In the wavelength range $1\,\mu\text{m} \lesssim \lambda \lesssim 300\,\mu\text{m}$, observations from the Earth's surface are always subject to very difficult conditions, if they are possible at all. The atmosphere has some windows in the near-infrared (NIR, $1\,\mu\text{m} \lesssim \lambda \lesssim 2.4\,\mu\text{m}$) which render ground-based observations possible. In the mid-infrared (MIR, $2.4\,\mu\text{m} \lesssim \lambda \lesssim 20\,\mu\text{m}$) and far-infrared (FIR, $20\,\mu\text{m} \lesssim \lambda \lesssim 300\,\mu\text{m}$) regimes, observations need to be carried out from outside the atmosphere, i.e., using balloons, high-flying airplanes, or satellites. The instruments have to be cooled to very low temperatures, otherwise their own thermal radiation would outshine any signal.

The first noteworthy observations in the far-infrared were made by the Kuiper Airborne Observatory (KAO), an airplane equipped with a 91 cm mirror which operated at altitudes up to 15 km. However, the breakthrough for IR

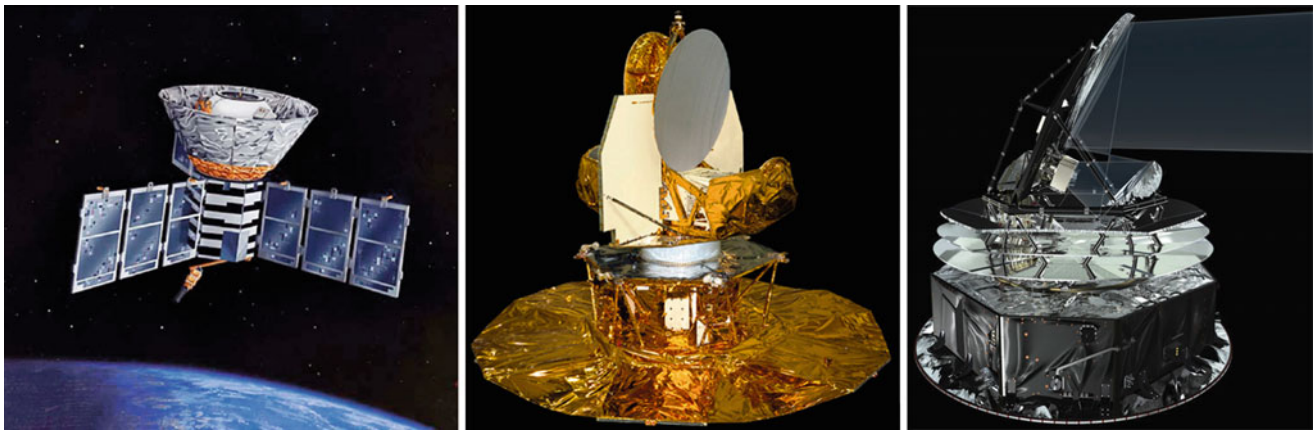


Fig. 1.30 *Left*: Artist's conception of the Cosmic Background Explorer (COBE) spacecraft, launched in 1989 into an Earth orbit, which discovered the temperature fluctuations in the cosmic microwave background. *Middle*: The Wilkinson Microwave Anisotropy Probe (WMAP) satellite was launched in 2001 and observed the microwave sky for 9 years. *Right*: The Planck satellite, launched in 2009 into an orbit at L2 (as for WMAP), has yielded the widest frequency range and highest angular resolution map of the full microwave sky yet. Results from these three satellite missions will be described in Sect. 8.6. Credit: *Left*: NASA/COBE Science Team. *Middle*: NASA/WMAP Science Team. *Right*: ESA (Image by AOES Medialab)

Fig. 1.31 The South Pole Telescope, a 10 m dish located at 2800 m altitude on the Antarctic Plateau. Its off-axis design and shielding minimize the effects from ground spill-over and scattering off the telescope optics. Credit: Glenn Grant, National Science Foundation

astronomy had to wait until the launch of IRAS, the InfraRed Astronomical Satellite (Fig. 1.32). In 1983, with its 60 cm telescope, IRAS compiled the first IR map of the sky at 12, 25, 60, and 100 μm, at an angular resolution of 30″ (2′) at 12 μm (100 μm). It discovered about a quarter of a million point sources as well as about 20 000 extended sources. The positional accuracy for point sources of better than $\sim$ 20″ allowed an identification of these sources at optical wavelengths. Arguably the most important discovery by IRAS was the identification of galaxies which emit the major fraction of their energy in the FIR part of the spectrum. These sources, often called IRAS galaxies, have a very high star-formation rate where the UV light of the young stars is absorbed by dust and then re-emitted as thermal radiation in the FIR. IRAS discovered about 75 000 of these so-called ultra-luminous IR galaxies (ULIRGs).

In contrast to the IRAS mission with its prime task of mapping the full sky, the Infrared Space Observatory ISO (Fig. 1.32) was dedicated to observations of selected objects and sky regions in a wavelength range of 2.5–240 μm. Although the telescope had the same diameter as IRAS, its angular resolution at 12 μm was about a hundred times better than that of IRAS, since the latter was limited by the size of the detector elements. The sensitivity of ISO topped that of

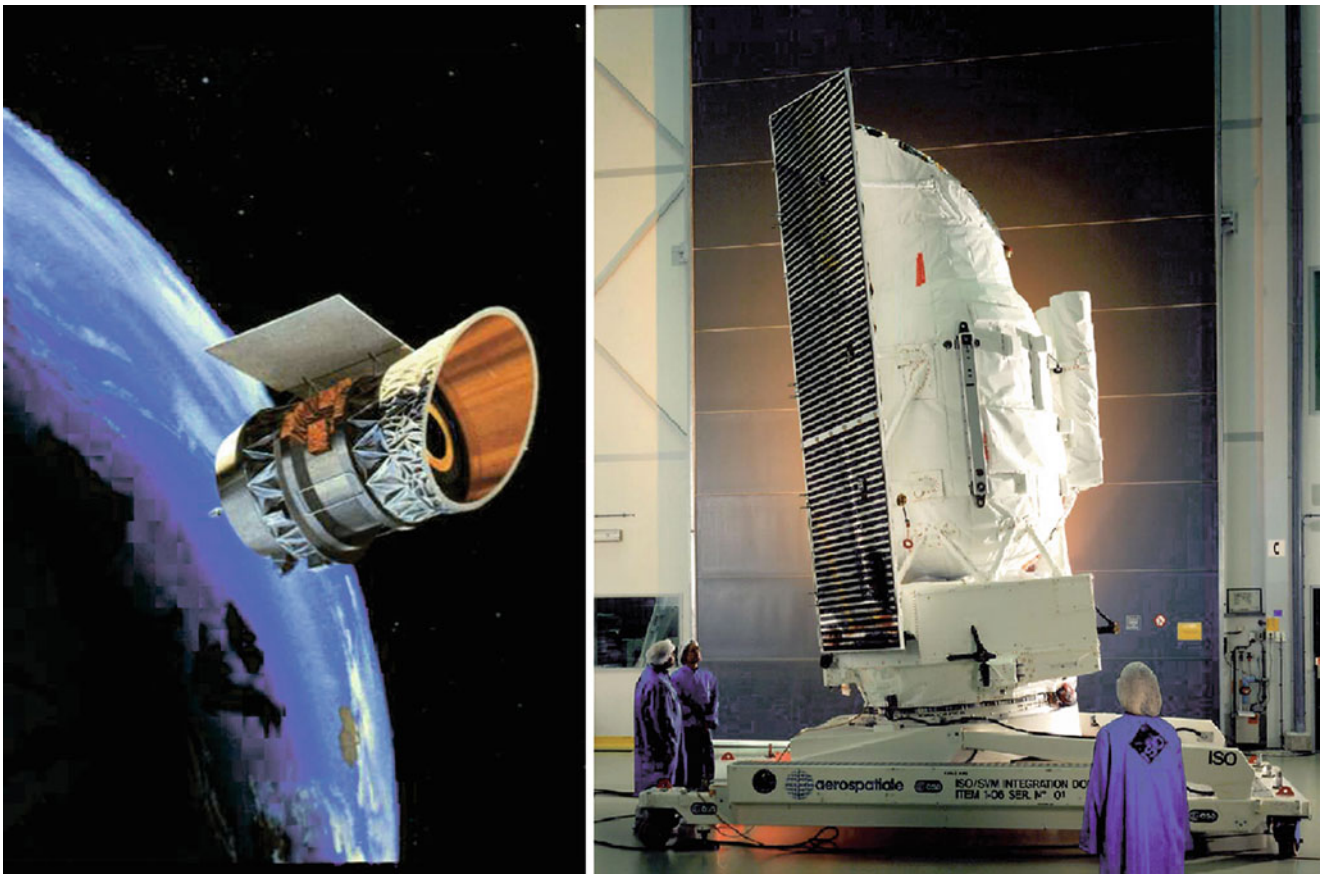


Fig. 1.32 The *left-hand picture* shows an artist's impression of IRAS in orbit. The project was a cooperation of the Netherlands, the USA, and Great Britain. IRAS was launched in 1983 and operated for 10 months; after that the supply of liquid helium, needed to cool the detectors, was exhausted. During this time IRAS scanned 96 % of the sky at four wavelengths. The ISO satellite, shown *on the right*, was an ESA project and observed between 1995 and 1998. Compared to IRAS it covered a larger wavelength range, had a better angular resolution and a thousand times higher sensitivity. Credit: NASA's Infrared Astrophysics Data Center, Caltech/JPL; ESA

IRAS by a factor $\sim$ 1000. ISO carried four instruments: two cameras and two spectrographs. Among the most important results from ISO in the extragalactic domain are the spatially-resolved observations of the dust-enshrouded star formation regions of ULIRGs.

In 2003 a new infrared satellite was launched (the Spitzer Space Telescope, see Fig. 1.33) with capabilities that by far outperform those of ISO. With its 85 cm telescope, Spitzer observes at wavelengths between 3.6 and 160 μm. Its Infrared Array Camera (IRAC) takes images at 3.6, 4.5, 5.8 and 8.0 μm simultaneously, and has a field of view of $5'.2 \times 5'.2$ and 256×256 pixels, significantly more than the 32×32 pixels of ISOCAM on ISO that had a comparable wavelength coverage. The Multiband Imaging Photometer for Spitzer (MIPS) operated at 24, 70 and 160 μm, and the Infrared Spectrograph (IRS) was a spectrometer covering the wavelength regime between 5.3 and 40 μm, with a spectral resolution of about $R = \lambda/\Delta\lambda \sim 100$. In 2009, the helium of the cooler was exhausted, which rendered observations at longer wavelength impossible. Since then, the Spitzer Warm Mission continues to observe in the two short wavelengths of IRAC. Spitzer has made important contributions to all fields of astronomy, including the first direct detection of light from an extrasolar planet. Spitzer has provided information about the thermal dust emission of many nearby and distant galaxies, and thus of their star-formation activity.

In May 2009, the Herschel Space Observatory (Fig. 1.33) was launched, together with the Planck satellite, into an L2 orbit.[10] With its 3.5 m diameter, it was the largest astronomical telescope in space up to then. In total, the satellite

[10] In the Earth-Sun system, there are five points—called Lagrange points, where the total force, i.e., the sum of the gravitational forces of Sun plus Earth and the centrifugal force all balance to zero. Objects located at these Lagrangian point will thus orbit at the same angular velocity around the Sun as the Earth. The Lagrange point L2 is located on the line connecting Sun and Earth, about 1.5×10^6 km outside the Earth orbit. A satellite located there always sees Sun and Earth in the same direction and can thus be shielded from their radiation. The L2 point is therefore a preferred location for astronomical satellites.

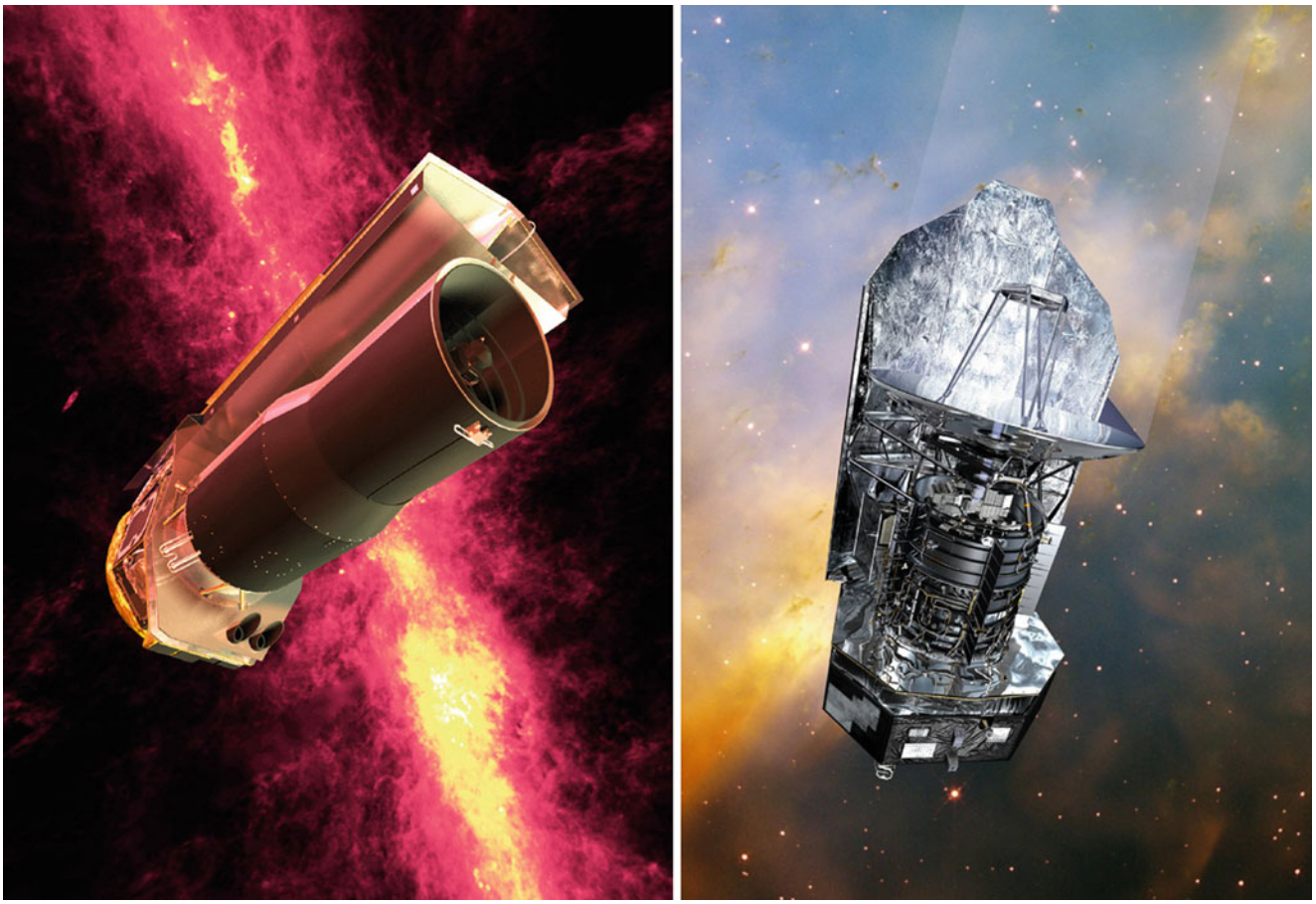


Fig. 1.33 *Left:* The Spitzer Space Telescope, launched in 2003. *Right*: The Herschel Space Observatory, launched in 2009 together with the Planck satellite. Herschel is equipped with a 3.5-m mirror and three instruments, observing between 60 and 650 μm. Credit: *Left*: NASA/JPL-Caltech. *Right*: Max-Planck-Institut für Astronomie/European Space Agency

had a diameter of 4 m, a height of 7.5 m, and weighs 3.4 tons. Herschel covered the spectral range from far-infrared to sub-millimeter wavelengths (55 μm to 670 μm), using three instruments: (1) the Photodetector Array Camera and Spectrometer (PACS), a camera and a low- to medium-resolution spectrometer for wavelengths up to about 205 μm, (2) the Spectral and Photometric Imaging REceiver (SPIRE) operated at three bands longward of $\lambda = 200\,\mu$m, and (3) the Heterodyne Instrument for the Far Infrared (HIFI), a high-resolution spectrometer operating at $\lambda \geq 150\,\mu$m. The large aperture, the sensitivity of the instruments and the wide frequency coverage made Herschel the by far most powerful FIR observatory yet. Herschel observed until end of April 2013, when it ran out of coolant.

Following the IRAS satellite, two more recent missions conducted all-sky surveys in the infrared. The Japanese satellite AKARI, launched in February 2006, mapped the entire sky at six wavelengths between 9 and 160 μm, and thus produced the first all-sky survey in the infrared after that of IRAS. The source catalogs extracted from this survey were publicly released in March 2010 and contain more than 1.3 million sources. The AKARI survey is about ten times more sensitive than IRAS, and can locate the position of a point source with an accuracy of better than 2″ at the shorter wavelengths. Given the scientific impact of the IRAS survey, one can easily foresee that of the new all-sky results. In addition, AKARI carried out pointed observations. After its liquid helium ran out, only observations at shorter wavelengths could be conducted. The Wide-field Infrared Survey Explorer (WISE) was launched in December 2009 and mapped the full sky at four wavebands between 3 and 25 μm, with an angular resolution between 6″ and 12″, and a more than hundred times larger sensitivity than IRAS. The resulting catalog from WISE contains positional and photometric information for over 563 million objects.

Another kind of infrared observatory is the Stratospheric Observatory for Infrared Astronomy (SOFIA), a 2.5 m telescope mounted onboard a refurbished Boeing 747 (see Fig. 1.34). Flying at an altitude of 12 km, the observations happen above most of the Earth atmosphere. SOFIA, a

Fig. 1.34 Stratospheric Observatory for Infrared Astronomy (SOFIA), a 2.5-m telescope onboard a refurbished Boeing 747 aircraft, designed to fly at 12 km altitude. A huge door was installed which opens at the high cruising altitude to allow for astronomical observations—a substantial challenge for the structural stability of the aircraft. Regular observations with SOFIA started in 2010. Credit: NASA/Jim Ross

US-German collaboration, has the advantage that technical developments in instrumentation can be implemented during the duration of the project.

1.3.3 Optical telescopes

The atmosphere is largely transparent in the optical part of the electromagnetic spectrum ($0.3\,\mu\text{m} \lesssim \lambda \lesssim 1\,\mu\text{m}$), and thus we are able to conduct observations from the ground. Since for the atmospheric windows in the NIR one normally uses the same telescopes as for optical astronomy, we will not distinguish between these two ranges here. Despite the tremendous progress made in all wavelength regimes, the optical and NIR spectral region is arguably the single most informative for astronomy, for a combination of two reasons: first, most of the radiation emitted by galaxies is light from stars which has its maximum in the optical regime or, in case of star formation obscured by dust, in the FIR regime, and second, the efficiency of optical detectors is highest, much more than those of infrared detectors. Together, these two points cause the observable number density of sources on the sky to be highest for optical observations.

Although optical astronomy has been pursued for many decades, it has evolved very rapidly in recent years. This is linked to a large number of technical achievements. A good illustration of this is the 10-m Keck telescope which was put into operation in 1993; this was the first optical telescope with a mirror diameter of more than 6 m. Constructing telescopes of this size became possible by the development of active optics, a method to control the surface of the mirror. A mirror of this size no longer has a stable shape but is affected, e.g., by gravitational deformation as the telescope is steered. Such a large mirror, in order to have a stable shape, would need to have a thickness comparable with its diameter, and producing and operating such mirrors is infeasible. It was also realized that part of the air turbulence that generates the seeing is caused by the telescope and its dome itself. By improving the thermal condition of telescopes and dome structures a reduction of the seeing could be achieved. The aforementioned replacement of photographic plates by CCDs, together with improvements to the latter, resulted in a vastly enhanced quantum efficiency of $\sim$ 70 % (at maximum even more than 90 %), barely leaving room for further improvements.

The throughput of optical telescopes has been immensely increased by designing wide-field CCD cameras, the largest of which is currently that of PanSTARRS, with $\sim 1.4 \times 10^9$ pixels, covering $\sim 7\,\text{deg}^2$. Furthermore, multi-object spectrographs have been built which allow us to observe the spectra of a large number of objects simultaneously. The largest of them are able to get spectra for several hundred sources in one exposure. Finally, with the Hubble Space Telescope the angular resolution of optical observations was increased by a factor of ~ 10 compared to the best sites on Earth. Further developments that will revolutionize the field even more, such as interferometry in the near IR/optical and adaptive optics, have recently been added to these achievements.

Currently, 13 optical telescopes of the 4-m class exist worldwide. They differ mainly in their location and their instrumentation. For example, the Canada-France-Hawaii Telescope (CFHT) on Mauna Kea (Fig. 1.35) has been a leader in wide-field photometry for many years, due to its extraordinarily good seeing. This is again emphasized by the installation of Megacam, a one square degree camera with $18\,000 \times 18\,000$ pixels. The Anglo-Australian Telescope (AAT) in Australia, in contrast, has distinctly worse seeing and has therefore specialized, among other things, in multi-object spectroscopy, for which the 2dF (two degree field) instrument was constructed. Most of these telescopes are also equipped with NIR instruments. The New Technology Telescope (NTT, see Fig. 1.36) made its largest contributions with its SOFI camera, a near-IR instrument with a large field-of-view of $\sim 5' \times 5'$ and an excellent image quality.

Hubble Space Telescope. To avoid the greatest problem in ground-based optical astronomy, the rocket scientist Hermann Oberth speculated already in the 1920s about telescopes in space which would not be affected by the influence of the Earth's atmosphere. In 1946 the astronomer Lyman Spitzer took up this issue again and discussed the possibilities for the realization of such a project.

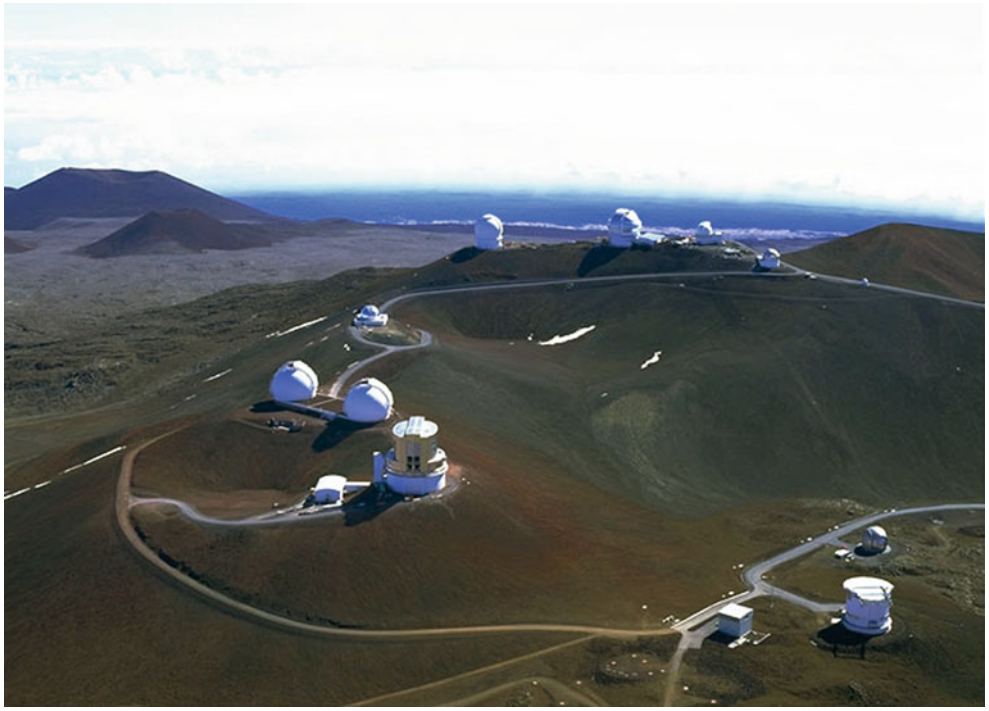


Fig. 1.35 Telescopes at the summit of Mauna Kea, Hawaii, at an altitude of 4200 m. The cylindrical dome to the left and below the center of the image contains the Subaru 8-m telescope; just behind it are the two 10-m Keck telescopes. The two large domes at the back house the Canada-France-Hawaii telescope (CFHT, 3.6 m) and the 8-m Gemini North. The telescope at the lower right is the 15-m James Clerk Maxwell sub-millimeter telescope (JCMT). Credit: R. Wainscoat, University of Hawaii

Fig. 1.36 The La Silla Observatory of ESO in Chile. On the peak in the middle, one can see the New Technology Telescope (NTT), a 3.5-m prototype of the VLT. The silvery shining dome to its left is the MPG/ESO 2.2-m telescope whose Wide Field Imager, a 8096^2 pixel camera with a 0.5° field-of-view, has been a very competitive instrument in the past decade. The picture was taken from the location of the 3.6-m telescope, the largest one on La Silla. Credit: European Southern Observatory

Shortly after NASA was founded in 1958, the construction of a large telescope in space was declared a long-term goal. After several feasibility studies and ESA's agreement to join the project, the HST was finally built. However, the launch was delayed by the explosion of the space shuttle 'Challenger' in 1986, so that it did not take place until April 24, 1990. An unpleasant surprise came as soon as the first images were taken: it was found that the 2.4 m main mirror was ground into the wrong shape. This problem was remedied in December 1993 during the first "servicing mission" (a series of Space Shuttle missions to the HST; see Fig. 1.37), when a correction lens was installed. After this, the HST could observe at its diffraction limit, i.e., with an angular resolution of better than $0''.1$, and became one of the most successful and best-known scientific instruments. In fact, the HST was far more important for extragalactic

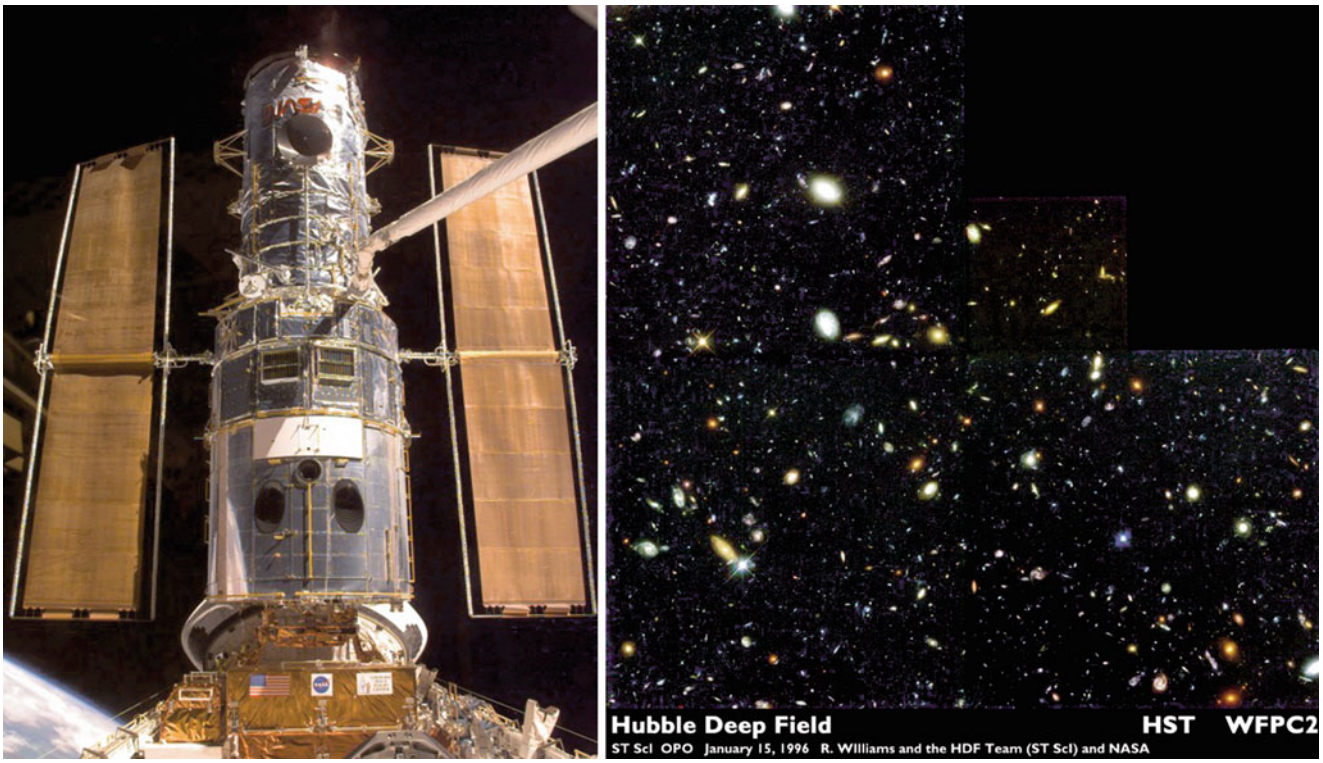


Fig. 1.37 *Left*: The HST mounted on the manipulator arm of the Space Shuttle during one of the servicing missions. *Right*: The Hubble Deep Field (North) was observed in December 1995 and the data released 1 month later. To compile this multicolor image, which at that time was the deepest image of the sky, images from four different filters were combined. The geometry of the field is due to the arrangement of the CCD detector arrays in the Wide Field and Planetary Camera 2, where one of the four chips is smaller than the other three and due to a smaller pixel scale. Credit: STScI, NASA

astronomy than anticipated before its launch, due to the fact that distant galaxies turned out to be much more compact than their local counterparts, and thus have a higher surface brightness. A main contributor to the success (and cost) of HST was a series of five visits to the telescope, where not only parts that failed (e.g., gyroscopes) were replaced and thus the lifetime of the observatory extended to well over 20 years, but also new, increasingly more powerful instruments were installed.

After the final servicing mission SM-4 in 2009, the HST carries a powerful suite of scientific instruments. The ACS (Advanced Camera for Surveys) has a field-of-view of $3'.4 \times 3'.4$ and a pixel scale of $0''.05$. It was installed on Hubble in 2002, had to be shut down after a malfunction in 2007, but after the repair in 2009, it continues to be a powerful workhorse for high-resolution imaging in the optical wavebands. Several imaging surveys have been carried out with ACS, some of which will be considered in some detail in later chapters. Newly installed during SM-4 was the Wide Field Camera 3 (WFC3), replacing the WFPC2 (Wide Field and Planetary Camera), which has been the most active instrument of HST since then. WFC3 covers a very broad range of wavebands, distributed over two different 'channels'. The UVIS channel operates between 0.2 and $1\,\mu$m with a $2'.6$ field of view, whereas the NIR channel operates between 0.85 and $1.7\,\mu$m and has a $\sim 2'$ field. This field is considerably larger than that of the other NIR instrument on HST, the NICMOS (Near Infrared Camera and Multi Object Spectrograph) instrument. Already in its first months of operation, WFC3 led to great progress in the field of very high redshift galaxies, by increasing the number of known galaxies with redshifts $z \geq 6$ by a large factor. The SM-4 mission also brought the Cosmic Origins Spectrograph (COS) to HST, operating in the UV region of the spectrum. The spectroscopic capability of HST in the UV-range, unavailable from the ground, was increased with COS by a large factor compared to the other UV instrument STIS (Space Telescope Imaging Spectrograph).

HST has provided important insights into our Solar System and the formation of stars, but it has achieved milestones in extragalactic astronomy. With HST observations of the nucleus of M87 (Fig. 1.11), one derived from the Doppler shift of the gas emission that the center of this galaxy contains a black hole of 2 billion Solar masses. HST has also proven that black holes exist in other galaxies and AGNs. The enormously improved angular resolution has

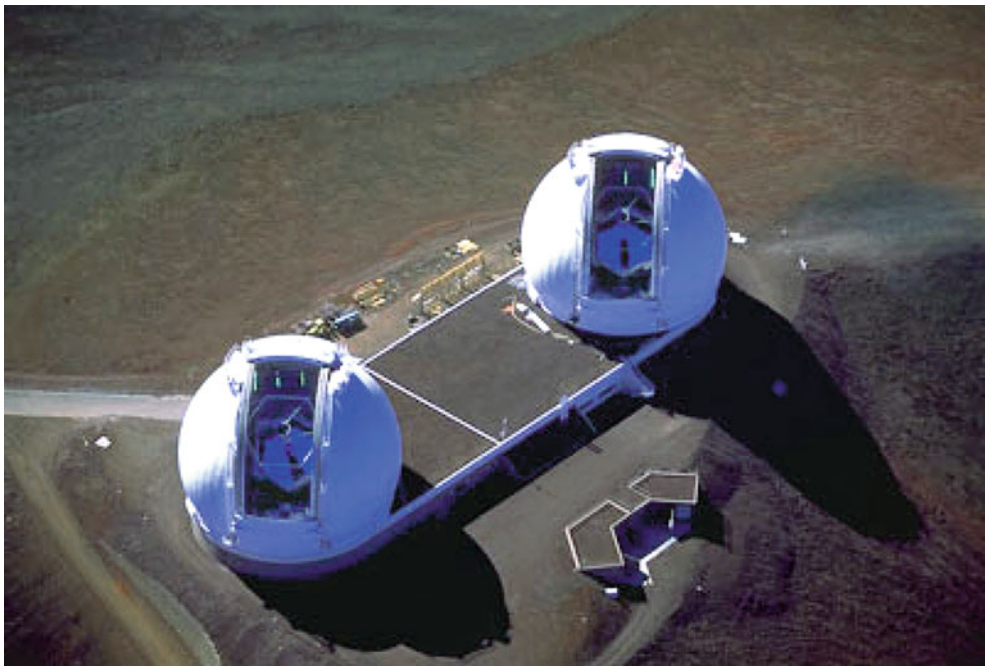


Fig. 1.38 The two Keck telescopes on Mauna Kea. With Keck I the era of large telescopes was heralded in 1993. Credit: R. Wainscoat, University of Hawaii

allowed us to study galaxies to a hitherto unknown level of detail.

Arguably the most important contribution of the HST to extragalactic astronomy are the Hubble Deep Fields. Scientists managed to convince Robert Williams, then director of the Space Telescope Science Institute, to use the HST to take a very deep image in an empty region of the sky, a field with (nearly) no foreground stars and without any known clusters of galaxies. At that time it was not clear whether anything interesting at all would result from these observations. Using the observing time that is allocated to the Director, the 'director discretionary time', HST was pointed at such a field in the Big Dipper, taking data for 10 days December 1995. The outcome was the Hubble Deep Field North (HDFN), one of the most important astronomical data sets, displayed in Fig. 1.37. From the HDFN and its southern counterpart, the HDFS, one obtains information about the early states of galaxies and their evolution. One of the first conclusions was that most of the early galaxies are very small and classified as irregulars. In 2002, the Hubble Ultra Deep Field (HUDF) was observed with the then newly installed ACS camera. Not only did it cover about twice the area of the HDFN but it was even deeper, by about one magnitude, owing to the higher sensitivity of ACS compared to WFPC2. We will discuss some of the imaging surveys of HST in more detail in Sect. 9.2.1.

Large Telescopes. For more than 40 years the 5-m telescope on Mt. Palomar was the largest telescope in the western world—the Russian 6-m telescope suffered from major problems from the outset. 1993 saw the birth of a new class of telescopes, of which the two Keck telescopes (see Fig. 1.38) were the first, each with a mirror diameter of 10 m.[11]

The site of the two Kecks at the summit of Mauna Kea (see Fig. 1.35) provides ideal observing conditions for many nights per year. This summit is now home to several large telescopes. The Japanese Subaru telescope and Gemini North are also located here, as well as the aforementioned CFHT and JCMT. The significant increase in sensitivity obtained by Keck, especially in spectroscopy, permitted completely new insights, for instance through absorption line spectroscopy of quasars. Keck was also essential for the spectroscopic verification of innumerable galaxies of redshift $z \gtrsim 3$, which are normally so dim that they cannot be examined with smaller telescopes.

The largest ground-based telescope project to date was the construction of the Very Large Telescope (VLT) of the European Southern Observatory (ESO), consisting of four telescopes each with a diameter of 8.2 m. ESO already operated the La Silla Observatory in Chile (see Fig. 1.36), but a better location was found for the VLT, the Cerro Paranal (at an altitude of 2600 m). This mountain is located in the Atacama desert, one of the driest regions on Earth. To build the telescopes on the mountain a substantial part of the mountain top first had to be cut off (Fig. 1.39).

[11] The 13 telescopes with diameter >8 m are: The two Keck telescopes, the four VLTs, Gemini-North (Mauna Kea) and Gemini-South (Chile), Subaru (Mauna Kea), the Hobby–Eberly Telescope (McDonald Observatory, Texas), the Large Binocular Telescope in Arizona consisting of two telescopes (see Fig. 1.44 below), and the Gran Telescopio Canarias at the Roque de los Muchachos Observatory on La Palma, Spain.

Fig. 1.39 The Paranal observatory after completion. The four large domes host one of the VLTs each. The smaller dome, seen to the left of the rightmost VLT, hosts the VLT Survey Telescope (VST), a dedicated 2.6 m telescope equipped with the wide-field optical camera Omega-CAM. The four much smaller domes host 1.5 m auxiliary telescopes, which are used in combination with the VLTs for optical interferometry. In the background, the VISTA (Visible and Infrared Survey Telescope for Astronomy) telescope is seen, equipped with a wide-field near-infrared camera. The buildings in front contain the control room for the telescopes and instruments, and a guest house for observers. Before the observatory was constructed, the top of the mountain was flattened to get a leveled surface of diameter $\sim$ 300 m, large enough to accommodate the telescopes and the facilities used for optical interferometry (VLTI). Credit: European Southern Observatory/G.Hüdepohl

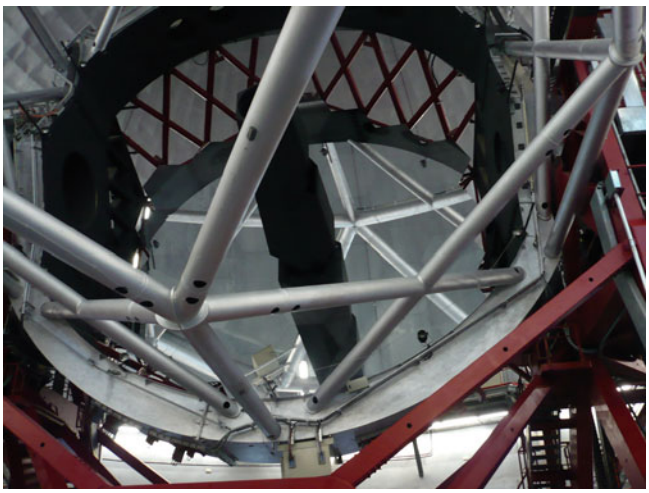


Fig. 1.40 The segmented mirror of the Gran Telescopio Canarias (GTC) at the Roque de los Muchachos Observatory on La Palma, inside its support structure. The hexagonal elements can clearly be seen. At an altitude of 2400 m, this observatory is the best astronomical site in Europe. Credit: P. Schneider, Argelander-Institut für Astronomie, Universität Bonn

In contrast to the Keck telescopes and the Gran Telescopio Canarias (GTC; see Fig. 1.40), which have a primary mirror that is segmented into 36 hexagonal elements, the mirrors of the VLT are monolithic, i.e., they consist of a single piece. However, they are very thin compared to the 5-m mirror on Mt. Palomar, far too thin to be intrinsically stable against gravity and other effects such as thermal deformations. Therefore, as for the Kecks, the shape of the mirrors has to be controlled electronically (see Fig. 1.41). The monolithic structure of the VLT mirrors yields a better image quality than that of the Keck telescopes, resulting in an appreciably simpler point-spread function.[12]

Each of the four telescopes has three accessible foci; this way, 12 different instruments can be installed at the VLT at any time. Switching between the three instruments is done with a deflection mirror. The permanent installation of the instruments allows their stable operation.

[12]The point-spread function (PSF) $P(\boldsymbol{\theta})$ describes the shape of the brightness profile of a point source as seen in the detector. Owing to seeing and diffraction effects, it has a finite width. The images of extended sources are also affected by the PSF: each small part of an extended sources can be considered as a point source, whose brightness in the detector is smeared by the PSF. Thus, if $I(\boldsymbol{\theta})$ is the true brightness profile, the observed one is given by

$$I^{\rm obs}(\boldsymbol{\theta}) = \int \mathrm{d}^2\theta'\; I(\boldsymbol{\theta}')\, P(\boldsymbol{\theta} - \boldsymbol{\theta}')\;,$$

where P is normalized to unity, $\int \mathrm{d}^2\theta\; P(\boldsymbol{\theta}) = 1$.

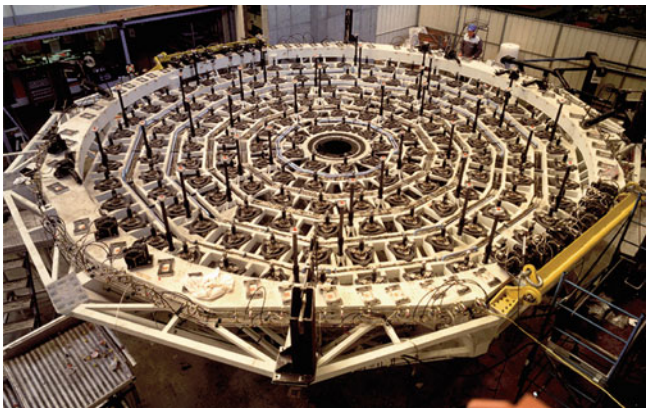


Fig. 1.41 The active optics system at the VLT. Each mirror is supported at 150 points; at these points, the mirror is adjusted to correct for deformations. The primary mirror is always shaped such that the light is focused in an optimal way, with its form being corrected for the changing gravitational forces when the telescope changes the pointing direction. In adaptive optics, in contrast to active optics, the wave front is controlled: the mirrors are deformed with high frequencies in such a way that the wave front is as planar as possible after passing through the optical system. In this way one can correct for the permanently changing atmospheric conditions and achieve images at diffraction-limited resolution, though only across a fairly small region of the focal plane. Credit: European Southern Observatory

The VLT (Fig. 1.42) also marked the beginning of a new form of ground-based observation with large optical telescopes. Whereas until recently an astronomer proposing an observation was assigned a certain number and dates of nights in which she could observe with the telescope, the VLT is mainly operated in the so-called service mode. The observations are performed by local astronomers according to detailed specifications that must be provided by the principal investigator of the observing program, and the data are then transmitted to the astronomer at her home institution. A significant advantage of this procedure is that one can better account for special requirements for observing conditions. For example, observations that require very good seeing can be carried out during the appropriate atmospheric conditions. With service observing the chances of getting a useful data set are increased. More than half of the observations with the VLT are performed in service mode.

Another aspect of service observing is that the astronomer does not have to make the long journey, at the expense of also missing out on the adventure and experience of observing. As mentioned before, the best astronomical sites are at quite remote places, and traveling to such an observatory is a physical experience. A trip from, e.g., Bonn to the Paranal starts with a train ride to Düsseldorf airport, a flight from there to Paris or Madrid, followed by a more than 12 h flight to Santiago de Chile. From there, another 2 h flight towards the Northern coast of Chile takes one to the city of Antofagasta. The final part of the journey is a 2 h car ride through the desert—literally, the last plant one sees in the outskirts of Antofagasta. And finally, from the distance, one can see the four majestic domes on the top of a mountain, the final destination of the journey.

Adaptive optics, integral field spectroscopy, and interferometry. New technical innovations in astronomy lead to improved data products; two examples should be mentioned

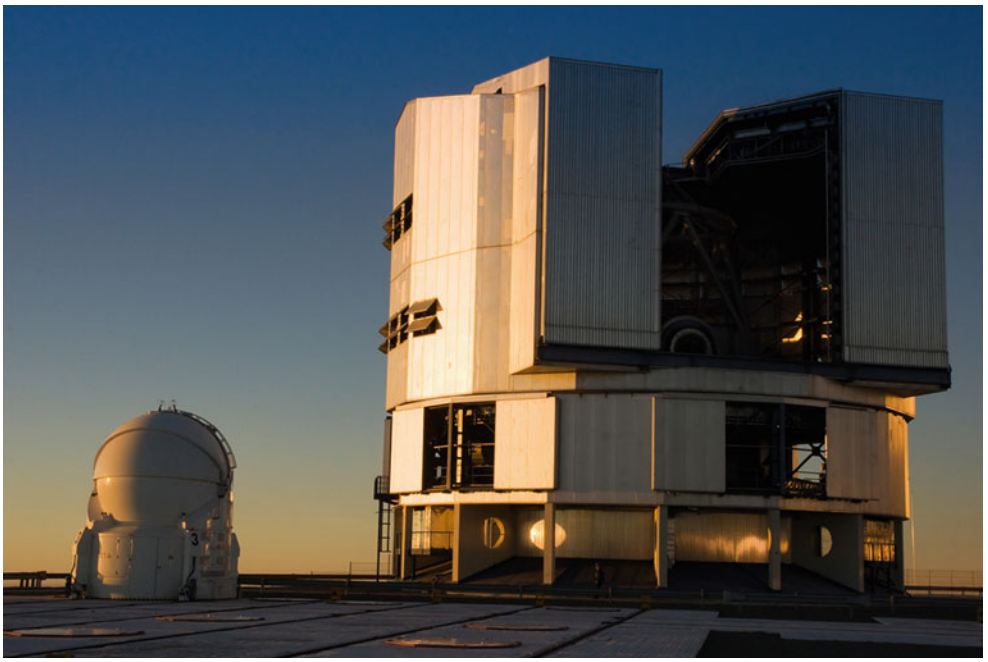


Fig. 1.42 One of the four Unit Telescopes of the VLT on the right, together with one of the auxiliary telescopes at the left. These auxiliary telescopes can be moved to allow for variable baseline configurations in optical interferometry. Credit: Iztok Boncina/European Southern Observatory

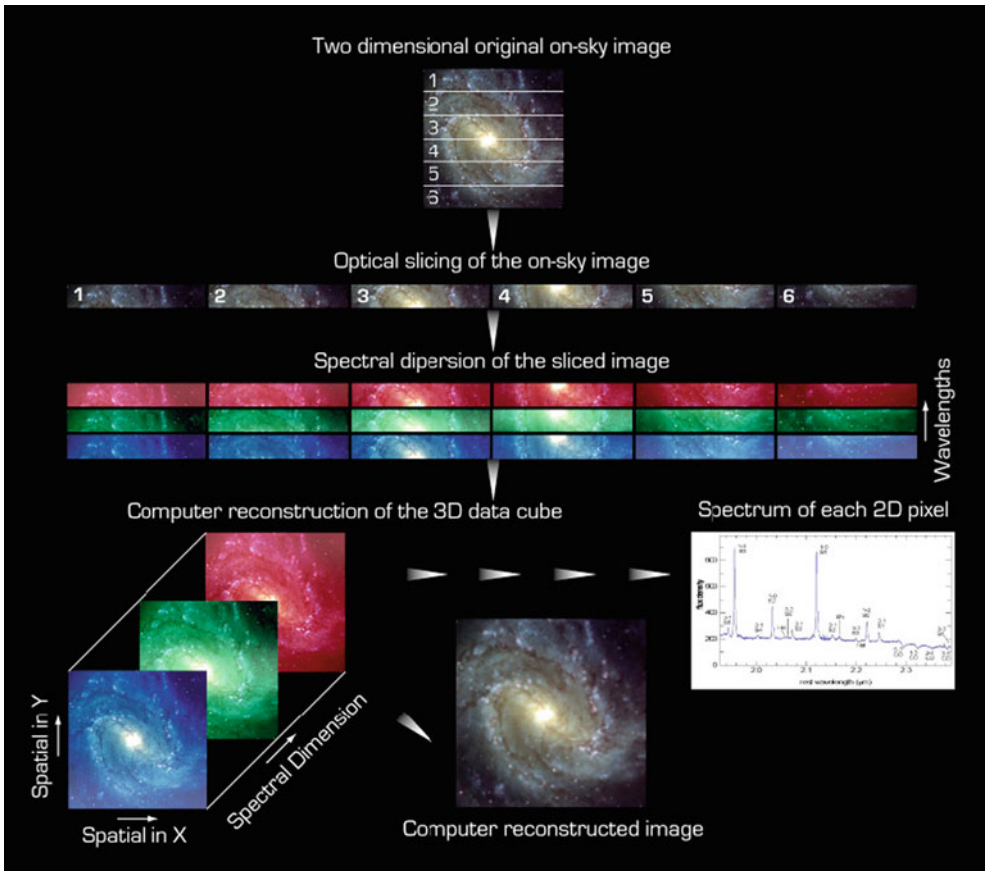


Fig. 1.43 The principle of an integral field unit (IFU) based on image slicing is illustrated. In a first step, the image of an object is 'sliced', using an arrangement of mirrors. The light of each slice is then send through a long-slit spectrograph, so that a spectrum from each pixel of a slice is obtained. Hence, in this way one obtains a spectrum of each pixel element in the area of the extended object. Adding up the light from all wavelengths in a pixel then yields the original image of the object. By weighting the contributions from the various wavelengths by an appropriate filter function, images in different wavebands can be obtained. Using lines in the spectrum of each image element, the corresponding Doppler shift can be obtained and a two-dimensional velocity field of the object (such as a rotation curve) can be reconstructed. Credit: European Southern Observatory

here. One is adaptive optics, a technique to obtain an angular resolution approaching the diffraction limit of the telescope. This is usually not the case, as turbulence in the atmosphere leads to a blurring of an image. In adaptive optics, one accounts for this by deforming the mirror at a high frequency, as to counteract the changing image position on the sky due to the turbulence. The wavefront of the incoming light is controlled by observing a bright reference source located closely on the sky to the target of interest. The motion of the reference source on the CCD then yields the necessary information about the wavefront deformation, which thus can be corrected for. In many cases, the source to be observed does not have a bright source close-by. One way to use adaptive optics in this situation is to generate an artificial source on the sky, by pointing a laser upwards. By tuning the laser to a wavelength of 5892 Å, sodium atoms in the upper atmosphere at $\sim$ 90 km altitude are excited and re-emit the light. In this way, an artificial light source (called laser guide star) is created which is viewed through the same atmosphere as the source of interest.

Another innovative technique, called integral field spectroscopy, allows one to obtain the spectral energy distribution for different regions of an extended source, pixel by pixel. Several different methods for this are used; a particular one, based on image slicing by an arrangement of mirrors, is explained in more detail in Fig. 1.43.

Optical interferometry is also reaching a state of maturity. Like in radio astronomy, the light received from several telescopes can be combined to obtain a higher-resolution image of a source. In contrast to VLBI techniques, the light at the different telescopes is not recorded and correlated afterwards; instead, the light beams from the different telescopes need to be combined directly. Below the plateau of the Paranal observatory is a large tunnel system where the light beams of the VLTs and/or the auxiliary telescopes are combined. The two Keck telescopes were built on a common structure to enable interferometry. The latest development in optical interferometry is the Large Binocular Telescope (LBT), where the two mirrors are mounted on a single structure (see Fig. 1.44).

1.3.4 UV telescopes

Radiation with a wavelength shorter than $\lambda \lesssim 0.3\,\mu\mathrm{m} = 3000$ Å cannot penetrate the Earth's atmosphere but is instead absorbed by the ozone layer, whereas radiation at wavelengths below 912 Å is absorbed by neutral hydrogen in the

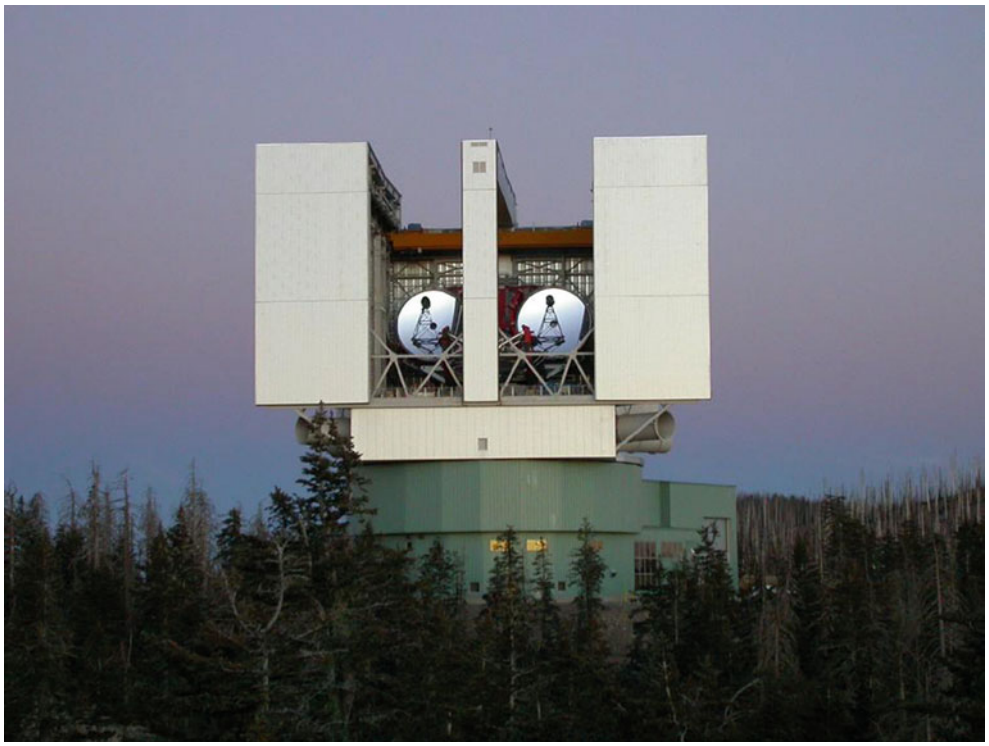


Fig. 1.44 The Large Binocular Telescope (LBT) on Mount Graham in Arizona. The two 8.4 m primary mirrors are mounted on a single structure, share one gigantic dome, and has been built specifically for optical interferometry. Credit: Large Binocular Telescope Observatory; courtesy NASA/JPL-Caltech

interstellar medium. The range between these two wavelengths is the UV part of the spectrum, in which observation is only possible from space.

The Copernicus satellite (also known as the Orbiting Astronomical Observatory 3, OAO-3) was the first long-term orbital mission designed to observe high-resolution spectra at ultraviolet wavelengths. In addition, the satellite contained an X-ray detector. Launched on August 21, 1972, it obtained UV spectra of 551 sources until its decommissioning in 1981. Among the achievements of the Copernicus mission are the first detection of interstellar molecular hydrogen H_2 and of CO, and measurements of the composition of the interstellar medium as well as of the distribution of OVI, i.e., five-time ionized oxygen.

The IUE (International Ultraviolet Explorer) operated between 1978 and 1996 and proved to be a remarkably productive observatory. During its 18 years of operation more than 10^5 spectra of galactic and extragalactic sources were obtained. In particular, the IUE contributed substantially to our knowledge of AGN.

The HST, with its much larger aperture, marks the next substantial step in UV astronomy. Many new insights were gained with the HST, especially through spectroscopy of quasars in the UV, insights into both the quasars themselves and, through the absorption lines in their spectra, into the intergalactic medium along the line-of-sight towards the sources. In 1999 the FUSE (Far Ultraviolet Spectroscopic Explorer) satellite was launched. From UV spectroscopy of absorption lines in luminous quasars this satellite provided us with a plethora of information on the state and chemical composition of the intergalactic medium.

While the majority of observations with UV satellites were dedicated to high-resolution spectroscopy of stars and AGNs, the prime purpose of the Galaxy Evolution Explorer (GALEX) satellite mission (Fig. 1.45), launched in 2003, was to compile extended photometric surveys. GALEX observed at wavelengths $1350\,\text{Å} \lesssim \lambda \lesssim 2830\,\text{Å}$ and performed a variety of surveys. Amongst them are the All-Sky Imaging Survey, covering 26 000 square degrees of the extragalactic sky, the Medium Imaging survey, with considerably deeper imaging of 1000 square degrees in field where spectroscopic redshift surveys (like SDSS) were available, and the Deep Imaging Survey of 80 square degrees and an exposure time of about 8 h per field. Several more specialized surveys were carried out, including one on nearby galaxies and on the Milky Way. In addition, GALEX performed several spectroscopic surveys. The results from GALEX are of great importance, especially for the study of the star-formation rate in nearby and distant galaxies. In June 2013, the operation of GALEX was terminated.

1.3.5 X-ray telescopes

As mentioned before, interstellar gas absorbs radiation at wavelengths shortward of 912 Å, the so-called Lyman edge. This corresponds to the ionization energy of hydrogen in its

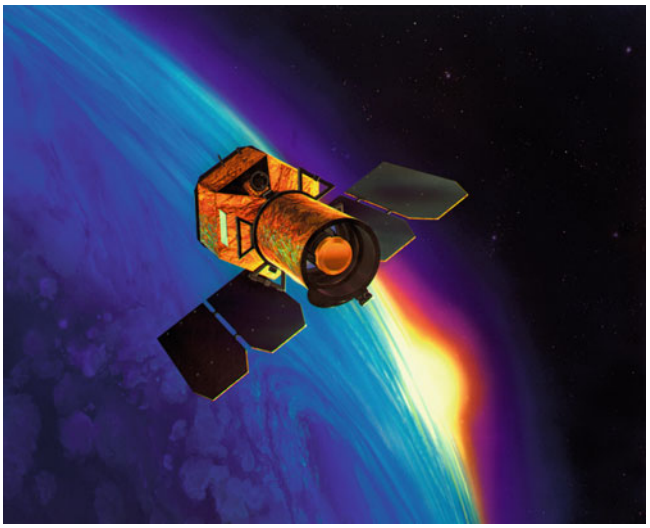


Fig. 1.45 The Galaxy Evolution Explorer GALEX was launched in 2003, and decommissioned after 10 years of astronomical observations in June 2013. GALEX had a 50 cm mirror and operated in the UV spectral range. Credit: NASA/JPL-Caltech

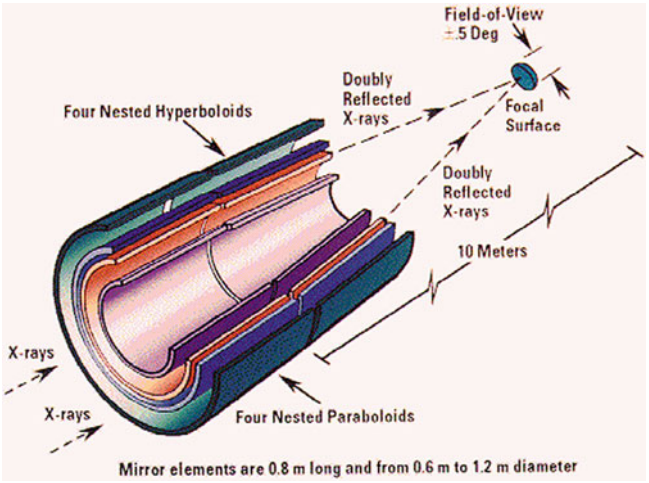


Fig. 1.46 Principle of a Wolter telescope for X-ray astronomy. X-rays are reflected by a metal surface only if the angle of incidence is very small, i.e., if the photon direction is almost parallel (within $\sim 2°$) to the metal surface. This principle can be employed for X-ray mirrors, by constructing a tube, consisting of a paraboloidal shaped surface combined with a hyperboloidal one in a way that X-ray photons are focussed. The effective area of such a mirror is very small, as only a small annular region of the photon beam hits this mirror, due to the small projected surface. The X-ray telescopes on XMM and Chandra are therefore composed of nested layers of such surfaces, each one acting like a focusing surface by itself, thereby multiplying the effective area. The figure illustrates the mirror of the Chandra observatory. Source: Wikipedia

ground state, which is 13.6 eV. Only at energies about ten times this value does the ISM become transparent again[13] and this denotes the low-energy limit of the domain of X-ray astronomy. Typically, X-ray astronomers do not measure the frequency of light in Hertz (or the wavelength in μm), but instead photons are characterized by their energy, measured in electron volts (eV).

The birth of X-ray astronomy was in the 1960s. Rocket and balloon-mounted telescopes which were originally only supposed to observe the Sun in X-rays also received signals from outside the Solar System. UHURU (the Swahili word for 'freedom') was the first satellite to observe exclusively the cosmic X-ray radiation and compiled the first X-ray map of the sky, discovering about 340 sources. This catalog of point sources was expanded in several follow-up missions, especially by NASA's High Energy Astrophysical Observatory (HEAO-1) which also detected a diffuse X-ray background radiation. On HEAO-2, also known as the Einstein satellite, the first Wolter telescope (see Fig. 1.46) was used for imaging, increasing the sensitivity by a factor of nearly a thousand compared to earlier missions. The Einstein observatory also marked a revolution in X-ray astronomy because of its high angular resolution, about $2''$ in the range of 0.1–4 keV. Among the great discoveries of the Einstein satellite is the X-ray emission of many clusters of galaxies that traces the presence of hot gas in the space between the cluster galaxies. The total mass of this gas significantly exceeds the mass of the stars in the cluster galaxies and therefore represents the main contribution to the baryonic content of clusters.

The next major step in X-ray astronomy was ROSAT (ROentgen SATellite; Fig. 1.47), launched in 1990. During the first 6 months of its 9-year mission ROSAT produced an all-sky map at far higher resolution than UHURU; this is called the ROSAT All Sky Survey. More than 10^5 individual sources were detected in this survey, the majority of them being AGNs. In the subsequent period of pointed observations ROSAT examined, among other types of sources, clusters of galaxies and AGNs. One of its instruments (the Position Sensitive Proportional Counter PSPC) provided spectral information in the range between 0.1 and 2.4 keV at an angular resolution of $\sim 20''$, while the other (High-Resolution Instrument HRI) instrument had a much better angular resolution ($\sim 3''$) but did not provide any spectral information. The Japanese X-ray satellite ASCA (Advanced Satellite for Cosmology and Astrophysics), launched in 1993, was able to observe in a significantly higher energy range of 0.5–12 keV and provided spectra of higher energy resolution, though at reduced angular resolution.

Since 1999 two new powerful satellites are in operation: NASA's Chandra observatory and ESA's XMM-Newton (X-ray Multi-Mirror Mission; see Fig. 1.47). Both have a large photon-collecting area and a high angular resolution, and they also set new standards in X-ray spectroscopy.

[13] This is due to the fact that the ionization cross section behaves approximately as $\propto \nu^{-3}$, where ν is the photon frequency; hence, the probability that a photon will be absorbed through photo-ionization of neutral hydrogen is ~ 1000 times smaller at 0.1 keV than at the Lyman limit.

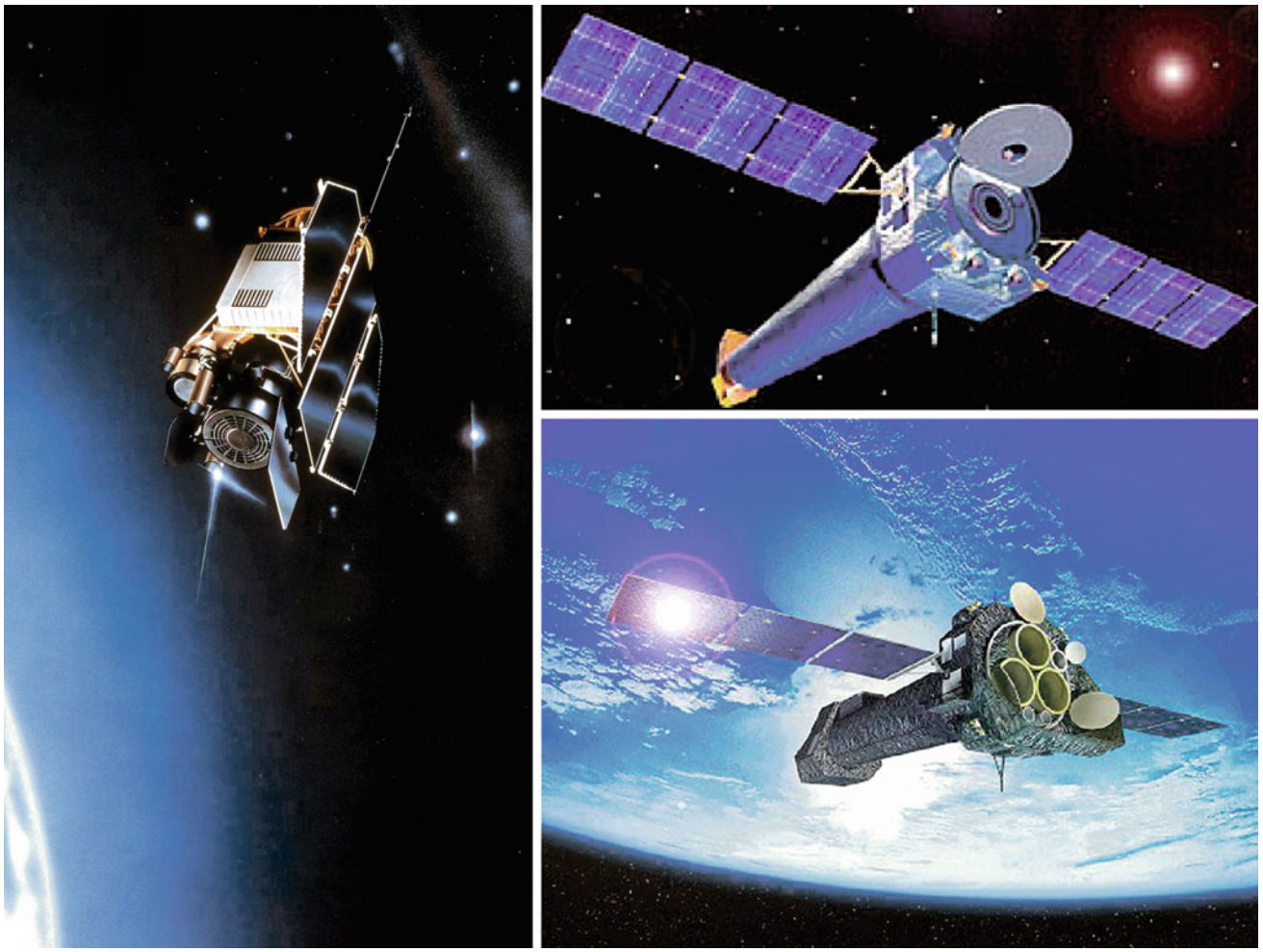


Fig. 1.47 *Left*: ROSAT, a German-US-British cooperation, observed from 1990 to 1999 in the energy range between 0.1 and 2.5 keV (soft X-ray). *Upper right*: Chandra was launched in July 1999. The energy range of its instruments lies between 0.1 and 10 keV. Its highly elliptical orbit permits long uninterrupted exposures. *Lower right*: XMM-Newton was launched in December 1999 and has since been in operation; it is the most successful ESA mission up to today, as measured from the number of publications resulting from its data. Observations are carried out with three telescopes at energies between 0.1 and 15 keV. Credit: *Left*: Max-Planck-Intitut für extraterrestrische Forschung (MPE) & DLR. *Top right*: NASA/CXC/SAO. *Bottom right*: European Space Agency

Compared to ROSAT, the energy range accessible with these two satellites is larger, from 0.1 to $\sim$ 10 keV. The angular resolution of Chandra is about $0''.5$ and thus, for the first time, comparable to that of optical telescopes. This high angular resolution led to major discoveries in the early years of operation. For instance, well-defined sharp structures in the X-ray emission from gas in clusters of galaxies were discovered, as well as X-ray radiation from the jets of AGNs which had been previously observed in the radio. Furthermore, Chandra discovered a class of X-ray sources, termed Ultra-luminous Compact X-ray Sources (ULXs), in which we may be observing the formation of black holes (Sect. 9.3.1). XMM-Newton has a larger sensitivity compared to Chandra, however at a somewhat smaller angular resolution. Among the most important observations of XMM-Newton are the spectroscopy of AGNs and of clusters of galaxies.

With Suzaku, a Japanese satellite launched in July 2005, a new X-ray observatory became available. One of its advantages compared to Chandra and XMM-Newton is its low orbit around the Earth, which keeps it inside the Earth magnetosphere and thus shields it from most of the Solar wind. Therefore, the radiation background of Suzaku is lower than for the other two X-ray observatories. This renders Suzaku particularly useful to study the low surface brightness outer regions of galaxy clusters.

1.3.6 Gamma-ray telescopes

The existence of gamma radiation was first postulated in the 1950s. This radiation is absorbed by the atmosphere, which is fortunate for the lifeforms on Earth. The first γ-ray

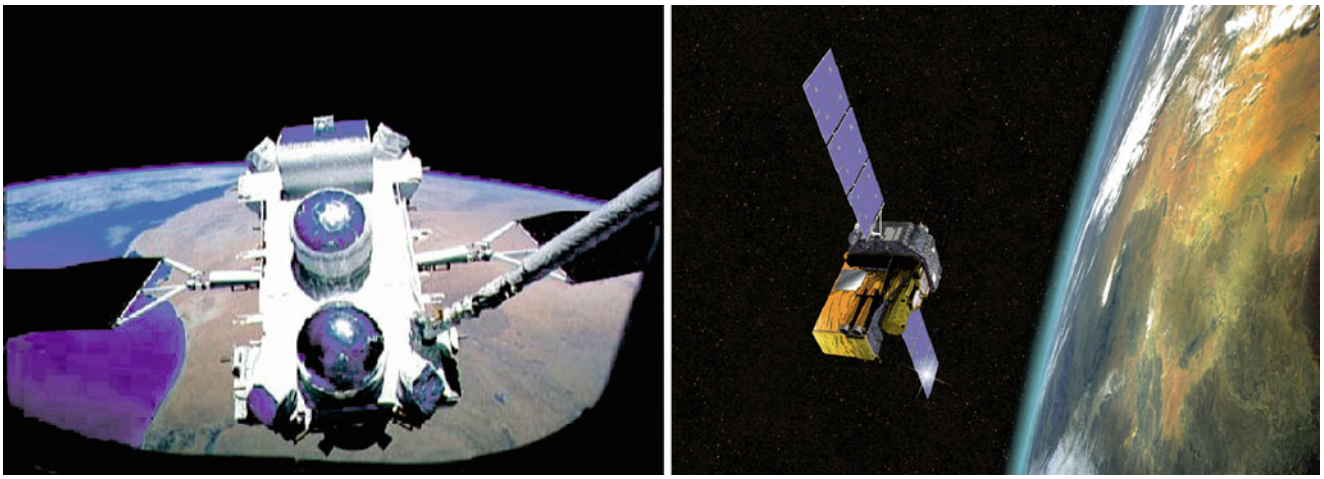


Fig. 1.48 The *left image* shows the Compton Gamma Ray Observatory (CGRO) mounted on the Space Shuttle manipulator arm. This NASA satellite carried out observations between 1991 and 2000. It was finally shut down after a gyroscope failed, and it burned up in the Earth's atmosphere in a controlled re-entry. ESA's Integral observatory, in operation since 2002, is shown *on the right*. Credit: *Left*: NASA. *Right*: ESA

telescopes were mounted on balloons, rockets, and satellites, and picked up fewer than 100 photons from cosmic sources. Those gamma photons had energies in the GeV range and above.

Detailed observations became possible with the satellites SAS-2 and COS-B. They compiled a map of the galaxy, confirmed the existence of a gamma background radiation, and for the first time observed pulsars in the gamma range. The first Gamma Ray Bursts (GRB), extremely bright and short-duration flashes on the gamma-ray sky, were detected from 1967 onwards by military satellites. Only the Italian-Dutch satellite Beppo-SAX (1996–2002) managed to localize a GRB with sufficient accuracy to allow an identification of the source in other wavebands, and thus to reveal its physical nature; we will come back to this subject in Sect. 9.7.

An enormous advance in high-energy astronomy was made with the launch of the Compton Gamma Ray Observatory (CGRO; Fig. 1.48) in 1991; the observatory was operational for 9 years. It carried four different instruments, among them the Burst And Transient Source Experiment (BATSE) and the Energetic Gamma Ray Experiment Telescope (EGRET). During its lifetime BATSE discovered about 3000 GRBs and contributed substantially to the understanding of the nature of these mysterious gamma-ray flashes. EGRET discovered many AGNs at very high energies above 20 MeV, which hints at extreme processes taking place in these objects.

One of the successors of the CGRO, the Integral satellite, was put into orbit as an ESA mission by a Russian Proton rocket at the end of 2002. At a weight of two tons, it was the heaviest ESA satellite that had been launched by then. It observes primarily at energies between 15 keV and 10 MeV in the gamma range, but has additional instruments for observation in the optical and X-ray regimes.

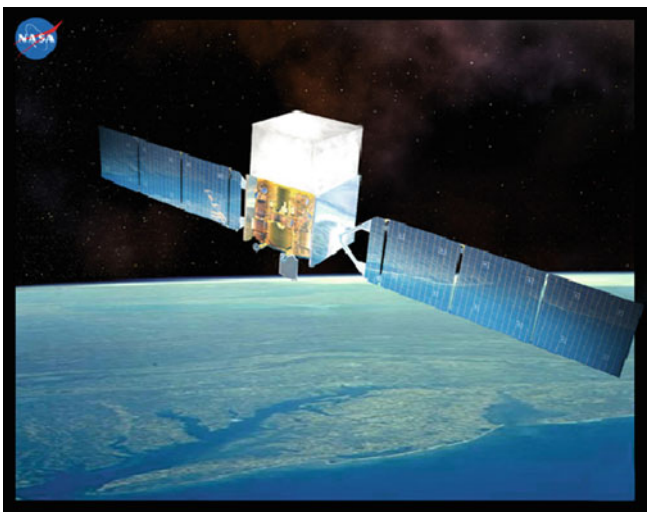


Fig. 1.49 Artist's drawing of the Fermi satellite, launched in 2008. Fermi observes at energies between 10 keV and 300 GeV. Credit: NASA E/PO, Sonoma State University, Aurore Simonnet

The other successor of the CGRO was launched in June 2008, the Fermi Gamma-ray Space telescope (Fig. 1.49). It carries two main instruments, the gamma-ray burst monitor (GBM) and the Large Area Telescope (LAT). With an energy range between 30 MeV and 300 GeV, comparable to, but wider than that of EGRET on CGRO, the LAT observes in a higher energy regime than Integral. Compared to EGRET, Fermi-LAT has higher sensitivity, a fourfold field-of-view (corresponding to 1/6 of the full sky), a better angular and energy resolution and a much better time resolution.

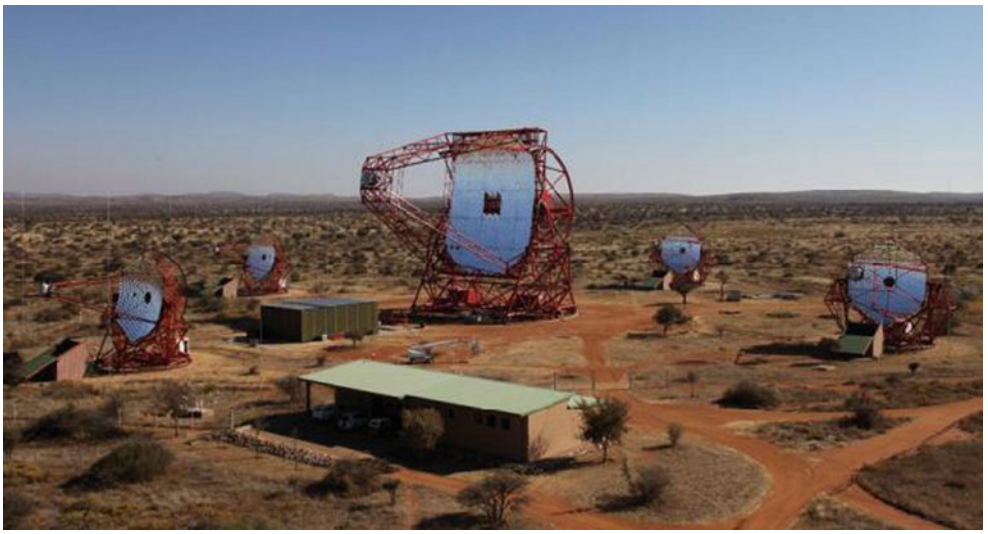


Fig. 1.50 The four 13 m-diameter telescopes, and the new 28-m telescope of the High Energy Stereoscopic System (H.E.S.S.) in Namibia, in South-West Africa. The optical quality of these mirrors is much lower than that of other optical telescopes. The angular resolution of the telescopes needs not to be better than a few arcminutes, as given by the width of the cone into which the Cherenkov light is emitted. Correspondingly, the pixel size is also adapted to this required angular resolution. Credit: H.E.S.S.-collaboration/Max-Planck-Institut für Kernphysik, Heidelberg

Observations are conducted in the sky survey mode, mapping the whole sky every 3 h. Fermi-LAT made significant discoveries in its first years of operations in the field of pulsar, gamma-ray bursts, active galactic nuclei and the gamma-ray background radiation.

At energies above 100 GeV, the photon flux even from the brightest sources is too small to be detectable with space-based instruments—characteristic values are $\sim 10^{-11}$ photons $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ above 1 TeV, or a few photons per m^2 per year. However, photons of these energies can be detected from the ground, using methods derived from cosmic ray physics. They make use of the fact that a high-energy photon hits an atom in the upper atmosphere, and in the interaction process, additional particles are generated, which by themselves are energetic enough to produce charged particles in further collisions. In this way, a number of particles are produced that are ultra-relativistic (meaning that their velocity is very close to that of the velocity of light) and all propagate in essentially the same direction as the incoming photon. Such a process is called an air shower. There are two different ways these air showers can be detected. First, some particles from the air shower propagate to the ground, and can there be detected in particle detectors.

The second method makes use of the fact that the propagation velocity of the particles in the air shower is higher than the local velocity of light in the atmosphere; in this case, charged particles emit Cherenkov radiation, and this radiation can be detected with optical telescopes. For TeV photons, the maximum of the air shower is found at altitudes around 10 km, and the minimum energy of an electron to emit Cherenkov radiation there is $\sim$ 40 MeV. About 100 such electrons need to be produced in the air shower in order to detect it unambiguously, which sets a threshold of the energy a photon must have to be detectable in this way. The typical energy threshold for current experiments is $\sim$ 100 GeV $=$ 0.1 TeV; one thus calls this the regime of TeV-astronomy. A major problem is the fact that charged particles from space—that is, cosmic rays (see Sect. 2.3.4)—also initiate air showers which have to be separated from the photon-induced ones. This is achieved by noting that air showers caused by charged particles are more complex, less collimated than the photon-induced ones. Another aspect of this technique is that one needs a dark and clear sky for the detection of the Cherenkov radiation, which limits the fraction of time in which such observations can be carried out.

Several such Cherenkov imaging telescopes are in operation, the most productive one up to now being H.E.S.S. (High Energy Stereoscopic System), shown in Fig. 1.50. Located in Namibia, H.E.S.S. consists of four telescopes each having a diameter of 13 m, recently complemented by a 28-m telescope. They observe the same region in the sky and detect the Cherenkov light from air showers, though under different viewing angles—they offer a stereoscopic view of the air shower. That permits the reconstruction of the geometry of the shower, in particular its direction and its intensity, resulting in an estimate of the direction of the incoming photon and its energy. In effect, the detector for the ultra-high energy photons is the atmosphere, and the effective area of the experiment is given by the field-of-view that the telescopes can cover, projected to an altitude of $\sim$ 10 km. In case of H.E.S.S., this is about $5 \times 5\,\mathrm{deg}^2$ yielding an effective area of $\sim 10^5\,\mathrm{m}^2$. The accuracy with which the direction

Fig. 1.51 One of the two MAGIC telescopes, located at the Roque de los Muchachos Observatory on La Palma. With their 17 m diameter, the MAGIC telescopes observe the Cherenkov radiation from air showers generated by TeV photons. The detector consists of >450 photo-multipliers

and energy of an individual photon can be determined is about $5'$ and $\sim$ 15 %, respectively. This direction accuracy allows one to determine the position of strong sources with an accuracy of $\sim$ $10''$. Figure 1.51 shows one telescope of the MAGIC (Major Atmospheric Gamma-ray Imaging Cherenkov Telescopes) experiment on La Palma, another of the Cherenkov imaging telescopes.

Up to the present, more than 100 sources have been detected at energies >100 GeV, most of them close to the Galactic plane. Many of them are as yet unidentified, but TeV radiation is observed from supernova remnants, pulsars and their immediate environment, and compact binary systems. Away from the Galactic plane, all of the (more than 50) detected TeV sources are active galactic nuclei, most of them blazars. Their very high energy emission provides insight into the processes that power these active objects.

1.4 Surveys

Modern astronomical research is partly based on large data sets, for example the spectra of a large number of objects used to investigate their spectral properties statistically. In some cases, these surveys are carried out for a single scientific objective, but frequently the same data are useful also for other branches of astronomy. Indeed, some of these data sets are very versatile and form an essential tool for a wide range of scientific applications. A few of these surveys which are of great relevance for several of the topics covered by this book are briefly described here, whereas more specialized ones will be discussed in connection with their (major) application in later chapters.

All-sky surveys. The most obvious example for versatile surveys are all-sky imaging surveys. If one finds a new source in a given waveband, then an obvious first step is to find out whether this source is also seen in other wavebands. For that, one initially uses all-sky surveys in these other wavebands and checks whether in them a source is seen at the same sky position. Optical all-sky surveys played an important role in the development of astronomy. The first optical Northern sky survey, the Bonner Durchmusterung, was carried out by Friedrich Wilhelm Argelander between 1852 and 1862 in Bonn, well before photographic plates became available for astronomical observations. It contains some 325 000 stars brighter than $m \approx 9.5$ and was later extended as Cordoba Survey to the Southern sky.

The Palomar Observatory Sky Survey (POSS) is a photographic atlas of the Northern ($\delta > -30°$) sky. It consists of 879 pairs of photoplates observed in two color bands and was completed in 1960. The coverage of the Southern part of the sky was completed in 1980 in the ESO/SERC Southern Sky Surveys, where this survey is about two magnitudes deeper ($B \lesssim 23$, $R \lesssim 22$) than POSS. Later, the photoplates from both surveys were digitized, forming the Digitized Sky Survey (DSS) that covers the full sky. Sections from the DSS can be obtained directly via the Internet, with the full DSS having a data volume of some 600 GB. Using photographic plates with finer grain and higher sensitivity, the second Palomar Sky Survey (POSS-II) was carried out in the 1980s and 1990s. It is about one magnitude deeper compared to the first one and consists of images in three (instead of two) color filters. This will probably be the last photographic atlas of the sky because, with the development of large CCD mosaic cameras, we are now able to perform such surveys digitally.

The first digital optical survey which covers a substantial fraction of the sky is the Sloan Digital Sky Survey (SDSS). Its first phase was carried out between 2000 and 2008 with a dedicated 2.5 m telescope at Apache Point Observatory in New Mexico, equipped with two instruments. The first is a camera with 30 CCDs which scanned nearly a quarter of the sky in five photometric bands. The amount of data collected in this survey is enormous, and its storage and reduction required a tremendous effort. For this photometric part of the Sloan Survey, a new photometric system was developed, with its five filters (u, g, r, i, z) chosen such that their transmission curves overlap as little as possible (see Appendix A.4). The second instrument is a multi-object spectrograph, using optical fibers which have to be manually installed in holes that had been punched into a metal plate. With it, about 640 spectra could be recorded simultaneously. Within the SDSS, the spectra of more than a million objects (mostly galaxies and quasars) were recorded, and the data products of the SDSS were made publicly available in a sequence of seven data releases. The scientific yield of the SDSS has been enormous, well beyond the scientific purpose which formed the prime motivation for carrying out the survey in the first place, namely to measure the large-scale structure of the Universe (see Sect. 8.1.2). The telescope and its instruments are currently used for a number of additional surveys.

All-sky surveys have been carried out in other wavebands as well. We mentioned before the all-sky surveys carried out by several satellites, e.g., ROSAT in the X-ray regime, IRAS, AKARI and WISE in the infrared, and COBE, WMAP and Planck in the microwave regime. A ground-based survey in the near-IR was carried out in the 1990s and released to the public in 2002. This Two Micron All Sky Survey (2MASS) imaged the whole sky in three near-IR bands (J, H and K_s). More than half a billion stars were detected, as well as about 1.6 million resolved sources, of which more than 98 % are galaxies and which are published in the Extended Source Catalog (XSC). This catalog is more than 90 % complete at Galactic latitudes $|b| > 20°$ for sources with diameters $\gtrsim 10''$ and magnitudes $K_s \leq 13.5$. One of the prime science drivers was to penetrate the dust of the Milky Way—at near-IR wavelength, the opacity is only one tenth of that of visible radiation. Therefore, these near-IR images allow us to see galaxies much closer to the Galactic disk than possible in optical images (see Fig. 1.52). Hence, 2MASS has given us the first rather complete map of galaxies in the nearby ($z \lesssim 0.1$) Universe.

The Leiden-Argentine-Bonn (LAB) survey covers the whole sky in the 21-cm emission line of neutral hydrogen, within a velocity range of ± 400 km/s. With its angular resolution of $\sim 36'$ and velocity resolution of 1.3 km/s it maps the neutral hydrogen in and around our Milky Way. Other surveys at radio frequencies include the NRAO VLA Sky Survey (NVSS), which covers the sky North of $-40°$ declination at 1.4 GHz, and the Bonn 408-MHz All-Sky Survey.

Deep multi-waveband Surveys. In order to understand the properties of galaxies and QSOs, data from a large range of wavelengths are required. Accounting for that, surveys are conducted on selected regions in the sky where several observatories make a coordinated effort to obtain a multi-wavelength data set. Perhaps the best-know example, and kind of a prototype, is the Hubble Deep Field, mentioned before and discussed in more detail in Sect. 9.2.1, where the deep HST data were soon supplemented by deep observations over all wavelengths, from the radio to the gamma-ray regime. Encouraged by the success of the scientific exploration of the HDF, several additional such surveys were carried out, including the Hubble Ultradeep Field (HUDF), the Great Observatories Origins Deep Survey (GOODS), the Galaxy Evolution from Morphology and Spectral Energy Distributions (GEMS) and the Cosmological Evolution Survey (COSMOS), with a sky coverage between that of the field-of-view of the ACS camera (HUDF) and nearly two square degrees (COSMOS). In all cases, a large number of other observatories acquired deep images in the same sky region to obtain the broadest waveband coverage possible, in addition to extended spectroscopic campaigns.

These deep multi-waveband surveys have yielded a large suite of results, mainly for distant objects. For each source in the survey field, the broad-band energy distribution is available and can be used to interpret the physical nature of the object—for example, normal galaxies can be distinguished from galaxies with weak nuclear activity which may show up in X-rays only. Unobscured star formation can be detected in the UV-part of the spectrum, whereas star formation hidden by dust shows up at infrared wavelengths. These surveys are also used for statistical properties of the galaxy population—for example, one can study the fraction of optically bright galaxies which show signs of nuclear activity in X-rays or in the radio. Using several optical and infrared bands, the redshift of an object can be estimated rather accurately, using the method of photometric redshifts (see Sect. 9.1.2). To end this highly incomplete list, one can search for rare objects in such surveys—for example those which are well detected in several near-IR bands, but show no sign of optical flux. As we will learn lateron (see Sect. 9.2.4), such an object is a very strong candidate for a galaxy with very high redshift.

Data bases and Virtual Observatories. Many of the surveys are designed to be of versatile use for many different research fields. In order for other astronomers to make use of the data, they have to be stored in well-organized data bases where the requested information can be readily searched for. But not only surveys are a valuable source of information for a wide range of applications; targeted observations of

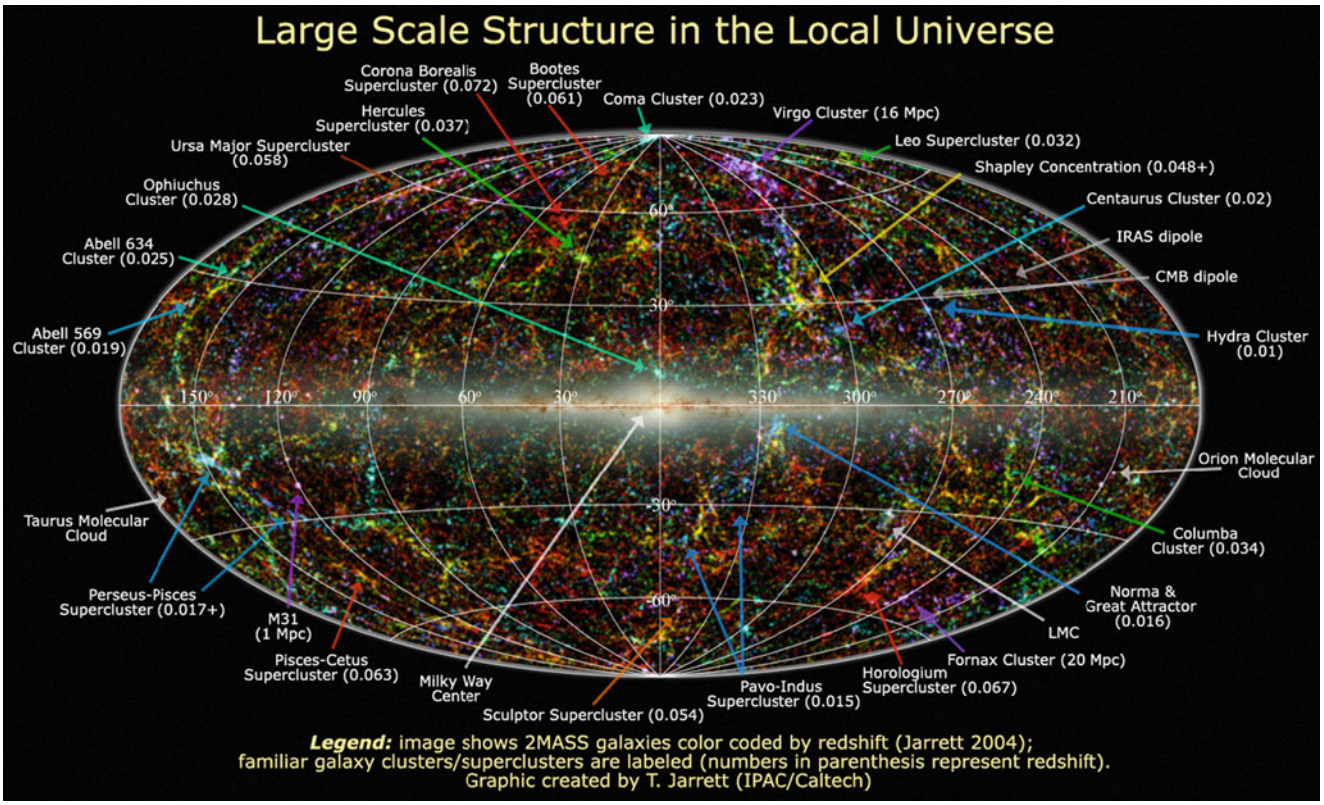


Fig. 1.52 A view of the full sky in near-IR light, as obtained from the 2MASS Extended Source Catalog (XSC). Color codes the distance of the galaxies, obtained from spectroscopic redshifts of $\sim 10^5$ galaxies, and estimated from their apparent brightness for the other $>10^6$ objects: *blue* indicates nearby galaxies (with $z < 0.01$), *green* ($0.01 < z < 0.04$) and *red* ($0.04 < z < 0.1$) show increasingly more distant galaxies. In addition, the intensity is related to the brightness of objects to enhance the contrast. The map is shown in Galactic coordinates, such that the Galactic center is at the center of the image—there, the Milky Way is intransparent even to near-IR radiation, and thus the distribution of stars as observed in 2MASS is shown. Several of the most obvious structures are labeled. In particular one sees that the distribution of galaxies is highly structured. Credit: T. Jarrett, IPAC/Caltech

particular objects yield data sets which may be of interest to researchers other than the proposer of that observations. For space observatories, it has become standard for a long time that the data are stored in archives, and that every astronomer has access to the data in these archives, usually with a proprietary period of 1 year in which only the proposer has this access, to allow her to scientifically exploit the data and publish the results. Also several major ground-based facilities operate in this manner, where data (science plus calibration data) are made available to the community via dedicated archives.

In order to make the different data archives mutually compatible, new standards for data formats and storage systems are being developed. The goal of these virtual observatory initiatives is to develop a common platform in which astronomers can get with little effort a multi-wavelength image of selected regions in the sky, where the data at different wavelengths are stored in different archives spread around the world. The data included in a virtual observatory is not restricted only to observations, but can include simulation results as well. For example, results from cosmological simulations of the large-scale matter distribution in the Universe can be transformed into a mock sky map of the simulated sources, which can be analyzed in the same way as real data to allow for comparisons between observations and model predictions. In this way, one aims for making optimal use of valuable (and expensive) data.

1.5 Problems

1.1. Age of the Universe. Based on the Hubble law (1.2), we can get a simple first estimate of the age of the Universe. Consider a galaxy at distance D whose radial velocity is given by (1.2), and assume that this velocity was the same throughout cosmic time. In this case, at some time in the past the separation was zero, and we can identify that instant as the Big Bang. Under these assumptions, calculate the current age of the Universe using (1.7). Does it depend on the choice of the galaxy, i.e., the current distance D? Compare your result with the age of the oldest stars found in our Galaxy,

which is about 12×10^9 yr. Since no signal can propagate faster than the speed of light c, the age of the Universe times the speed of light is often called the 'size of the visible Universe'. How large is that?

1.2. Sky fraction filled with nearby galaxies. The mean number density of luminous galaxies in the local Universe is about $2 \times 10^{-2}\, h^3\,\mathrm{Mpc}^{-3}$. Assume that they are uniformly distributed, and that the diameter of their luminous region is about 20 kpc, comparable to that of the Milky Way. How many of these galaxies are contained in a sphere with radius $r_0 = 1\, h^{-1}$ Gpc around us? How many of these galaxies will be seen per square degree on the sky? What is the fraction of the sky which luminous galaxies within a distance of r_0 subtend?

1.3. Density of the Universe. Our Universe has a mean matter density ρ_m given by (1.10), with $\Omega_\mathrm{m} \approx 0.3$. About 15 % of this mass density in contributed by baryonic matter, i.e., protons, neutrons and electrons, yielding a baryonic mass density of $\rho_\mathrm{b} \approx 0.15\, \rho_\mathrm{m}$. Use $h \approx 0.71$, or $h^2 \approx 1/2$ for this exercise.

1. The closest star to the Sun has a distance slightly larger than 1 pc, so that we can estimate the local mass density to be $\rho_{*\mathrm{local}} \approx 1 M_\odot\,\mathrm{pc}^{-3}$. Compare this value with ρ_b.
2. From the rotational velocity $V_0 \approx 220\,\mathrm{km/s}$ of the Sun around the center of the Milky Way, we obtain the mass $M(R_0)$ of the Milky Way contained in a sphere of radius $R_0 \approx 8\,\mathrm{kpc}$ from the law (1.1) of Kepler rotation. By which factor is the mean density inside R_0 larger than ρ_m?
3. If you place a cube of 1 m side-length at a random point in the Universe, how many baryons do you expect to find in it on average?

1.4. Free-fall time. Consider a sphere of mass M and initial radius r_0. If there is no pressure acting against gravity, and if there is no outward-directed motion of the matter in the sphere, the sphere will reduce its radius, and in the idealized case considered here, collapse to a single point in a finite time. According to Newton's law of gravity, the radius evolves in time obeying the equation of motion

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} = -\frac{GM}{r^2}\,.$$

The solution of this equation depends on the initial velocity, as well the initial radius, and requires some algebra. However, a simple solution can be obtained by the ansatz $r(t) = r_0\,(1 - t/t_\mathrm{f})^\alpha$, where $r(t = 0) = r_0$.

1. Show that this ansatz leads indeed to a solution of the equation of motion, and determine the two parameters α and t_f. Describe the qualitative behavior of the solution $r(t)$.
2. Show that the time-scale t_f, i.e., the time it takes the sphere to collapse to a point, depends only on the mean initial density of the sphere. What is this time-scale for a density corresponding to the mean density inside the inner 8 kpc of our Galaxy? Hint: Make use of the fact that you know the orbital time of the Sun around the center of the Milky Way, $t_\mathrm{orb} \approx 2.3 \times 10^8$ yr.
3. What is the time-scale t_f if the density is the mean density of the current Universe? How does this compare to the age of the Universe estimated in Problem 1.1? What is t_f for the Einstein–de Sitter model with density (1.14), and how does this compare to the current age of the Universe in the EdS model, given by (1.13)? Can you interpret your finding?

The Milky Way as a galaxy

2

The Earth is orbiting around the Sun, which itself is orbiting around the center of the Milky Way. Our Milky Way, the Galaxy, is the only galaxy in which we are able to study astrophysical processes in detail. Therefore, our journey through extragalactic astronomy will begin in our home Galaxy, with which we first need to become familiar before we are ready to take off into the depths of the Universe. Knowing the properties of the Milky Way is indispensable for understanding other galaxies.

2.1 Galactic coordinates

On a clear night, and sufficiently far away from cities, one can see the magnificent band of the Milky Way on the sky (Fig. 2.1). This observation suggests that the distribution of light, i.e., that of the stars in the Galaxy is predominantly that of a thin disk, as is also clearly seen in Fig. 1.52. A detailed analysis of the geometry of the distribution of stars and gas confirms this impression. This geometry of the Galaxy suggests the introduction of two specially adapted coordinate systems which are particularly convenient for quantitative descriptions.

Spherical Galactic coordinates (ℓ, b). We consider a spherical coordinate system, with its center being "here", at the location of the Sun (see Fig. 2.2). The *Galactic plane* is the plane of the Galactic disk, i.e., it is parallel to the band of the Milky Way. The two *Galactic coordinates* ℓ and b are angular coordinates on the sphere. Here, b denotes the *Galactic latitude*, the angular distance of a source from the Galactic plane, with $b \in [-90°, +90°]$. The great circle $b = 0°$ is then located in the plane of the Galactic disk. The direction $b = 90°$ is perpendicular to the disk and denotes the North Galactic Pole (NGP), while $b = -90°$ marks the direction to the South Galactic Pole (SGP). The second angular coordinate is the *Galactic longitude* ℓ, with $\ell \in [0°, 360°]$. It measures the angular separation between the position of a source, projected perpendicularly onto the Galactic disk (see Fig. 2.2), and the Galactic center, which itself has angular coordinates $b = 0°$ and $\ell = 0°$. Given ℓ and b for a source, its location on the sky is fully specified. In order to specify its three-dimensional location, the distance of that source from us is also needed.

The conversion of the positions of sources given in Galactic coordinates (b, ℓ) to that in equatorial coordinates

Fig. 2.1 An unusual optical image of the Milky Way: This total view of the Galaxy is composed of a large number of individual images. Credit: Stephan Messner

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_2,

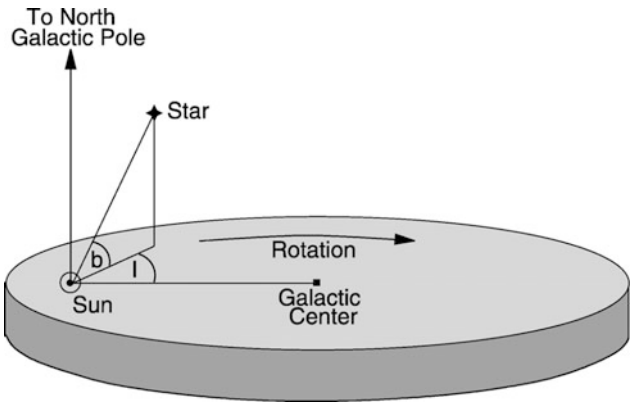


Fig. 2.2 The Sun is at the origin of the Galactic coordinate system. The directions to the Galactic center and to the North Galactic Pole (NGP) are indicated and are located at $\ell = 0°$ and $b = 0°$, and at $b = 90°$, respectively. Adopted from: B.W. Carroll & D.A. Ostlie 1996, *Introduction to Modern Astrophysics*, Addison-Wesley

(α, δ) and vice versa is obtained from the rotation between these two coordinate systems, and is described by spherical trigonometry.[1] The necessary formulae can be found in numerous standard texts. We will not reproduce them here, since nowadays this transformation is done nearly exclusively using computer programs. Instead, we will give some examples. The following figures refer to the Epoch 2000: due to the precession of the rotation axis of the Earth, the equatorial coordinate system changes with time, and is updated from time to time. The position of the Galactic center (at $\ell = 0° = b$) is $\alpha = 17^{\rm h}45.6^{\rm m}$, $\delta = -28°56'.2$ in equatorial coordinates. This immediately implies that at the La Silla Observatory, located at geographic latitude $-29°$, the Galactic center is found near the zenith at local midnight in May/June. Because of the high stellar density in the Galactic disk and the large extinction due to dust this is therefore not a good season for extragalactic observations from La Silla. The North Galactic Pole has coordinates $\alpha_{\rm NGP} = 192.85948° \approx 12^{\rm h}51^{\rm m}$, $\delta_{\rm NGP} = 27.12825° \approx 27°7'.7$.

Zone of Avoidance. As already mentioned, the absorption by dust and the presence of numerous bright stars render optical observations of extragalactic sources in the direction of the disk difficult. The best observing conditions are found at large $|b|$, while it is very hard to do extragalactic astronomy in the optical regime at $|b| \lesssim 10°$; this region is therefore often called the 'Zone of Avoidance'. An illustrative example is the galaxy Dwingeloo 1, which was already mentioned in Sect. 1.1 (see Fig. 1.9). This galaxy was only discovered in the 1990s despite being in our immediate vicinity: it is located at low $|b|$, right in the Zone of Avoidance. As mentioned before, one of the prime motivations for carrying out the 2MASS survey (see Sect. 1.4) was to 'peek' through the dust in the Zone of Avoidance by observing in the near-IR bands.

Cylindrical Galactic coordinates (R, θ, z). The angular coordinates introduced above are well suited to describing the angular position of a source relative to the Galactic disk. However, we will now introduce another three-dimensional coordinate system for the description of the Milky Way geometry that will prove very convenient in the study of its kinematic and dynamic properties. It is a cylindrical coordinate system, with the Galactic center at the origin (see also Fig. 2.22 below). The radial coordinate R measures the distance of an object from the Galactic center in the disk, and z specifies the height above the disk (objects with negative z are thus located below the Galactic disk, i.e., south of it). For instance, the Sun has a distance from the Galactic center of $R = R_0 \approx 8\,{\rm kpc}$. The angle θ specifies the angular separation of an object in the disk relative to the position of the Sun, as seen from the Galactic center. The distance of an object with coordinates R, θ, z from the Galactic center is then $\sqrt{R^2 + z^2}$, independent of θ. If the matter distribution in the Milky Way was axially symmetric, the density would then depend only on R and z, but not on θ. Since this assumption is a good approximation, this coordinate system is very well suited for the physical description of the Galaxy.

2.2 Determination of distances within our Galaxy

A central problem in astronomy is the estimation of distances. The position of sources on the sphere gives us a two-dimensional picture. To obtain three-dimensional information, measurements of distances are required. We need to know the distance to a source if we want to draw conclusions about its physical parameters. For example, we can directly observe the angular diameter of an object, but to derive the physical size we need to know its distance. Another example is the determination of the *luminosity* L of a source, which can be derived from the observed *flux* S only by means of its distance D, using

$$\boxed{L = 4\pi S\, D^2.} \qquad (2.1)$$

It is useful to consider the dimensions of the physical parameters in this equation. The unit of the luminosity is $[L] = {\rm erg\,s^{-1}}$, and that of the flux $[S] = {\rm erg\,s^{-1}\,cm^{-2}}$. The flux is the energy passing through a unit area per unit time

[1] The equatorial coordinates are defined by the direction of the Earth's rotation axis and by the rotation of the Earth. The intersections of the Earth's axis and the sphere define the northern and southern poles. The great circles on the sphere through these two poles, the meridians, are curves of constant *right ascension* α. Curves perpendicular to them and parallel to the projection of the Earth's equator onto the sky are curves of constant *declination* δ, with the poles located at $\delta = \pm 90°$.

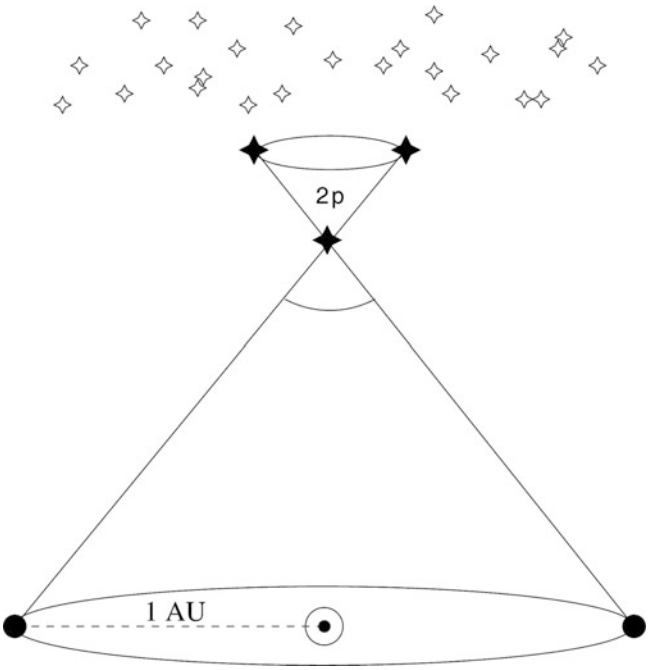


Fig. 2.3 Illustration of the parallax effect: in the course of the Earth's orbit around the Sun the apparent positions of nearby stars on the sky seem to change relative to those of very distant background sources

(see Appendix A). Of course, the physical properties of a source are characterized by the luminosity L and not by the flux S, which depends on its distance from the Sun.

Here we will review various methods for the estimation of distances of objects in our Milky Way, postponing the discussion of methods for estimating extragalactic distances to Sect. 3.9.

2.2.1 Trigonometric parallax

The most important method of distance determination is the *trigonometric parallax*, not only from a historical point-of-view. This method is based on a purely geometric effect and is therefore independent of any physical assumptions. Due to the motion of the Earth around the Sun the positions of nearby stars on the sphere change relative to those of very distant sources (e.g., extragalactic objects such as quasars). The latter therefore define a fixed reference frame on the sphere (see Fig. 2.3). In the course of a year the apparent position of a nearby star follows an ellipse on the sphere, the semi-major axis of which is called the *parallax* p.[2] The axis ratio of this ellipse depends on the direction of the star relative to the ecliptic (the plane that is defined by the orbits of the Earth and the other planets) and is of no further interest here. The parallax depends on the radius r of the Earth's orbit, hence on the Earth-Sun distance which is, by definition, one astronomical unit.[3] Furthermore, the parallax depends on the distance D of the star,

$$\frac{r}{D} = \tan p \approx p \,, \tag{2.2}$$

where we used $p \ll 1$ in the last step, and p is measured in radians as usual. The trigonometric parallax is also used to define the common unit of distance in astronomy: one *parsec* (pc) is the distance of a hypothetical source for which the parallax is exactly $p = 1''$. With the conversion of arcseconds to radians ($1'' \approx 4.848 \times 10^{-6}$ radians) one gets $p/1'' = 206265p$, which for a parsec yields

$$\boxed{1\mathrm{pc} = 206265\mathrm{AU} = 3.086 \times 10^{18}\mathrm{cm}} \,. \tag{2.3}$$

The distance corresponding to a measured parallax is then calculated as

$$\boxed{D = \left(\frac{p}{1''}\right)^{-1} \mathrm{pc}} \,. \tag{2.4}$$

To determine the parallax p, precise measurements of the position of an object at different times are needed, spread over a year, allowing us to measure the ellipse drawn on the sphere by the object's apparent position. For ground-based observations the accuracy of this method is limited by the atmosphere. The seeing causes a blurring of the images of astronomical sources and thus limits the accuracy of position measurements. From the ground this method is therefore limited to parallaxes larger than $\approx 0''.01$, implying that the trigonometric parallax yields distances to stars only within ~ 30 pc.

An extension of this method towards smaller p, and thus larger distances, became possible with the astrometric satellite Hipparcos. It operated between November 1989 and March 1993 and measured the positions and trigonometric parallaxes of about 120 000 bright stars, with a precision of $\sim 0''.001$ for the brighter targets. With Hipparcos the method of trigonometric parallax could be extended to stars up to distances of ~ 300 pc. The satellite Gaia, the successor mission to Hipparcos, was launched on Dec. 19, 2013. Gaia will compile a catalog of $\sim 10^9$ stars up to $V \approx 20$ in four broad-band and eleven narrow-band filters. It will measure parallaxes for these stars with an accuracy of $\sim 2 \times 10^{-4}$ arcsec, and a considerably better accuracy for the brightest stars. Gaia will thus determine the distances for $\sim 2 \times 10^8$ stars with a precision of 10 %, and tangential velocities (see next section) with a precision of better than 3 km/s.

[2] In general, since the star also has a spatial velocity different from that of the Sun, the ellipse is superposed on a linear track on the sky; this linear motion is called *proper motion* and will be discussed below.

[3] To be precise, the Earth's orbit is an ellipse, and one astronomical unit is its semi-major axis, being $1\,\mathrm{AU} = 1.496 \times 10^{13}$ cm.

The trigonometric parallax method forms the basis of nearly all distance determinations owing to its purely geometrical nature. For example, using this method the distances to nearby stars have been determined, allowing the production of the Hertzsprung–Russell diagram (see Appendix B.2). Hence, all distance measures that are based on the properties of stars, such as will be described below, are calibrated by the trigonometric parallax.

2.2.2 Proper motions

Stars are moving relative to us or, more precisely, relative to the Sun. To study the kinematics of the Milky Way we need to be able to measure the velocities of stars. The radial component v_r of the velocity is easily obtained from the Doppler shift of spectral lines,

$$\boxed{v_\mathrm{r} = \frac{\Delta\lambda}{\lambda_0}\, c}\,, \tag{2.5}$$

where λ_0 is the rest-frame wavelength of an atomic transition and $\Delta\lambda = \lambda_\mathrm{obs} - \lambda_0$ the Doppler shift of the wavelength due to the radial velocity of the source. The sign of the radial velocity is defined such that $v_\mathrm{r} > 0$ corresponds to a motion away from us, i.e., to a redshift of spectral lines.

In contrast, the determination of the other two velocity components is much more difficult. The tangential component, v_t, of the velocity can be obtained from the *proper motion* of an object. In addition to the motion caused by the parallax, stars also change their positions on the sphere as a function of time because of the transverse component of their velocity relative to the Sun. The proper motion μ is thus an angular velocity, e.g., measured in milliarcseconds per year (mas/yr). This angular velocity is linked to the tangential velocity component via

$$\boxed{v_\mathrm{t} = D\mu \quad \text{or} \quad \frac{v_\mathrm{t}}{\mathrm{km/s}} = 4.74 \left(\frac{D}{1\,\mathrm{pc}}\right)\left(\frac{\mu}{1''/\mathrm{yr}}\right)}\,. \tag{2.6}$$

Therefore, one can calculate the tangential velocity from the proper motion and the distance. If the latter is derived from the trigonometric parallax, (2.6) and (2.4) can be combined to yield

$$\frac{v_\mathrm{t}}{\mathrm{km/s}} = 4.74 \left(\frac{\mu}{1''/\mathrm{yr}}\right)\left(\frac{p}{1''}\right)^{-1}\,. \tag{2.7}$$

Hipparcos measured proper motions for $\sim 10^5$ stars with an accuracy of up to a few mas/yr; however, they can be translated into physical velocities only if their distance is known.

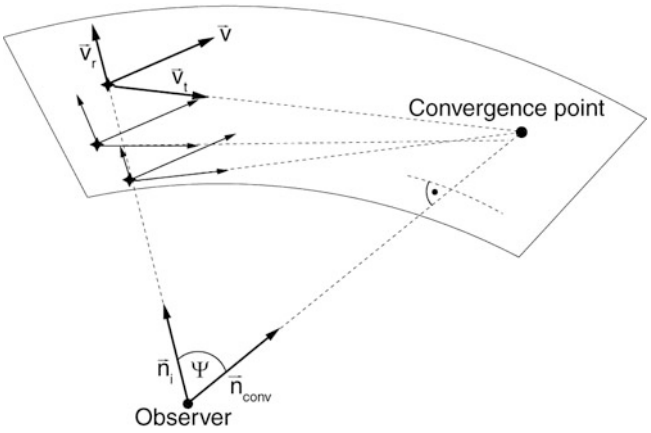


Fig. 2.4 The moving cluster parallax is a projection effect, similar to that known from viewing railway tracks. The directions of velocity vectors pointing away from us seem to converge and intersect at the convergence point. The connecting line from the observer to the convergence point is parallel to the velocity vector of the star cluster

Of course, the proper motion has two components, corresponding to the absolute value of the angular velocity and its direction on the sphere. Together with v_r this determines the three-dimensional velocity vector. Correcting for the known velocity of the Earth around the Sun, one can then compute the velocity vector $\boldsymbol{v}$ of the star relative to the Sun, called the *heliocentric velocity*.

2.2.3 Moving cluster parallax

The stars in an (open) star cluster all have a very similar spatial velocity. This implies that their proper motion vectors should be similar. To what accuracy the proper motions are aligned depends on the angular extent of the star cluster on the sphere. Like two railway tracks that run parallel but do not appear parallel to us, the vectors of proper motions in a star cluster also do not appear parallel. They are directed towards a convergence point, as depicted in Fig. 2.4. We shall demonstrate next how to use this effect to determine the distance to a star cluster.

We consider a star cluster and assume that all stars have the same spatial velocity $\boldsymbol{v}$. The position of the i-th star as a function of time is then described by

$$\boldsymbol{r}_i(t) = \boldsymbol{r}_i + \boldsymbol{v}t\,, \tag{2.8}$$

where $\boldsymbol{r}_i$ is the current position if we identify the origin of time, $t = 0$, with 'today'. The direction of a star relative to us is described by the unit vector

$$\boldsymbol{n}_i(t) := \frac{\boldsymbol{r}_i(t)}{|\boldsymbol{r}_i(t)|}\,. \tag{2.9}$$

From this, one infers that for large times, $t \to \infty$, the direction vectors are identical for all stars in the cluster,

$$\boldsymbol{n}_i(t) \to \frac{\boldsymbol{v}}{|\boldsymbol{v}|} =: \boldsymbol{n}_\mathrm{conv}\,. \tag{2.10}$$

Hence for large times all stars will appear at the same point $\boldsymbol{n}_{\rm conv}$: the convergence point. This only depends on the direction of the velocity vector of the star cluster. In other words, the direction vector of the stars is such that they are all moving towards the convergence point. Thus, $\boldsymbol{n}_{\rm conv}$ (and hence $\boldsymbol{v}/|\boldsymbol{v}|$) can be measured from the direction of the proper motions of the stars in the cluster. On the other hand, one component of $\boldsymbol{v}$ can be determined from the (easily measured) radial velocity $v_{\rm r}$. With these two observables the three-dimensional velocity vector $\boldsymbol{v}$ is completely determined, as is easily demonstrated: let ψ be the angle between the line-of-sight $\boldsymbol{n}$ towards a star in the cluster and $\boldsymbol{v}$. The angle ψ is directly read off from the direction vector $\boldsymbol{n}$ and the convergence point, $\cos\psi = \boldsymbol{n}\cdot\boldsymbol{v}/|\boldsymbol{v}| = \boldsymbol{n}_{\rm conv}\cdot\boldsymbol{n}$. With $v \equiv |\boldsymbol{v}|$ one then obtains

$$v_{\rm r} = v\cos\psi \quad , \quad v_{\rm t} = v\sin\psi \ ,$$

and so

$$v_{\rm t} = v_{\rm r}\tan\psi \ . \tag{2.11}$$

This means that the tangential velocity $v_{\rm t}$ can be measured without determining the distance to the stars in the cluster. On the other hand, (2.6) defines a relation between the proper motion, the distance, and $v_{\rm t}$. Hence, a distance determination for the star is now possible with

$$\mu = \frac{v_{\rm t}}{D} = \frac{v_{\rm r}\tan\psi}{D} \quad \rightarrow \quad D = \frac{v_{\rm r}\tan\psi}{\mu} \ . \tag{2.12}$$

This method yields accurate distance estimates of star clusters within $\sim$ 200 pc. The accuracy depends on the measurability of the proper motions. Furthermore, the cluster should cover a sufficiently large area on the sky for the convergence point to be well defined. For the distance estimate, one can then take the average over a large number of stars in the cluster if one assumes that the spatial extent of the cluster is much smaller than its distance to us. Targets for applying this method are the Hyades, a cluster of about 200 stars at a mean distance of $D \approx 45$ pc, the Ursa-Major group of about 60 stars at $D \approx 24$ pc, and the Pleiades with about 600 stars at $D \approx 130$ pc.

Historically the distance determination to the Hyades, using the moving cluster parallax, was extremely important because it defined the scale to all other, larger distances. Its constituent stars of known distance are used to construct a calibrated Hertzsprung–Russell diagram which forms the basis for determining the distance to other star clusters, as will be discussed in Sect. 2.2.4. In other words, it is the lowest rung of the so-called distance ladder that we will discuss in Sect. 3.9. With Hipparcos, however, the distance to the Hyades stars could also be measured using the trigonometric parallax, yielding more accurate values. Hipparcos was even able to differentiate the 'near' from the 'far' side of the cluster—this star cluster is too close for the assumption of an approximately equal distance of all its stars to be still valid. A recent value for the mean distance of the Hyades is

$$\bar{D}_{\rm Hyades} = 46.3 \pm 0.3{\rm pc} \ . \tag{2.13}$$

2.2.4 Photometric distance; extinction and reddening

Most stars in the color-magnitude diagram are located along the main sequence. This enables us to compile a calibrated main sequence of those stars whose trigonometric parallaxes are measured, thus with known distances. Utilizing photometric methods, it is then possible to derive the distance to a star cluster, as we will demonstrate in the following.

The stars of a star cluster define their own main sequence (color-magnitude diagrams for some star clusters are displayed in Fig. 2.5); since they are all located at the same distance, their main sequence is already defined in a color-magnitude diagram in which only apparent magnitudes are plotted. This cluster main sequence can then be fitted to a calibrated main sequence[4] by a suitable choice of the distance, i.e., by adjusting the distance modulus $m - M$,

$$m - M = 5\log(D/{\rm pc}) - 5 \ ,$$

where m and M denote the apparent and absolute magnitude, respectively.

In reality this method cannot be applied so easily since the position of a star on the main sequence does not only depend on its mass but also on its age and metallicity. Furthermore, only stars of luminosity class V (i.e., dwarf stars) define the main sequence, but without spectroscopic data it is not possible to determine the luminosity class.

Extinction and reddening. Another major problem is extinction. Absorption and scattering of light by dust affect the relation of absolute to apparent magnitude: for a given M, the apparent magnitude m becomes larger (fainter) in the case of absorption, making the source appear dimmer. Also, since extinction depends on wavelength, the spectral energy distribution of the source is modified and the observed color of the star changes. Because extinction by dust is always associated with such a change in color, one can estimate the absorption—provided one has sufficient information on the intrinsic color of a source or of an ensemble of sources. We will now show how this method can be used to estimate the distance to a star cluster.

We consider the equation of radiative transfer for pure absorption or scattering (see Appendix A),

$$\boxed{\frac{{\rm d}I_\nu}{{\rm d}s} = -\kappa_\nu I_\nu} \ , \tag{2.14}$$

where I_ν denotes the specific intensity at frequency ν, κ_ν the absorption coefficient, and s the distance coordinate along

[4] i.e., to the main sequence in a color-magnitude diagram in which absolute magnitudes are plotted.

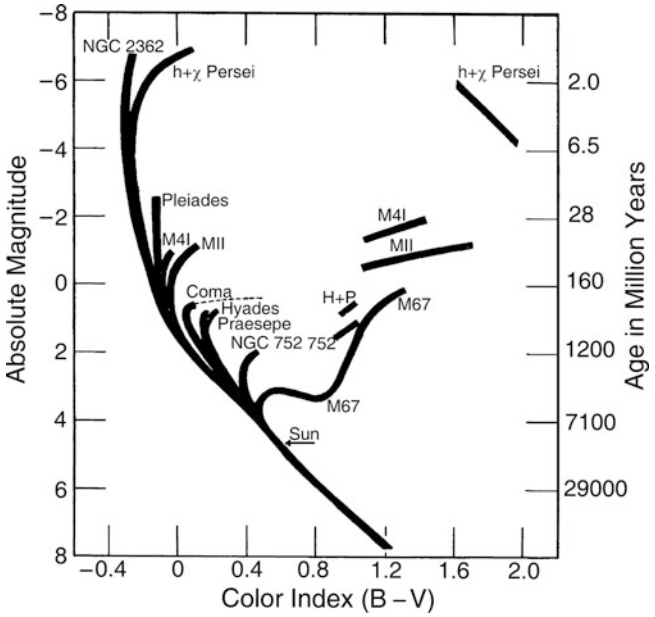


Fig. 2.5 Color-magnitude diagram (CMD) for different star clusters. Such a diagram can be used for the distance determination of star clusters because the absolute magnitudes of main sequence stars are known (by calibration with nearby clusters, especially the Hyades). One can thus determine the distance modulus by vertically 'shifting' the main sequence. Also, the age of a star cluster can be estimated from a CMD: luminous main sequence stars have a shorter lifetime on the main sequence than less luminous ones. The turn-off point in the stellar sequence away from the main sequence therefore corresponds to that stellar mass for which the lifetime on the main sequence equals the age of the star cluster. Accordingly, the age is specified on the right axis as a function of the position of the turn-off point; the Sun will leave the main sequence after about 10×10^9 yr. Credit: Allan Sandage, Carnegie

the light beam. The absorption coefficient has the dimension of an inverse length. Equation (2.14) says that the amount by which the intensity of a light beam is diminished on a path of length $\mathrm{d}s$ is proportional to the original intensity and to the path length $\mathrm{d}s$. The absorption coefficient is thus defined as the constant of proportionality. In other words, on the distance interval $\mathrm{d}s$, a fraction $\kappa_\nu \, \mathrm{d}s$ of all photons at frequency ν is absorbed or scattered out of the beam. The solution of the transport equation (2.14) is obtained by writing it in the form $\mathrm{d} \ln I_\nu = \mathrm{d}I_\nu / I_\nu = -\kappa_\nu \, \mathrm{d}s$ and integrating from 0 to s,

$$\ln I_\nu(s) - \ln I_\nu(0) = -\int_0^s \mathrm{d}s' \, \kappa_\nu(s') \equiv -\tau_\nu(s) \; ,$$

where in the last step we defined the *optical depth*, τ_ν, which depends on frequency. This yields

$$\boxed{I_\nu(s) = I_\nu(0) \, \mathrm{e}^{-\tau_\nu(s)}} \; . \tag{2.15}$$

The specific intensity is thus reduced by a factor $\mathrm{e}^{-\tau}$ compared to the case of no absorption taking place. Accordingly, for the flux we obtain

$$\boxed{S_\nu = S_\nu(0) \, \mathrm{e}^{-\tau_\nu(s)}} \; , \tag{2.16}$$

where S_ν is the flux measured by the observer at a distance s from the source, and $S_\nu(0)$ is the flux of the source without absorption. Because of the relation between flux and magnitude $m = -2.5 \log S + \text{const.}$, or $S \propto 10^{-0.4m}$, one has

$$\frac{S_\nu}{S_{\nu,0}} = 10^{-0.4(m-m_0)} = \mathrm{e}^{-\tau_\nu} = 10^{-\log(\mathrm{e})\tau_\nu} \; ,$$

or

$$\begin{aligned} A_\nu &:= m - m_0 = -2.5 \, \log(S_\nu / S_{\nu,0}) \\ &= 2.5 \, \log(\mathrm{e}) \, \tau_\nu = 1.086 \tau_\nu \; . \end{aligned} \tag{2.17}$$

Here, A_ν is the *extinction coefficient* describing the change of apparent magnitude m compared to that without absorption, m_0. Since the absorption coefficient κ_ν depends on frequency, absorption is always linked to a change in color. This is described by the *color excess* which is defined as follows:

$$\begin{aligned} E(X-Y) &:= A_X - A_Y = (X - X_0) - (Y - Y_0) \\ &= (X - Y) - (X - Y)_0 . \end{aligned} \tag{2.18}$$

The color excess describes the change of the color index $(X - Y)$, measured in two filters X and Y that define the corresponding spectral windows by their transmission curves. The ratio $A_X / A_Y = \tau_{\nu(X)} / \tau_{\nu(Y)}$ depends only on the optical properties of the dust or, more specifically, on the ratio of the absorption coefficients in the two frequency bands X and Y considered here. Thus, the color excess is proportional to the extinction coefficient,

$$\boxed{E(X-Y) = A_X - A_Y = A_X \left(1 - \frac{A_Y}{A_X}\right) \equiv A_X \, R_X^{-1}} \; , \tag{2.19}$$

where in the last step we introduced the factor of proportionality R_X between the extinction coefficient and the color excess, which depends only on the properties of the dust and the choice of the filters. Usually, one considers a blue and a visual filter (see Appendix A.4.2 for a description of the filters commonly used) and writes

$$A_V = R_V \, E(B-V) \; . \tag{2.20}$$

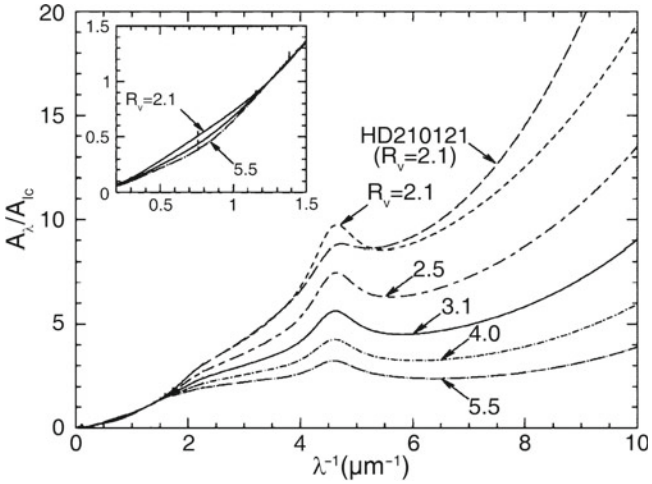


Fig. 2.6 Wavelength dependence of the extinction coefficient A_ν, normalized to the extinction coefficient A_I at $\lambda = 9000\,\text{Å} = 0.9\,\mu\text{m}$. Different kinds of clouds, characterized by the value of R_V, i.e., by the reddening law, are shown. On the x-axis the inverse wavelength is plotted, so that the frequency increases to the right. The *solid curve* specifies the mean Galactic extinction curve. The extinction coefficient, as determined from the observation of an individual star, is also shown; clearly the observed law deviates from the model in some details. The figure *insert* shows a detailed plot at relatively large wavelengths in the NIR range of the spectrum; at these wavelengths the extinction depends only weakly on the value of R_V. Source: B. Draine 2003, *Interstellar Dust Grains*, ARA&A 41, 241. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 41 ©2003 by Annual Reviews www.annualreviews.org

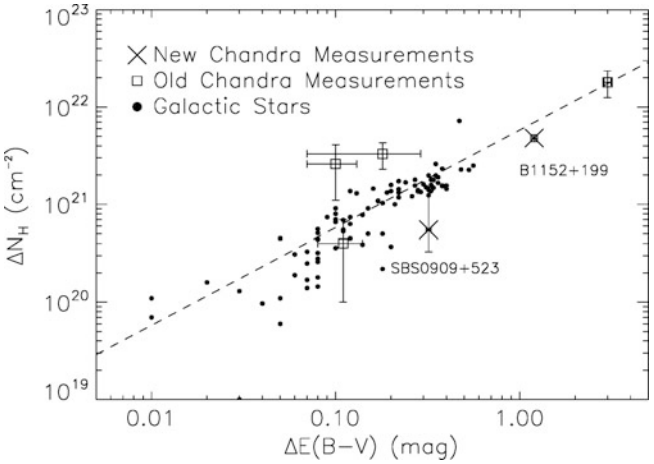


Fig. 2.7 The column density of neutral hydrogen along the line-of-sight to Galactic stars, plotted as a function of the corresponding color excess $E(B-V)$, as shown by the points. The *dashed line* is the best-fitting linear relation as given by (2.22). The *other symbols* correspond to measurements of both quantities in distant galaxies and will be discussed in Sect. 3.11.4. Source: X. Dai & C.S. Kochanek 2009, *Differential X-Ray Absorption and Dust-to-Gas Ratios of the Lens Galaxies SBS 0909+523, FBQS 0951+2635, and B 1152+199*, ApJ 692, 677, p. 682, Fig. 5. ©AAS. Reproduced with permission

For example, for dust in our Milky Way we have the characteristic relation

$$\boxed{A_V = (3.1 \pm 0.1) E(B-V)}\,. \qquad (2.21)$$

This relation is not a universal law, but the factor of proportionality depends on the properties of the dust. They are determined, e.g., by the chemical composition and the size distribution of the dust grains. Figure 2.6 shows the wavelength dependence of the extinction coefficient for different kinds of dust, corresponding to different values of R_V. In the optical part of the spectrum we have approximately $\tau_\nu \propto \nu$, i.e., blue light is absorbed (or scattered) more strongly than red light. The extinction therefore always causes a reddening.[5]

The extinction coefficient A_V is proportional to the optical depth towards a source, see (2.17), and according to (2.21), so is the color excess. Since the extinction is due to dust along the line-of-sight, the color excess is proportional to the column density of dust towards the source. If we assume that the dust-to-gas ratio in the interstellar medium does not vary greatly, we expect that the column density of neutral hydrogen N_H is proportional to the color excess. The former can be measured from the Lyman-α absorption in the spectra of stars, whereas the latter is obtained by comparing the observed color of these stars with the color expected for the type of star, given its spectrum (and thus, its spectral classification). One finds indeed that the color excess is proportional to the HI column density (see Fig. 2.7), with

$$E(B-V) = 1.7\,\text{mag}\left(\frac{N_\text{H}}{10^{22}\,\text{atoms}\,\text{cm}^{-2}}\right)\,, \qquad (2.22)$$

and a scatter of about 30 % around this relation. The fact that this scatter is so small indicates that the assumption of a constant dust-to-gas ratio is reasonable.

In the Solar neighborhood the extinction coefficient for sources in the disk is about

$$A_V \approx 1\,\text{mag}\,\frac{D}{1\,\text{kpc}}\,, \qquad (2.23)$$

but this relation is at best a rough approximation, since the absorption coefficient can show strong local deviations from this law, for instance in the direction of molecular clouds (see, e.g., Fig. 2.8).

Color-color diagram. We now return to the distance determination for a star cluster. As a first step in this measurement, it is necessary to determine the degree of extinction, which can only be done by analyzing the reddening. The stars of the cluster are plotted in a *color-color diagram*,

[5] With what we have just learned we can readily answer the question of why the sky is blue and the setting Sun red.

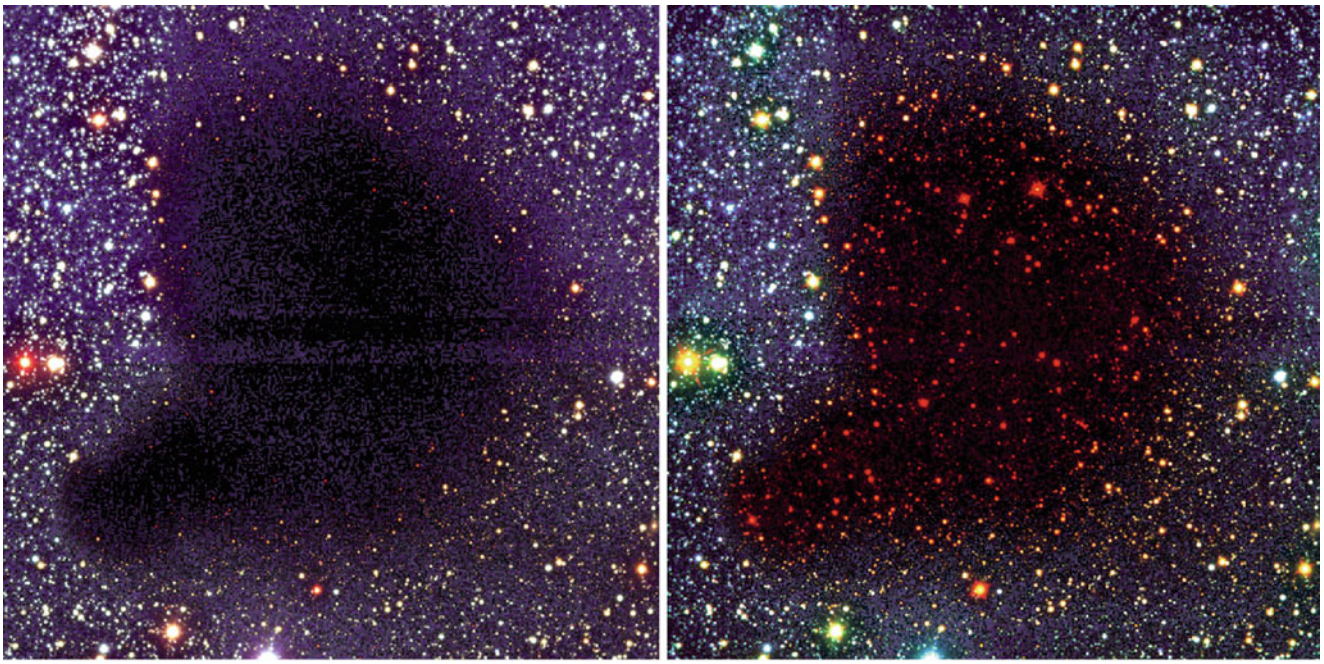


Fig. 2.8 These images of the molecular cloud Barnard 68 show the effects of extinction and reddening: the *left image* is a composite of exposures in the filters B, V, and I. At the center of the cloud essentially all the light from the background stars is absorbed. Near the edge it is dimmed and visibly shifted to the *red*. In the *right-hand image* observations in the filters B, I, and K have been combined (red is assigned here to the near-infrared K-band filter); we can clearly see that the cloud is more transparent at longer wavelengths. Credit: European Southern Observatory

for example by plotting the colors $(U-B)$ and $(B-V)$ on the two axes (see Fig. 2.9). A color-color diagram also shows a main sequence along which the majority of the stars are aligned. The wavelength-dependent extinction causes a reddening *in both colors*. This shifts the positions of the stars in the diagram. The direction of the reddening vector depends only on the properties of the dust and is here assumed to be known, whereas the *amplitude* of the shift depends on the extinction coefficient. In a similar way to the CMD, this amplitude can now be determined if one has access to a calibrated, unreddened main sequence for the color-color diagram which can be obtained from the examination of nearby stars. From the relative shift of the main sequence in the two diagrams one can then derive the reddening and thus the extinction. The essential point here is the fact that *the color-color diagram is independent of the distance.*

This then defines the procedure for the distance determination of a star cluster using photometry: in the first step we determine the reddening $E(B-V)$, and thus with (2.21) also A_V, by shifting the main sequence in a color-color diagram along the reddening vector until it matches a calibrated main sequence. In the second step the distance modulus is determined by vertically (i.e., in the direction of M) shifting the main sequence in the color-magnitude diagram until it matches a calibrated main sequence. From this, the distance is finally obtained according to

$$\boxed{m - M = 5\log(D/1\text{pc}) - 5 + A}\,. \tag{2.24}$$

2.2.5 Spectroscopic distance

From the spectrum of a star, the spectral type as well as its luminosity class can be obtained. The former is determined from the strength of various absorption lines in the spectrum, while the latter is obtained from the width of the lines. From the line width the surface gravity of the star can be derived, and from that its radius (more precisely, M/R^2). The spectral type and the luminosity class specify the position of the star in the HRD unambiguously. By means of stellar evolution models, the absolute magnitude M_V can then be determined. Furthermore, the comparison of the observed color with that expected from theory yields the color excess $E(B-V)$, and from that we obtain A_V. With this information we are then able to determine the distance using

$$\boxed{m_V - A_V - M_V = 5\log(D/\text{pc}) - 5}\,. \tag{2.25}$$

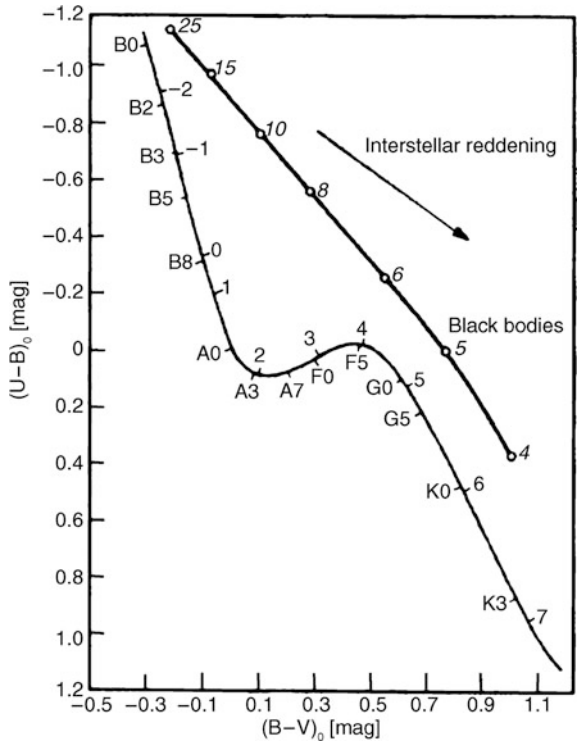


Fig. 2.9 Color-color diagram for main sequence stars. Spectral types and absolute magnitudes are specified along the *lower curve*. The *upper curve* shows the location of black bodies in the color-color diagram, with the temperature in units of 10^3 K labeled along the curve. Interstellar reddening shifts the measured stellar locations parallel to the reddening vector indicated by the *arrow*. Source: A. Unsöld & B. Baschek, *The New Cosmos*, Springer-Verlag

2.2.6 Distances of visual binary stars

Kepler's third law for a two-body problem,

$$\boxed{P^2 = \frac{4\pi^2}{G(m_1 + m_2)} a^3} \,, \tag{2.26}$$

relates the orbital period P of a binary star to the masses m_i of the two components and the semi-major axis a of the ellipse. The latter is defined by the separation vector between the two stars in the course of one period. This law can be used to determine the distance to a visual binary star. For such a system, the period P and the angular diameter 2θ of the orbit are direct observables. If one additionally knows the mass of the two stars, for instance from their spectral classification, a can be determined according to (2.26), and from this the distance follows with $D = a/\theta$.

2.2.7 Distances of pulsating stars

Several types of pulsating stars show periodic changes in their brightnesses, where the period of a star is related to its mass, and thus to its luminosity. This *period-luminosity (PL) relation* is ideally suited for distance measurements: since the determination of the period is independent of distance, one can obtain the luminosity directly from the period if the calibrated PL-relation is known. The distance is thus directly derived from the measured magnitude using (2.25), if the extinction can be determined from color measurements.

The existence of a relation between the luminosity and the pulsation period can be expected from simple physical considerations. Pulsations are essentially radial density waves inside a star that propagate with the speed of sound, c_s. Thus, one can expect that the period is comparable to the sound crossing time through the star, $P \sim R/c_s$. The speed of sound c_s in a gas is of the same order of magnitude as the thermal velocity of the gas particles, so that $k_B T \sim m_p c_s^2$, where m_p is the proton mass (and thus a characteristic mass of particles in the stellar plasma) and k_B is Boltzmann's constant. According to the virial theorem, one expects that the gravitational binding energy of the star is about twice the kinetic (i.e., thermal) energy, so that for a proton

$$\frac{GMm_p}{R} \sim k_B T \,.$$

Combining these relations, we obtain for the pulsation period

$$P \sim \frac{R}{c_s} \sim \frac{R\sqrt{m_p}}{\sqrt{k_B T}} \sim \frac{R^{3/2}}{\sqrt{GM}} \propto \bar{\rho}^{-1/2} \,, \tag{2.27}$$

where $\bar{\rho}$ is the mean density of the star. This is a remarkable result—the pulsation period depends only on the mean density. Furthermore, the stellar luminosity is related to its mass by approximately $L \propto M^3$. If we now consider stars of equal effective temperature $T_{\rm eff}$ (where $L \propto R^2 T_{\rm eff}^4$), we find that

$$P \propto \frac{R^{3/2}}{\sqrt{M}} \propto L^{7/12} \,, \tag{2.28}$$

which is the relation between period and luminosity that we were aiming for.

One finds that a well-defined period-luminosity relation exists for three types of pulsating stars:

- *δ Cepheid stars (classical Cepheids)*. These are young stars found in the disk population (close to the Galactic plane) and in young star clusters. Owing to their position in or near the disk, extinction always plays a role in the determination of their luminosity. To minimize the effect of extinction it is particularly useful to look at the period-luminosity relation in the near-IR (e.g., in the K-band at $\lambda \sim 2.4\,\mu$m). Furthermore, the scatter around the period-luminosity relation is smaller for longer wavelengths of the applied filter, as is also shown in Fig. 2.10. The period-luminosity relation is also steeper for longer wavelengths, resulting in a more accurate determination of the absolute magnitude.
- *W Virginis stars*, also called population II Cepheids (we will explain the term of stellar populations in Sect. 2.3.2). These are low-mass, metal-poor stars located in the halo of the Galaxy, in globular clusters, and near the Galactic center.
- *RR Lyrae stars*. These are likewise population II stars and thus metal-poor. They are found in the halo, in

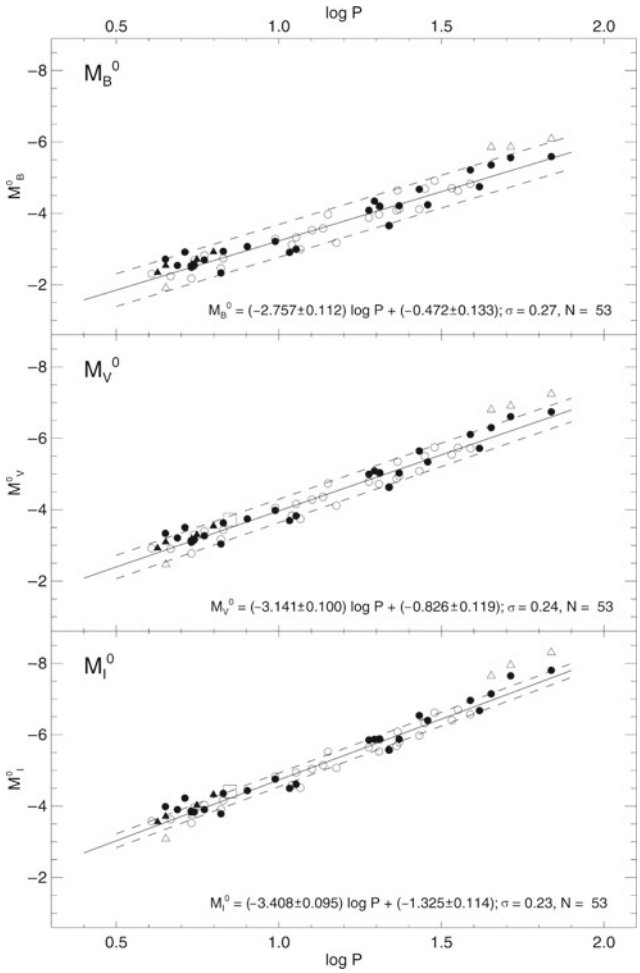


Fig. 2.10 Period-luminosity relation for Galactic Cepheids, measured in three different filters bands (B, V, and I, *from top to bottom*). The absolute magnitudes were corrected for extinction by using colors. The period is given in days. *Open* and *solid circles* denote data for those Cepheids for which distances were estimated using different methods; the three objects marked by *triangles* have a variable period and are discarded in the derivation of the period-luminosity relation. The latter is indicated by the *solid line*, with its parametrization specified in the plots. The *broken lines* indicate the uncertainty range of the period-luminosity relation. The slope of the period-luminosity relation increases, and the dispersion of the individual measurements around the mean PL-relation decreases, if one moves to redder filters. Source: G.A. Tammann et al. 2003, *New Period-Luminosity and Period-Color relations of classical Cepheids: I. Cepheids in the Galaxy*, A&A 404, 423, p. 436, Fig. 11. ©ESO. Reproduced with permission

globular clusters, and in the Galactic bulge. Their absolute magnitudes are confined to a narrow interval, $M_V \in [0.5, 1.0]$, with a mean value of about 0.6. This obviously makes them very good distance indicators. More precise predictions of their magnitudes are possible with the following dependence on metallicity and period:

$$\langle M_K \rangle = -(2.0 \pm 0.3)\log(P/1\mathrm{d}) + (0.06 \pm 0.04)[\mathrm{Fe/H}] - 0.7 \pm 0.1 \,. \tag{2.29}$$

Metallicity. In the last equation, the metallicity of a star was introduced, which needs to be defined. In astrophysics, all chemical elements heavier than helium are called *metals*. These elements, with the exception of some traces of lithium, were not produced in the early Universe but rather later in the interior of stars. The metallicity is thus also a measure of the chemical evolution and enrichment of matter in a star or gas cloud. For an element X, the *metallicity index* of a star is defined as

$$\boxed{[\mathrm{X/H}] \equiv \log\left(\frac{n(\mathrm{X})}{n(\mathrm{H})}\right)_* - \log\left(\frac{n(\mathrm{X})}{n(\mathrm{H})}\right)_\odot} \,, \tag{2.30}$$

thus it is the logarithm of the ratio of the fraction of X relative to hydrogen in the star and in the Sun, where n is the number density of the species considered. For example, $[\mathrm{Fe/H}] = -1$ means that iron has only a tenth of its Solar abundance. The *metallicity* Z is the total mass fraction of all elements heavier than helium; the Sun has $Z \approx 0.02$, meaning that about 98 % of the Solar mass is composed of hydrogen and helium.

The period-luminosity relations are not only of significant importance for distance determinations within our Galaxy. They also play an essential role in extragalactic astronomy, since the Cepheids (which are by far the most luminous of the three types of pulsating stars listed above) are also found and observed outside the Milky Way; they therefore enable us to directly determine the distances of other galaxies, which is essential for measuring the Hubble constant. These aspects will be discussed in detail in Sect. 3.9.

2.3 The structure of the Galaxy

Roughly speaking, the Galaxy consists of the disk, the central bulge, and the Galactic halo—a roughly spherical distribution of stars and globular clusters that surrounds the disk. The disk, whose stars form the visible band of the Milky Way, contains spiral arms similar to those observed in other spiral galaxies. The Sun, together with its planets, orbits around the Galactic center on an approximately circular orbit. The distance R_0 to the Galactic center is not very accurately known, as we will discuss later. To have a reference value, the International Astronomical Union (IAU) officially defined the value of R_0 in 1985,

$$\boxed{R_0 = 8.5\,\mathrm{kpc}} \quad \text{official value, IAU 1985} \,. \tag{2.31}$$

More recent examinations have, however, found that the real value is slightly smaller, $R_0 \approx 8.0\,\mathrm{kpc}$. The diameter of the disk of stars, gas, and dust is $\sim 50\,\mathrm{kpc}$. A schematic depiction of our Galaxy is shown in Fig. 1.6.

2.3.1 The Galactic disk: Distribution of stars

By measuring the distances of stars in the Solar neighborhood one can determine the three-dimensional stellar distribution. From these investigations, one finds that there are different stellar components, as we will discuss below. For each of them, the number density in the direction perpendicular to the Galactic disk is approximately described by an exponential law,

$$n(z) \propto \exp\left(-\frac{|z|}{h}\right) , \tag{2.32}$$

where the *scale-height* h specifies the thickness of the respective component. One finds that h varies between different populations of stars, motivating the definition of different components of the Galactic disk. In principle, three components need to be distinguished: (1) The *young thin disk* contains the largest fraction of gas and dust in the Galaxy, and in this region star formation is still taking place today. The youngest stars are found in the young thin disk, which has a scale-height of about $h_{\rm ytd} \sim 100\,\mathrm{pc}$. (2) The *old thin disk* is thicker and has a scale-height of about $h_{\rm otd} \sim 325\,\mathrm{pc}$. (3) The *thick disk* has a scale-height of $h_{\rm thick} \sim 1.5\,\mathrm{kpc}$. The thick disk contributes only about 2 % to the total mass density in the Galactic plane at $z = 0$. This separation into three disk components is rather coarse and can be further refined if one uses a finer classification of stellar populations.

Molecular gas, out of which new stars are born, has the smallest scale-height, $h_{\rm mol} \sim 65\,\mathrm{pc}$, followed by the atomic gas. This can be clearly seen by comparing the distributions of atomic and molecular hydrogen in Fig. 1.8. The younger a stellar population is, the smaller its scale-height. Another characterization of the different stellar populations can be made with respect to the velocity dispersion of the stars, i.e., the amplitude of the components of their random motions. As a first approximation, the stars in the disk move around the Galactic center on circular orbits. However, these orbits are not perfectly circular: besides the orbital velocity (which is about 220 km/s in the Solar vicinity), they have additional random velocity components.

Velocity dispersion. The formal definition of the components of the velocity dispersion is as follows: let $f(\boldsymbol{v})\,\mathrm{d}^3v$ be the number density of stars (of a given population) at a fixed location, with velocities in a volume element d^3v around $\boldsymbol{v}$ in the vector space of velocities. If we use Cartesian coordinates, for example $\boldsymbol{v} = (v_1, v_2, v_3)$, then $f(\boldsymbol{v})\,\mathrm{d}^3v$ is the number of stars with the i-th velocity component in the interval $[v_i, v_i + \mathrm{d}v_i]$, and $\mathrm{d}^3v = \mathrm{d}v_1\,\mathrm{d}v_2\,\mathrm{d}v_3$. The mean velocity $\langle\boldsymbol{v}\rangle$ of the population then follows from this distribution via

$$\langle\boldsymbol{v}\rangle = n^{-1}\int_{\mathbf{R}^3}\mathrm{d}^3v\; f(\boldsymbol{v})\,\boldsymbol{v}\;, \quad \text{where} \quad n = \int_{\mathbf{R}^3}\mathrm{d}^3v\; f(\boldsymbol{v}) \tag{2.33}$$

denotes the total number density of stars in the population. The velocity dispersion σ then describes the root mean square deviations of the velocities from $\langle\boldsymbol{v}\rangle$. For a component i of the velocity vector, the dispersion σ_i is defined as

$$\sigma_i^2 = \left\langle (v_i - \langle v_i\rangle)^2\right\rangle = \left\langle v_i^2 - \langle v_i\rangle^2\right\rangle = n^{-1}\int_{\mathbf{R}^3}\mathrm{d}^3v\; f(\boldsymbol{v})\left(v_i^2 - \langle v_i\rangle^2\right) . \tag{2.34}$$

The larger σ_i is, the broader the distribution of the stochastic motions. We note that the same concept applies to the velocity distribution of molecules in a gas. The mean velocity $\langle\boldsymbol{v}\rangle$ at each point defines the bulk velocity of the gas, e.g., the wind speed in the atmosphere, whereas the velocity dispersion is caused by thermal motion of the molecules and is determined by the temperature of the gas.

The random motion of the stars in the direction perpendicular to the disk is the reason for the finite thickness of the population; it is similar to a thermal distribution. Accordingly, it has the effect of a pressure, the so-called *dynamical pressure* of the distribution. This pressure determines the scale-height of the distribution, which corresponds to the law of atmospheres. The larger the dynamical pressure, i.e., the larger the velocity dispersion σ_z perpendicular to the disk, the larger the scale-height h will be. The analysis of stars in the Solar neighborhood yields $\sigma_z \sim 16\,\mathrm{km/s}$ for stars younger than $\sim 3\,\mathrm{Gyr}$, corresponding to a scale-height of $h \sim 250\,\mathrm{pc}$, whereas stars older than $\sim 6\,\mathrm{Gyr}$ have a scale-height of $\sim 350\,\mathrm{pc}$ and a velocity dispersion of $\sigma_z \sim 25\,\mathrm{km/s}$.

The density distribution of the total star population, obtained from counts and distance determinations of stars, is to a good approximation described by

$$\boxed{n(R,z) = n_0\left(\mathrm{e}^{-|z|/h_{\rm thin}} + 0.02\mathrm{e}^{-|z|/h_{\rm thick}}\right)\mathrm{e}^{-R/h_R}} ; \tag{2.35}$$

here, R and z are the cylinder coordinates introduced above (see Sect. 2.1), with the origin at the Galactic center, and $h_{\rm thin} \approx h_{\rm otd} \approx 325\,\mathrm{pc}$ is the scale-height of the thin disk. The distribution in the radial direction can also be well described by an exponential law, where $h_R \approx 3.5\,\mathrm{kpc}$ denotes the *scale-length of the Galactic disk*. The normalization of the distribution is determined by the density $n \approx 0.2\ \mathrm{stars/pc^3}$ in the Solar neighborhood, for stars in the range of absolute magnitudes of $4.5 \le M_V \le 9.5$. The distribution described by (2.35) is not smooth at $z = 0$; it has a kink at this point and it is therefore unphysical. To get a smooth distribution which follows the exponential law for large z and is smooth in the plane of the disk, the distribution is slightly modified. As an example, for the luminosity density of the old thin disk (that is proportional to the number density of the stars), we can write:

$$\boxed{L(R,z) = \frac{L_0\,\mathrm{e}^{-R/h_R}}{\cosh^2(z/h_z)}} , \tag{2.36}$$

with $h_z = 2h_{\rm thin}$ and $L_0 \approx 0.05 L_\odot/{\rm pc}^3$. The Sun is a member of the young thin disk and is located above the plane of the disk, at $z \approx 30\,{\rm pc}$.

2.3.2 The Galactic disk: chemical composition and age; supernovae

Stellar populations. The chemical composition of stars in the thin and the thick disks differs: we observe the clear tendency that stars in the thin disk have a higher metallicity than those in the thick disk. In contrast, the metallicity of stars in the Galactic halo and in the bulge is smaller. To paraphrase these trends, one distinguishes between stars of population I (pop I) which have a Solar-like metallicity ($Z \sim 0.02$) and are mainly located in the thin disk, and stars of population II (pop II) that are metal-poor ($Z \sim 0.001$) and predominantly found in the thick disk, in the halo, and in the bulge. In reality, stars cover a wide range in Z, and the figures above are only characteristic values. For stellar populations a somewhat finer separation was also introduced, such as 'extreme population I', 'intermediate population II', and so on. The populations also differ in age (stars of pop I are younger than those of pop II), in scale height (as mentioned above), and in the velocity dispersion perpendicular to the disk (σ_z is larger for pop II stars than for pop I stars).

We shall now attempt to understand the origin of these different metallicities and their relation to the scale height and to age, starting with a brief discussion of the phenomenon that is the main reason for the metal enrichment of the interstellar medium.

Metallicity and supernovae. Supernovae (SNe) are explosive events. Within a few days, a SN can reach a luminosity of $10^9 L_\odot$, which is a considerable fraction of the total luminosity of a galaxy; after that the luminosity decreases again with a time-scale of weeks. In the explosion, a star is disrupted and (most of) the matter of the star is driven into the interstellar medium, enriching it with metals that were produced in the course of stellar evolution or in the process of the supernova explosion.

Classification of supernovae. Based on their spectral properties, SNe are divided into several classes. SNe of Type I do not show any Balmer lines of hydrogen in their spectrum, in contrast to those of Type II. The Type I SNe are further subdivided: SNe Ia show strong emission of SiII $\lambda\,6150\,$Å whereas no SiII at all is visible in spectra of Type Ib,c. Our current understanding of the supernova phenomenon differs from this spectral classification.[6] Following various observational results and also theoretical analyses, we are confident today that SNe Ia are a phenomenon which is intrinsically different from the other supernova types. For this interpretation, it is of particular importance that SNe Ia are found in all types of galaxies, whereas we observe SNe II and SNe Ib,c only in spiral and irregular galaxies, and here only in those regions in which blue stars predominate. As we will see in Chap. 3, the stellar population in elliptical galaxies consists almost exclusively of old stars, while spirals also contain young stars. From this observational fact it is concluded that the phenomenon of SNe II and SNe Ib,c is linked to a young stellar population, whereas SNe Ia occur also in older stellar populations. We shall discuss the two classes of supernovae next.

Core-collapse supernovae. SNe II and SNe Ib,c are the final stages in the evolution of massive ($\gtrsim 8 M_\odot$) stars. Inside these stars, ever heavier elements are generated by nuclear fusion: once all the hydrogen in the inner region is used up, helium will be burned, then carbon, oxygen, etc. This chain comes to an end when the iron nucleus is reached, the atomic nucleus with the highest binding energy per nucleon. After this no more energy can be gained from fusion to heavier elements so that the pressure, which is normally balancing the gravitational force in the star, can no longer be maintained. The star then collapse under its own gravity. This gravitational collapse proceeds until the innermost region reaches a density about three times the density of an atomic nucleus. At this point the so-called rebounce occurs: a shock wave runs towards the surface, thereby heating the infalling material, and the star explodes. In the center, a compact object probably remains—a neutron star or, possibly, depending on the mass of the iron core, a black hole. Such neutron stars are visible as pulsars[7] at the location of some historically observed SNe, the most famous of which is the Crab pulsar which has been identified with a supernovae explosion seen by Chinese astronomers in 1054. Presumably all neutron stars have been formed in such core-collapse supernovae.

The major fraction of the binding energy released in the formation of the compact object is emitted in the form of neutrinos: about 3×10^{53} erg. Underground neutrino detectors

[6] This notation scheme (Type Ia, Type II, and so on) is characteristic for phenomena that one wishes to classify upon discovery, but for which no physical interpretation is available at that time. Other examples are the spectral classes of stars which are not named in alphabetical order nor according to their mass on the main sequence; or the division of Seyfert galaxies into Type 1 and Type 2. Once such a notation is established, it often becomes permanent even if a later physical understanding of the phenomenon suggests a more meaningful classification.

[7] Pulsars are sources which show a *very* regular periodic radiation, most often seen at radio frequencies. Their periods lie in the range from $\sim 10^{-3}\,$s (milli-second pulsars) to $\sim 5\,$s. Their pulse period is identified as the rotational period of the neutron star—an object with about one Solar mass and a radius of $\sim 10\,$km. The matter density in neutron stars is about the same as that in atomic nuclei.

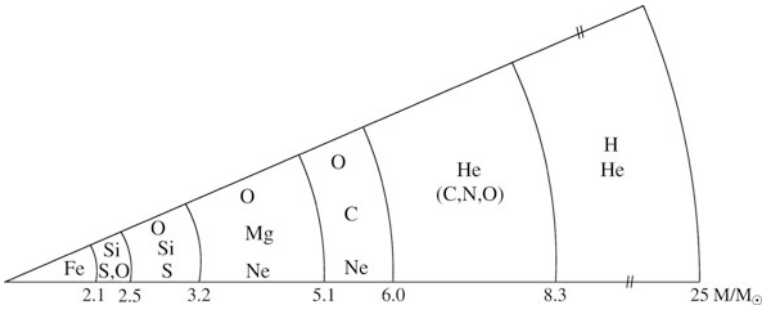


Fig. 2.11 Chemical shell structure of a massive star at the end of its life with the axis labeled by the mass within a given radius. The elements that have been formed in the various stages of the nuclear burning are ordered in a structure resembling that of an onion, with heavier elements being located closer to the center. This is the initial condition for a supernova explosion. Adapted from A. Unsöld & B. Baschek, *The New Cosmos*, Springer-Verlag

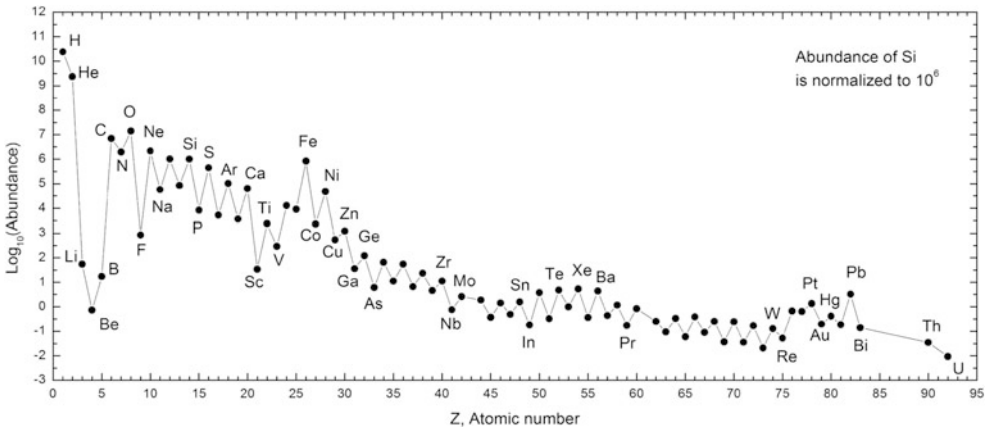


Fig. 2.12 The relative abundance of chemical elements in the Solar System, normalized such that silicon attains the value 10^6. By far the most abundant elements are hydrogen and helium; as we will see later, these elements were produced in the first 3 min of the cosmic evolution. Essentially all the other elements were produced later in stellar interiors. As a general trend, the abundances decrease with increasing atomic number, except for the light elements lithium (Li), beryllium (Be), and boron (B), which are generated in stars, but also easily destroyed due to their low binding energy. Superposed on this decrease, the abundances show an oscillating behavior: nuclei with an even number of protons are more abundant than those with an odd atomic number—this phenomenon is due to the production of alpha elements in core-collapse supernovae. Furthermore, iron (Fe), cobalt (Co) and nickel (Ni) stick out in their relatively high abundance, given their atomic number, which is due to their abundant production mainly in Type Ia SNe. Source: Wikipedia, numerical data from: Katharina Lodders

were able to trace about 10 neutrinos originating from SN 1987A in the Large Magellanic Cloud.[8] Due to the high density inside the star after the collapse, even neutrinos, despite their very small cross section, are absorbed and scattered, so that part of their outward-directed momentum contributes to the explosion of the stellar envelope. This shell expands at $v \sim 10\,000\,\mathrm{km/s}$, corresponding to a kinetic energy of $E_{\mathrm{kin}} \sim 10^{51}$ erg. Of this, only about 10^{49} erg is converted into photons in the hot envelope and then emitted—the energy of a SN that is visible in photons is thus only a small fraction of the total energy produced.

Owing to the various stages of nuclear fusion in the progenitor star, the chemical elements are arranged in shells: the light elements (H, He) in the outer shells, and the heavier elements (C, O, Ne, Mg, Si, Ar, Ca, Fe, Ni) in the inner ones—see Fig. 2.11. The explosion ejects them into the interstellar medium which is thus chemically enriched. It is important to note that mainly nuclei with an even number of protons and neutrons are formed. This is a consequence of the nuclear reaction chains involved, where successive nuclei in this chain are obtained by adding an α-particle (or ^{4}He-nucleus), i.e., two protons and two neutrons. Such elements are therefore called α-elements. The dominance of α-elements in the chemical abundance of the interstellar medium, as well as in the Solar System (see Fig. 2.12), is

[8]The name of a supernova is composed of the year of explosion, and a single capital letter or two lower case letters. The first detected supernova in a year gets the letter 'A', the second 'B' and so on; the 27th then obtains an 'aa', the 28th an 'ab' etc. Hence, SN 1987A was the first one discovered in 1987.

thus a clear indication of nuclear fusion occurring in the He-rich zones of stars where the hydrogen has been burnt.

Supernovae Type Ia. SNe Ia are most likely the explosions of white dwarfs (WDs). These compact stars which form the final evolutionary stages of less massive stars no longer maintain their internal pressure by nuclear fusion. Rather, they are stabilized by the degeneracy pressure of the electrons—a quantum mechanical phenomenon related to the Fermi exclusion principle. Such a white dwarf can be stable only if its mass does not exceed a limiting mass, the *Chandrasekhar mass*; it has a value of $M_{\rm Ch} \approx 1.44 M_\odot$. For $M > M_{\rm Ch}$, the degeneracy pressure can no longer balance the gravitational force.

A white dwarf can become unstable if its mass approaches the Chandrasekhar mass limit. There are two different scenarios with which this is possible: If the white dwarf is part of a close binary system, matter from the companion star may flow onto the white dwarf; this is called the 'single-degenerate' model. In this process, its mass will slowly increase and approach the limiting mass. At about $M \approx 1.3 M_\odot$, carbon burning will ignite in its interior, transforming about half of the star into iron-group elements, i.e., iron, cobalt, and nickel. The resulting explosion of the star will enrich the ISM with $\sim 0.6\ M_\odot$ of Fe, while the WD itself will be torn apart completely, leaving no remnant star. A second (so-called 'double-degenerate') scenario for the origin of SNe Ia is that of the merger of two white dwarfs for which the sum of their masses exceeds the Chandrasekhar mass. Of course, these two scenarios are not mutually exclusive, and both routes may be realized in nature.

Since the initial conditions are probably very homogeneous for the class of SNe Ia in the single-degenerate scenario (defined by the limiting mass prior to the trigger of the explosion), they are good candidates for *standard candles*: all SNe Ia have approximately the same luminosity. As we will discuss later (see Sect. 3.9.4), this is not really the case, but nevertheless SNe Ia play a very important role in the cosmological distance determination, and thus in the determination of cosmological parameters. On the other hand, in the double-degenerate scenario, the class of SNe Ia is not expected to be very homogeneous, as the mass prior to the explosion no longer attains a universal value. In fact, there are some SNe Ia which are clearly different from the majority of this class, by being far more luminous. It may be that such events are triggered by the merging of two white dwarfs, whereas the majority of the explosions is caused by the single-degenerate formation process.

This interpretation of the different types of SNe explains why one finds core-collapse SNe only in galaxies in which star formation occurs. They are the final stages of massive, i.e., young, stars which have a lifetime of not more than 2×10^7 yr. By contrast, SNe Ia can occur in all types of galaxies, since their progenitors are members of an old stellar population.

In addition to SNe, metal enrichment of the interstellar medium (ISM) also takes place in other stages of stellar evolution, by stellar winds or during phases in which stars eject part of their envelope which is then visible, e.g., as a planetary nebula. If the matter in the star has been mixed by convection prior to such a phase, so that the metals newly formed by nuclear fusion in the interior have been transported towards the surface of the star, these metals will then be released into the ISM.

Age-metallicity relation. Assuming that at the beginning of its evolution the Milky Way had a chemical composition with only low metal content, the metallicity should be strongly related to the age of a stellar population. With each new generation of stars, more metals are produced and ejected into the ISM, partially by stellar winds, but mainly by SN explosions. Stars that are formed later should therefore have a higher metal content than those that were formed in the early phase of the Galaxy. One would thus expect that a relation exists between the age of a star and its metallicity.

For instance, under this assumption the iron abundance [Fe/H] can be used as an age indicator for a stellar population, with the iron predominantly being produced and ejected in SNe of Type Ia. Therefore, a newly formed generation of stars has a higher fraction of iron than their predecessors, and the youngest stars should have the highest iron abundance. Indeed one finds $[{\rm Fe/H}] = -4.5$ (i.e., 3×10^{-5} of the Solar iron abundance) for extremely old stars, whereas very young stars have $[{\rm Fe/H}] = 1$, so their metallicity can significantly exceed that of the Sun.

However, this age-metallicity relation is not very tight. On the one hand, SNe Ia occur only $\gtrsim 10^9$ yr after the formation of a stellar population. The exact time-span is not known because even if one accepts the accretion scenario for SN Ia described above, it is unclear in what form and in what systems the accretion of material onto the white dwarf takes place and how long it typically takes until the limiting mass is reached. On the other hand, the mixing of the SN ejecta in the ISM occurs only locally, so that large inhomogeneities of the [Fe/H] ratio may be present in the ISM, and thus even for stars of the same age. An alternative measure for metallicity is [O/H], because oxygen, which is an α-element, is produced and ejected mainly in supernova explosions of massive stars. These happen just $\sim 10^7$yr after the formation of a stellar population, which is virtually instantaneous.

Origin of the thick disk. Characteristic values for the metallicity are $-0.5 \lesssim [{\rm Fe/H}] \lesssim 0.3$ in the thin disk, while for the thick disk $-1.0 \lesssim [{\rm Fe/H}] \lesssim -0.4$ is typical. From this, one can deduce that stars in the thin disk must be

significantly younger on average than those in the thick disk. This result can now be interpreted using the age-metallicity relation. Either star formation has started earlier, or ceased earlier, in the thick disk than in the thin disk, or stars that originally belonged to the thin disk have migrated into the thick disk. The second alternative is favored for various reasons. It would be hard to understand why molecular gas, out of which stars are formed, was much more broadly distributed in earlier times than it is today, where we find it well concentrated near the Galactic plane. In addition, the widening of an initially narrow stellar distribution in time is also expected. The matter distribution in the disk is not homogeneous and, along their orbits around the Galactic center, stars experience this inhomogeneous gravitational field caused by other stars, spiral arms, and massive molecular clouds. Stellar orbits are perturbed by such fluctuations, i.e., they gain a random velocity component perpendicular to the disk from local inhomogeneities of the gravitational field. In other words, the velocity dispersion σ_z of a stellar population grows in time, and the scale height of a population increases. In contrast to stars, the gas keeps its narrow distribution around the Galactic plane due to internal friction.

This interpretation is, however, not unambiguous. Another scenario for the formation of the thick disk is also possible, where the stars of the thick disk were formed outside the Milky Way and only became constituents of the disk later, through accretion of satellite galaxies. This model is supported, among other reasons, by the fact that the rotational velocity of the thick disk around the Galactic center is smaller by $\sim 50\,\mathrm{km/s}$ than that of the thin disk. In other spirals, in which a thick disk component was found and kinematically analyzed, the discrepancy between the rotation curves of the thick and thin disks is sometimes even stronger. In one case, the thick disk was observed to rotate around the center of the galaxy in the opposite direction to the gas disk. In such a case, the aforementioned model of the evolution of the thick disk by kinematic heating of stars would definitely not apply.

Mass-to-light ratio. The total stellar mass of the thin disk is $\sim 6 \times 10^{10} M_\odot$, to which $\sim 0.5 \times 10^{10} M_\odot$ in the form of dust and gas has to be added. The luminosity of the stars in the thin disk is $L_B \approx 1.8 \times 10^{10} L_\odot$. Together, this yields a mass-to-light ratio of

$$\boxed{\frac{M}{L_B} \approx 3 \frac{M_\odot}{L_\odot} \quad \text{in thin disk}}\,. \tag{2.37}$$

The M/L ratio in the thick disk is higher, as expected from an older stellar population. The relative contribution of the thick disk to the stellar budget of the Milky Way is quite uncertain; estimates range from $\sim$ 5 to $\sim$ 30 %, which reflects the difficulty to attribute individual stars to the thin vs. thick disk; also the criteria for this classification vary substantially. In any case, due to the larger mass-to-light ratio of the thick disk, its contribution to the luminosity of the Milky Way is small. Nevertheless, the thick disk is invaluable for the diagnosis of the dynamical evolution of the disk. If the Milky Way were to be observed from the outside, one would find a M/L value for the disk of about four in Solar units; this is a characteristic value for spiral galaxies.

2.3.3 The Galactic disk: dust and gas

Spatial distribution. The spiral structure of the Milky Way and other spiral galaxies is delineated by very young objects like O- and B-stars and HII-regions.[9] This is the reason why spiral arms appear blue. Obviously, star formation in our Milky Way takes place mainly in the spiral arms. Here, the *molecular clouds*—gas clouds which are sufficiently dense and cool for molecules to form in large abundance—contract under their own gravity and form new stars. The spiral arms are much less prominent in red light (see also Fig. 3.24 below). Emission in the red is dominated by an older stellar population, and these old stars have had time to move away from the spiral arms. The Sun is located close to, but not in, a spiral arm—the so-called Orion arm (see Fig. 2.13).

Open clusters. Star formation in molecular clouds leads to the formation of open star clusters, since stars are not born individually; instead, the contraction of a molecular cloud gives rise to many stars at the same time, which form an (open) star cluster. Its mass depends of course on the mass of the parent molecular cloud, ranging from $\sim 100\,M_\odot$ to $\sim 10^4\,M_\odot$. The stars in these clusters all have the same velocity—indeed, the velocity dispersion in open clusters is small, below $\sim 1\,\mathrm{km/s}$.

Since molecular gas is concentrated close to the Galactic plane, such star clusters in the Milky Way are born there. Most of the open clusters known have ages below 300 Myr, and those are found within $\sim$ 50 pc of the Galactic plane. Older clusters can have larger $|z|$, as they can move from their place of birth, similar to what we said about the stars in the thick disk. The reason why we see only a few open clusters with ages above 1 Gyr is that these are not strongly gravitationally bound, if at all. Hence, in the course of time, tidal gravitational forces dissolve such clusters, and this effect is more important at small galactocentric radii R.

[9] H II-regions are nearly spherical regions of fully ionized hydrogen (thus the name HII region) surrounding a young hot star which photoionizes the gas. They emit strong emission lines of which the Balmer lines of hydrogen are strongest.

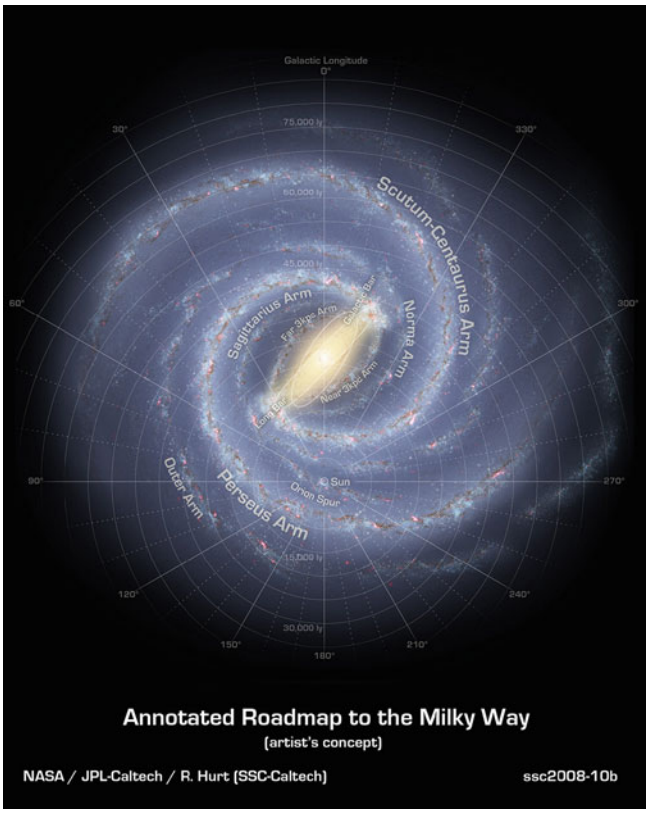


Fig. 2.13 A sketch of the plane of the Milky Way, based to a large degree on observations from the Spitzer Space Telescope. It shows the two major spiral arms which originate at the ends of the central bar, as well as two minor spiral arms. The Sun is located near the Orion arm, a partial spiral arm. Credit: NASA/JPL-Caltech/R. Hurt (SSC/Caltech)

Observing the gas in the Galaxy is made possible mainly by the 21 cm line emission of HI (neutral atomic hydrogen) and by the emission of CO, the second-most abundant molecule after H_2 (molecular hydrogen). H_2 is a symmetric molecule and thus has no electric dipole moment, which is the main reason why it does not radiate strongly. In most cases it is assumed that the ratio of CO to H_2 is a universal constant (called the 'X-factor'). Under this assumption, the distribution of CO can be converted into that of the total molecular gas. The Milky Way is optically thin at 21 cm, i.e., 21 cm radiation is not absorbed along its path from the source to the observer. With radio-astronomical methods it is thus possible to observe atomic gas throughout the entire Galaxy.

Distribution of dust. To examine the distribution of dust, two options are available. First, dust is detected by the extinction it causes. This effect can be analyzed quantitatively, for instance by star counts or by investigating the reddening of stars (an example of this can be seen in Fig. 2.8). Second, dust emits thermal radiation, observable in the FIR part of the spectrum, which was mapped by several satellites such as IRAS and COBE. By combining the sky maps of these two satellites at different frequencies, the Galactic distribution of dust was determined. The dust temperature varies in a relatively narrow range between $\sim$ 17 and $\sim$ 21 K, but even across this small range, the dust emission varies, for fixed column density, by a factor $\sim$ 5 at a wavelength of 100 μm. Therefore, one needs to combine maps at different frequencies in order to determine column densities and temperatures. In addition, the zodiacal light caused by the reflection of Solar radiation by dust inside our Solar system has to be subtracted before the Galactic FIR emission can be analyzed. This is possible with multi-frequency data because of the different spectral shapes. The resulting distribution of dust is displayed in Fig. 2.14. It shows the concentration of dust around the Galactic plane, as well as large-scale anisotropies at high Galactic latitudes. The dust map shown here is routinely used for extinction correction when observing extragalactic sources.

Besides a strong concentration towards the Galactic plane, gas and dust are preferentially found in spiral arms where they serve as raw material for star formation. Molecular hydrogen (H_2) and dust are generally found at $3\,\mathrm{kpc} \lesssim R \lesssim 8\,\mathrm{kpc}$, within $|z| \lesssim 90\,\mathrm{pc}$ of both sides of the Galactic plane. In contrast, the distribution of atomic hydrogen (HI) is observed out to much larger distances from the Galactic center ($R \lesssim 25\mathrm{kpc}$), with a scale height of $\sim 160\,\mathrm{pc}$ inside the Solar orbit, $R \lesssim R_0$. At larger distances from the Galactic center, $R \gtrsim 12\,\mathrm{kpc}$, the scale height increases substantially to $\sim 1\,\mathrm{kpc}$. The gaseous disk is warped at these large radii though the origin of this warp is unclear. For example, it may be caused by the gravitational field of the Magellanic Clouds. The total mass in the two components of hydrogen is about $M(\mathrm{HI}) \approx 4 \times 10^9 M_\odot$ and $M(\mathrm{H_2}) \approx 10^9 M_\odot$, respectively, i.e., the gas mass in our Galaxy is less than $\sim 10\,\%$ of the stellar mass. The density of the gas in the Solar neighborhood is about $\rho(\mathrm{gas}) \sim 0.04 M_\odot/\mathrm{pc}^3$.

Phases of the interstellar medium. Gas in the Milky Way exists at a range of different temperatures and densities. The coolest phase of the interstellar medium is that represented by molecular gas. Since molecules are easily destroyed by photons from hot stars, they need to be shielded from the interstellar radiation field, which is provided by the dust embedded in the gas. The molecules can cool the gas efficiently even at low temperatures: through collisions between particles, part of the kinetic energy can be used to put one of the particles into an excited state, and thus to remove kinetic energy from the particle distribution, thereby lowering their mean velocity and, thus, their temperature. This is possible only if the kinetic energy is high enough for this internal excitation. Molecules have excited levels at low energies—the rotational and vibrational excitations—so they are able to cool cold gas; in fact, this is the necessary condition for

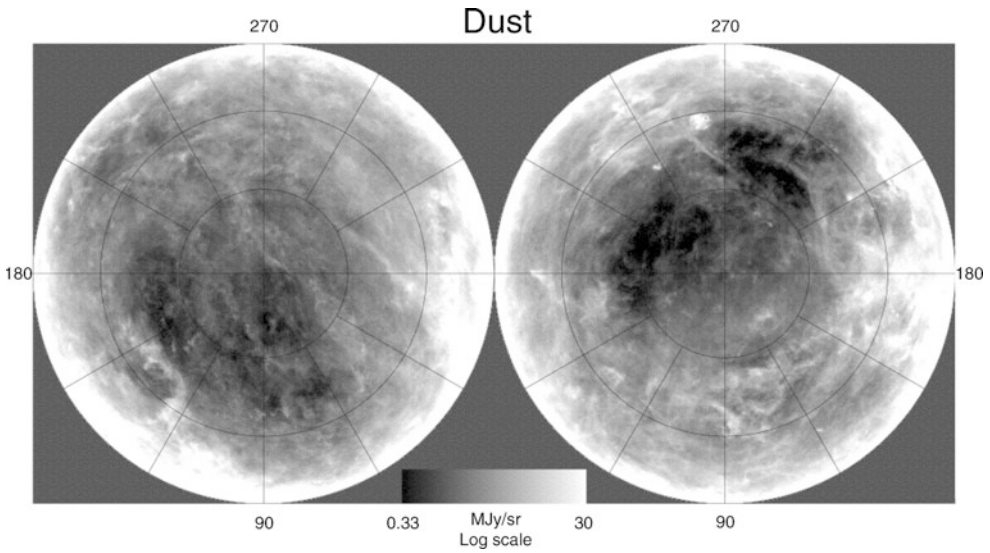


Fig. 2.14 Distribution of dust in the Galaxy, derived from a combination of IRAS and COBE sky maps. The northern Galactic sky in Galactic coordinates is displayed *on the left*, the southern *on the right*. We can clearly see the concentration of dust towards the Galactic plane, as well as regions with a very low column density of dust; these regions in the sky are particularly well suited for very deep extragalactic observations. Source: D.J. Schlegel, D.P. Finkbeiner & M. Davis 1998, *Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds*, ApJ 500, 525, p. 542, Fig. 8. ©AAS. Reproduced with permission

the formation of stars. The energy in the excited level is then released by the emission of a photon which can escape. The range of temperatures in the molecular gas phase extends from $\sim$ 10 K to about 70 K, with characteristic densities of 100 particles per cm^3.

A second prominent phase is the warm interstellar gas, with temperatures of a few thousand degrees. Depending on T, the fraction of atoms which are ionized, i.e., the ionization fraction, can range from 0.01 to 1. This gas can be heated by hydrodynamical processes or by photoionization. For example, gas near to a hot star will be ionized by the energetic photons. The kinetic energy of the electron released in this photoionization process is the difference between the energy of the ionizing photon and the binding energy of the electron. The energy of the electron is then transferred to the gas through collisions, thus providing an effective heating source. Cooling is provided by atomic transitions excited by collisions between atoms, or recombination of atoms with electrons, and the subsequent emission of photons from the excited states. Since hydrogen is by far the most abundant species, its atomic transitions dominate the cooling for $T \gtrsim$ 5000 K, and is then a very efficient coolant. Because of that, the temperature of this warm gas tends towards $T \sim$ 8000 K, almost independent of the intensity and spectrum of the ionizing radiation, at least over a wide range of these parameters. Perhaps the best known examples for this gas are the aforementioned HII regions around hot stars, and planetary nebulae.

2.3.4 Cosmic rays

The magnetic field of the Galaxy. Like many other cosmic objects, the Milky Way contains a magnetic field. The properties of this field can be analyzed using a variety of methods, and we list some of them in the following.

- *Polarization of stellar light.* The light of distant stars is partially polarized, with the degree of polarization being strongly related to the extinction, or reddening, of the star. This hints at the polarization being linked to the dust causing the extinction. The light scattered by dust particles is partially linearly polarized, with the direction of polarization depending on the alignment of the dust grains. If their orientation was random, the superposition of the scattered radiation from different dust particles would add up to a vanishing net polarization. However, a net polarization is measured, so the orientation of dust particles cannot be random, rather it must be coherent on large scales. Such a coherent alignment is provided by a large-scale magnetic field, whereby the orientation of dust particles, measurable from the polarization direction, indicates the (projected) direction of the magnetic field.
- *The Zeeman effect.* The energy levels in an atom change if the atom is placed in a magnetic field. Of particular importance in the present context is the fact that the 21 cm transition line of neutral hydrogen is split in a magnetic field. Because the amplitude of the line split is proportional to the strength of the magnetic field, the

field strength can be determined from observations of this Zeeman effect.

- *Synchrotron radiation.* When relativistic electrons move in a magnetic field they are subject to the Lorentz force. The corresponding acceleration is perpendicular both to the velocity vector of the particles and to the magnetic field vector. As a result, the electrons follow a helical (i.e., corkscrew) track, which is a superposition of circular orbits perpendicular to the field lines and a linear motion along the field. Since accelerated charges emit electromagnetic radiation, this helical movement is the source of the so-called synchrotron radiation (which will be discussed in more detail in Sect. 5.1.2). This radiation, which is observable at radio frequencies, is linearly polarized, with the direction of the polarization depending on the direction of the magnetic field.
- *Faraday rotation.* If polarized radiation passes through a magnetized plasma, the direction of the polarization rotates. The rotation angle depends quadratically on the wavelength of the radiation,

$$\Delta\theta = \mathrm{RM}\,\lambda^2 \,. \tag{2.38}$$

The *rotation measure* RM is the integral along the line-of-sight towards the source over the electron density and the component $B_\parallel$ of the magnetic field in direction of the line-of-sight,

$$\mathrm{RM} = 81\,\frac{\mathrm{rad}}{\mathrm{cm}^2}\int_0^D \frac{\mathrm{d}\ell}{\mathrm{pc}}\,\frac{n_\mathrm{e}}{\mathrm{cm}^{-3}}\,\frac{B_\parallel}{\mathrm{G}}\,. \tag{2.39}$$

The dependence of the rotation angle (2.38) on λ allows us to determine the rotation measure RM, and thus to estimate the product of electron density and magnetic field. If the former is known, one immediately gets information about B. By measuring the RM for sources in different directions and at different distances the magnetic field of the Galaxy can be mapped.

From applying the methods discussed above, we know that a magnetic field exists in the disk of our Milky Way. This field has a strength of about 4×10^{-6} G and mainly follows the spiral arms.

Cosmic rays. We obtain most of the information about our Universe from the electromagnetic radiation that we observe. However, we receive an additional radiation component, the energetic cosmic rays, which were discovered by Victor Hess in 1912 who carried out balloon flights and found that the degree of ionizing radiation increases with increasing height. Cosmic rays consist primarily of electrically charged particles, mainly electrons and nuclei. In addition to the particle radiation that is produced in energetic processes at the Solar surface, a much more energetic cosmic ray component exists that can only originate in sources outside the Solar system.

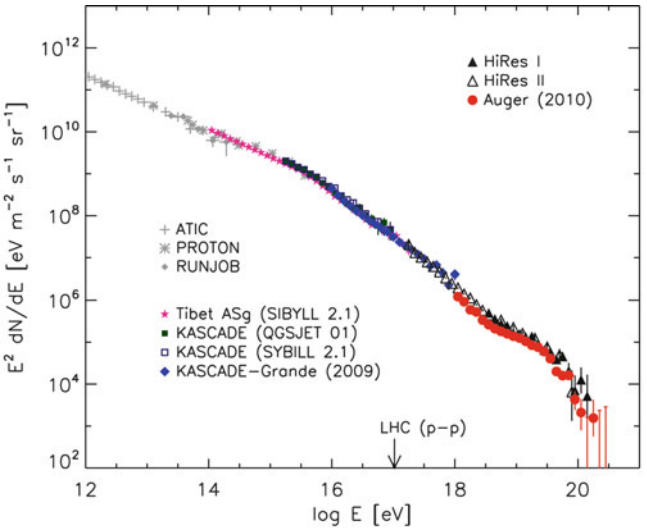


Fig. 2.15 The energy spectrum $\mathrm{d}N/\mathrm{d}E$ of cosmic rays, for better visibility multiplied by E^2. Data from different experiments are shown by *different symbols*. At energies below 10^{10} eV (not shown), the flux of cosmic rays is dominated by those from the Sun, whereas for higher energies, they are due to sources in our Galaxy or beyond. The energy spectrum is well described by piecewise power-law spectra, with a steepening at $E \sim 10^{15}$ eV (called the knee), and a flattening at $E \sim 3 \times 10^{18}$ eV. Beyond $E \sim 3 \times 10^{19}$ eV, the spectrum shows a cut-off. Also indicated is the energy of a cosmic ray proton whose collision with a proton in the Earth' atmosphere has the same center-of-mass energy as the highest energy collisions at the Large Hadron Collider at CERN. The cosmic ray fluxes are very small: cosmic rays with energies larger than $\sim 10^{15}$ eV arrive at the Earth at a rate of about 1 per m^2 per year, those with energies above 10^{18} eV come at a rate of approximately 1 km^{-2}yr^{-1}; this implies that one needs huge detectors to study these particles. Source: K. Kotera & A.V. Olinto 2011, *The Astrophysics of Ultrahigh-Energy Cosmic Rays*, ARA&A 49, 119, p. 120, Fig. 1. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 49 ©2011 by Annual Reviews `www.annualreviews.org`

The energy spectrum of the cosmic rays is, to a good approximation, a power law: the flux of particles with energy between E and $E + \mathrm{d}E$ can be written as $(\mathrm{d}N/\mathrm{d}E)\,\mathrm{d}E \propto E^{-q}\,\mathrm{d}E$, with $q \approx 2.7$. However, as can be seen in Fig. 2.15, the slope of the spectrum changes slightly, but significantly, at some energy scales: at $E \sim 10^{15}$ eV the spectrum becomes steeper, and at $E \gtrsim 10^{18}$ eV it flattens again[10]; these two energy scales in the cosmic ray spectrum have been given the suggestive names of 'knee' and 'ankle', respectively. Measurements of the spectrum at these high energies are rather uncertain, however, because of the strongly decreasing

[10] These energies should be compared with those reached in particle accelerators: the LHC at CERN reaches ~ 10 TeV $= 10^{13}$ eV. Hence, cosmic accelerators are much more efficient than man-made machines.

flux with increasing energy. This implies that only very few particles are detected.

Cosmic ray acceleration and confinement. To accelerate particles to such high energies, very energetic processes are necessary. For energies below 10^{15} eV, very convincing arguments suggest supernova remnants as the sites of the acceleration. The SN explosion drives a shock front[11] into the ISM with an initial velocity of $\sim$ 10 000 km/s. Plasma processes in a shock front can accelerate some particles to very high energies. The theory of this diffuse shock acceleration predicts that the resulting energy spectrum of the particles follows a power law, the slope of which depends only on the strength of the shock (i.e., the ratio of the densities on both sides of the shock front). This power law agrees very well with the slope of the observed cosmic ray spectrum below the knee, if additional effects caused by the propagation of particles in the Milky Way (e.g., energy losses, and the possibility for escaping the Galaxy) are taken into account. The presence of very energetic electrons in SN remnants is observed directly by their synchrotron emission, so that the slope of the produced spectrum can be inferred by observations.

Accelerated particles then propagate through the Galaxy where, due to the magnetic field, they move along complicated helical tracks. Therefore, the direction from which a particle arrives at Earth cannot be identified with the direction to its source of origin. The magnetic field is also the reason why particles do not leave the Milky Way along a straight path, but instead are stored for a long time ($\sim 10^7$ yr) before they eventually diffuse out, an effect called confinement.

The sources of the particles with energy between $\sim 10^{15}$ eV and $\sim 10^{18}$ eV are likewise presumed to be located inside our Milky Way, because the magnetic field is sufficiently strong to confine them in the Galaxy. It is not known, however, whether these particles are also accelerated in supernova remnants; if they are, the steepening of the spectrum may be related to the fact that particles with $E \gtrsim 10^{15}$ eV have a Larmor radius which no longer is small compared to the size of the remnant itself, and so they find it easier to escape from the accelerating region. Particles with energies larger than $\sim 10^{18}$ eV are probably of extragalactic origin. The radius of their helical tracks in the magnetic field of the Galaxy, i.e., their Larmor radius, is larger than the radius of the Milky Way itself, so they cannot be confined. Their origin is also unknown, but AGNs are the most probable source of these particles.

Ultra-high energy cosmic rays. Finally, one of the largest puzzles of high-energy astrophysics is the origin of cosmic rays with $E \gtrsim 10^{19}$ eV. The energy of these so-called ultra-high energy cosmic rays (UHECRs) is so large that they are able to interact with the cosmic microwave background to produce pions and other particles, losing much of their energy in this process. These particles cannot propagate much further than $\sim$ 100 Mpc through the Universe before they have lost most of their energy. This implies that their acceleration sites should be located in the close vicinity of the Milky Way. Since the curvature of the orbits of such highly energetic particles is very small, it should, in principle, be possible to identify their origin: there are not many AGNs within 100 Mpc that are promising candidates for the origin of these ultra-high energy cosmic rays. Furthermore, the maximal possible distance a cosmic ray particle can propagate through the Universe decreases strongly with increasing energy, so that the number of potential sources must decrease accordingly. Once this minimal distance is below the nearest AGN, there should be essentially no particle that can reach us. In other words, one expects to see a cut-off (called the Greisen–Zatsepin–Kuzmin, or GZK cut-off) in the energy spectrum at $E \sim 2 \times 10^{20}$ eV, but beginning already at $E \gtrsim 5 \times 10^{19}$ eV. Before 2007, this cut-off was not observed, and different cosmic ray experiments reported a different energy spectrum for these UHECRs—based, literally, on a handful of events.

The breakthrough came with the first results from the Auger experiment, the by far most sensitive experiment owing to its large effective area.[12] When the first results were published in 2007, the expected high-energy cut-off in the UHECR spectrum was detected—thereby erasing the necessity for many very exotic processes that had been proposed earlier to account for the apparent lack of this cut-off. With this detection the idea about the origin of the UHECRs from sources within a distance of $\sim$ 100 Mpc is strongly supported. But if this is indeed the case, these sources should be identified.

[11] Shock fronts are surfaces in a gas flow where the parameters of state for the gas, such as pressure, density, and temperature, change discontinuously. The standard example for a shock front is the bang in an explosion, where a spherical shock wave propagates outwards from the point of explosion. Another example is the sonic boom caused, for example, by airplanes that move at a speed exceeding the velocity of sound. Such shock fronts are solutions of the hydrodynamic equations. They occur frequently in astrophysics, e.g., in explosion phenomena such as supernovae or in rapid (i.e., supersonic) flows such as those we will discuss in the context of AGNs.

[12] The Pierre Auger Observatory in Argentina combines 1600 surface detectors for the detection of particles from air showers, generated by cosmic rays hitting the atmosphere, with 24 optical telescopes measuring the optical light produced by these air showers. The detectors are spread over an area of 3000 km^2, with a spacing between detectors of 1.5 km, small enough to resolve the structure of air showers which is needed to determine the direction of the incoming cosmic ray. Starting regular observations in 2004, Auger has already led to breakthroughs in cosmic ray research.

Indeed, a correlation between the arrival direction of UHECRs and the direction of nearby AGN has been found, providing evidence that these are the places in which particles can be accelerated to such high energies. From a statistical analysis of this correlation, the typical angular separation between the cosmic ray and the corresponding AGN is estimated to be $\sim 3°$, which may be identified with the deflection of direction that a cosmic ray experiences on its way to Earth, most likely due to magnetic fields. Whereas substantially increased statistics, possible with accumulating data, is needed to confirm this correlation, the big puzzle about the UHECRs may have found a solution.

Energy density. It is interesting to realize that the energy densities of cosmic rays, the magnetic field, the turbulent energy of the ISM, and the electromagnetic radiation of the stars are about the same—as if an equilibrium between these different components has been established. Since these components interact with each other—e.g., the turbulent motions of the ISM can amplify the magnetic field, and vice versa, the magnetic field affects the velocity of the ISM and of cosmic rays—it is not improbable that these interaction processes can establish an equipartition of the energy densities.

Gamma radiation from the Milky Way. The Milky Way emits γ-radiation, as can be seen in Fig. 1.8. There is diffuse γ-ray emission which can be traced back to the cosmic rays in the Galaxy. When these energetic particles collide with nuclei in the interstellar medium, radiation is released. This gives rise to a continuum radiation which closely follows a power-law spectrum, such that the observed flux S_ν is $\propto \nu^{-\alpha}$, with $\alpha \sim 2$. The quantitative analysis of the distribution of this emission provides the most important information about the spatial distribution of cosmic rays in the Milky Way.

Gamma-ray lines. In addition to the continuum radiation, one also observes line radiation in γ-rays, at energies below ~ 10 MeV. The first detected and most prominent line has an energy of 1.809 MeV and corresponds to a radioactive decay of the Al^{26} nucleus. The spatial distribution of this emission is strongly concentrated towards the Galactic disk and thus follows the young stellar population in the Milky Way. Since the lifetime of the Al^{26} nucleus is short ($\sim 10^6$ yr), it must be produced near the emission site, which then implies that it is produced by the young stellar population. It is formed in hot stars and released to the interstellar medium either through stellar winds or core-collapse supernovae. Gamma-lines from other radioactive nuclei have been detected as well.

Annihilation radiation from the Galaxy. Furthermore, line radiation with an energy of 511 keV has been detected in the Galaxy. This line is produced when an electron and a positron annihilate into two photons, each with an energy corresponding to the rest-mass energy of an electron, i.e., 511 keV.[13] This annihilation radiation was identified first in the 1970s. With the Integral satellite, its emission morphology has been mapped with an angular resolution of $\sim 3°$. The 511 keV line emission is detected both from the Galactic disk and the bulge. The angular resolution is not sufficient to tell whether the annihilation line traces the young stellar population (i.e., the thin disk) or the older population in the thick disk. However, one can compare the distribution of the annihilation radiation with that of Al^{26} and other radioactive species. In about 85 % of all decays Al^{26} emits a positron. If this positron annihilates close to its production site one can predict the expected annihilation radiation from the distribution of the 1.809 MeV line. In fact, the intensity and angular distribution of the 511 keV line from the disk are compatible with this scenario for the generation of positrons.

The origin of the annihilation radiation from the bulge, which has a luminosity larger than that from the disk by a factor ~ 5, is unknown. One needs to find a plausible source for the production of positrons in the bulge. There is no unique answer to this problem at present, but Type Ia supernovae and energetic processes near low-mass X-ray binaries are prime candidates for this source.

2.3.5 The Galactic bulge

The Galactic bulge is the central thickening of our Galaxy. Figure 1.2 shows another spiral galaxy from its side, with its bulge clearly visible. Compared to that, the bulge in the Milky Way is far more difficult to identify in the optical, as can be seen in Fig. 2.1, owing to obscuration. However, in the near-IR, it clearly sticks out (Fig. 1.8). The characteristic scale-length of the bulge is ~ 1 kpc. Owing to the strong extinction in the disk, the bulge is best observed in the IR. The extinction to the Galactic center in the visual is $A_V \sim 28$ mag. However, some lines-of-sight close to the Galactic center exist where A_V is significantly smaller, so that observations in optical and near-IR light are possible, e.g., in Baade's Window, located about 4° below the Galactic center at $\ell \sim 1°$, for which $A_V \sim 2$mag (also see Sect. 2.6).

From the observations by COBE, and also from Galactic microlensing experiments (see Sect. 2.5), we know that our bulge has the shape of a peanut-shaped bar, with the major axis pointing away from us by about 25°. The scale-height of the bulge is ~ 400 pc, with an axis-ratio of $\sim 1 : 0.35 : 0.26$.

As is the case for the exponential profiles that describe the light distribution in the disk, the functional form of the brightness distribution in the bulge is also suggested from

[13] In addition to the two-photon annihilation, there is also an annihilation channel in which three photons are produced; the corresponding radiation forms a continuum spectrum, i.e., no spectral lines.

observations of other spiral galaxies. The profiles of their bulges, observed from the outside, are much better determined than in our Galaxy where we are located amid its stars.

The de Vaucouleurs profile. The brightness profile of our bulge can be approximated by the de Vaucouleurs law which describes the surface brightness I as a function of the projected distance R from the center,

$$\boxed{\log\left(\frac{I(R)}{I_\mathrm{e}}\right) = -3.3307\left[\left(\frac{R}{R_\mathrm{e}}\right)^{1/4} - 1\right]}, \qquad (2.40)$$

with $I(R)$ being the measured surface brightness, e.g., in $[I] = L_\odot/\mathrm{pc}^2$. R_e is the effective radius, defined such that half of the luminosity is emitted from within R_e,

$$\int_0^{R_\mathrm{e}} \mathrm{d}R\; R\, I(R) = \frac{1}{2}\int_0^{\infty} \mathrm{d}R\; R\, I(R)\,. \qquad (2.41)$$

This definition of R_e also leads to the numerical factor on the right-hand side of (2.40). As one can easily see from (2.40), $I_\mathrm{e} = I(R_\mathrm{e})$ is the surface brightness at the effective radius. An alternative form of the de Vaucouleurs law is

$$\boxed{I(R) = I_\mathrm{e}\,\exp\left(-7.669\left[(R/R_\mathrm{e})^{1/4} - 1\right]\right)}\,. \qquad (2.42)$$

Because of its mathematical form, it is also called an $r^{1/4}$ law. The $r^{1/4}$ law falls off significantly more slowly than an exponential law for large R. For the Galactic bulge, one finds an effective radius of $R_\mathrm{e} \approx 0.7\,\mathrm{kpc}$. With the de Vaucouleurs profile, a relation between luminosity, effective radius, and surface brightness is obtained by integrating over the surface brightness,

$$\boxed{L = \int_0^{\infty} \mathrm{d}R\; 2\pi R\, I(R) = 7.215\pi I_\mathrm{e}\, R_\mathrm{e}^2}\,. \qquad (2.43)$$

Stellar age distribution in the bulge. The stars in the bulge cover a large range in metallicity, $-1 \lesssim [\mathrm{Fe/H}] \lesssim +0.6$, with a mean of about 0.3, i.e., the mean metallicity is about twice that of the Sun. The metallicity also changes as a function of distance from the center, with more distant stars having a smaller value of [Fe/H].

The high metallicity means that either the stars of the bulge formed rather late, according to the age-metallicity relation, or that it is an old population with very intense star formation activities at an early cosmic epoch. We can distinguish between these two possibilities from the chemical composition of stars in the bulge, obtained from spectroscopy. This is shown in Fig. 2.16, where the magnesium-to-iron ratio is shown for stars in the bulge and compared to disk stars. Obviously, bulge stars have a significantly higher abundance of Mg, relative to iron, than the stars from the thin disk, but much more similar to thick disk stars. Recalling the discussion of the chemical enrichment of the interstellar medium by supernovae in Sect. 2.3.2, this implies that the enrichment must have occurred predominantly by core-collapse supernovae, since they produce a high ratio of α-elements (like magnesium) compared to iron, whereas Type Ia SNe produce mainly iron-group elements. Therefore, most of the bulge stars must have formed before the Type Ia SNe exploded. Whereas the time lag between the birth of a stellar population and the explosion of the bulk of Type Ia SN is not well known (it depends on the evolution of binary systems), it is estimated to be between 1 and 3 Gyr. Hence, most of the bulge stars must have formed on a rather short time-scale: the bulge consists mainly of an old stellar population, formed within $\sim$ 1 Gyr. This is also confirmed with the color-magnitude diagram of bulge stars from which an age of 10 ± 2.5 Gyr is determined.

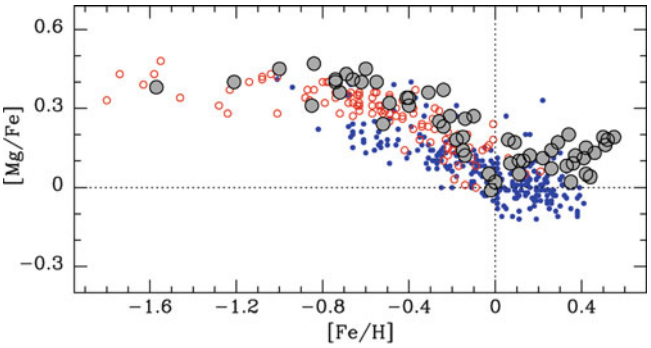


Fig. 2.16 The ratio of magnesium and iron, as a function of metallicity [Fe/H]. *Filled grey circles* correspond to bulge stars, *red* (*blue*) *circles* show nearby stars from the thick (thin) disk. The *dotted lines* corresponds to the Solar value. Source: T. Bensby et al. 2013, *Chemical evolution of the Galactic bulge as traced by microlensed dwarf and subgiant stars. V. Evidence for a wide age distribution and a complex MDF*, A&A 549, A147, Fig. 27. ©ESO. Reproduced with permission

However, in the region of the bulge, one also finds stars that kinematically belong to the disk and the halo, as both extend to the inner region of the Milky Way. The thousands of RR Lyrae stars found in the bulge, for example, have a much lower metallicity than typical bulge stars and may well belong to the innermost region of the stellar halo, and younger stars may be part of the disk population.

The mass of the bulge is about $M_\mathrm{bulge} \sim 1.6 \times 10^{10} M_\odot$ and its luminosity is $L_{B,\mathrm{bulge}} \sim 3 \times 10^9 L_\odot$, which results in a stellar mass-to-light ratio of

$$\boxed{\frac{M}{L} \approx 5\frac{M_\odot}{L_\odot} \text{ in the bulge}}\,, \qquad (2.44)$$

larger than that of the thin disk.

2.3.6 The stellar halo

The visible halo of our Galaxy consists of about 150 *globular clusters* and field stars with a high velocity component perpendicular to the Galactic plane. A globular cluster is a collection of typically several hundred thousand stars, contained within a spherical region of radius $\sim$ 20 pc. The stars in the cluster are gravitationally bound and orbit in the common gravitational field. The old globular clusters with [Fe/H] < -0.8 have an approximately spherical distribution around the Galactic center. A second population of globular clusters exists that contains younger stars with a higher metallicity, [Fe/H] > -0.8. They have a more oblate geometrical distribution and are possibly part of the thick disk because they show roughly the same scale-height. The total mass of the stellar halo in the radius range between 1 and 40 kpc is $\sim 4 \times 10^8 M_\odot$.

Most globular clusters are at a distance of $r \lesssim 35$ kpc (with $r = \sqrt{R^2 + z^2}$) from the Galactic center, but some are also found at $r > 60$ kpc. At these distances it is hard to judge whether these objects are part of the Galaxy or whether they have been captured from a neighboring galaxy, such as the Magellanic Clouds. Also, field stars have been found at distances out to $r \sim 50$ kpc, which is the reason why one assumes a characteristic value of $r_{\rm halo} \sim 50$ kpc for the extent of the visible halo.

The *density distribution* of metal-poor globular clusters and field stars in the halo is described by

$$n(r) \propto r^{-\gamma} \ , \tag{2.45}$$

with a slope γ in the range 3–3.5. Alternatively, one can fit a de Vaucouleurs profile to the density distribution, which results in an effective radius of $r_{\rm e} \sim 2.7$ kpc. Star counts from the Sloan Digital Sky Survey provided clear indications that the stellar halo of the Milky Way is flattened, i.e., it is oblate, with an axis ratio of the smallest axis (in the direction of the rotation axis) to the longer ones being $q \sim 0.6$.

Furthermore, the SDSS discovered the fact that the stellar halo is highly structured: the distribution of stars in the halo is not smooth, but local over- and underdensities are abundant. Several so-called stellar streams were found, regions of stellar overdensities with the shape of a long and narrow cylinder. These stellar streams can in some cases be traced back to the disruption of a low-mass satellite galaxy of the Milky Way by tidal gravitational forces, most noticeably to the Sagittarius dwarf spheroidal (Sgr dSph).

Tidal disruption. Consider a system of gravitationally bound particles, such as a star cluster, a star, or a gas cloud, moving in a gravitational field. The trajectory of the system is determined by the gravitational acceleration. However, since the system is extended, particles in the outer part of the system experience a different gravitational acceleration than the center of mass. Hence, in the rest frame of the moving system, there is a net acceleration of the particles away from the center, due to tidal gravitational forces. The best-known example of this are the tides on Earth: whereas the Earth is freely falling in the gravitational field caused by the Sun (and the Moon), matter on its surface experiences a net force, since the gravitational field is inhomogeneous, giving rise to the tides. If this net force for particles in the outer part of the system is directed outwards, and stronger than the gravitational force binding the particles to the system, these particles will be removed from the system—the system will lose particles due to this *tidal stripping*.

Condition for tidal disruption. We can consider this process more quantitatively. Consider a spherical system of mass M and radius R, so the gravitational acceleration on the surface is $a_{\rm s} = -GM/R^2$, directed inwards. If $\phi(\boldsymbol{r})$ is the gravitational potential in which this system moves, the tidal acceleration $\boldsymbol{a}_{\rm tid}$ is the difference between the acceleration $-\nabla\phi$ at the surface of the system and that at its center,

$$\boldsymbol{a}_{\rm tid}(\boldsymbol{R}) = \boldsymbol{a}(\boldsymbol{r} + \boldsymbol{R}) - \boldsymbol{a}(\boldsymbol{r}) \ ,$$

where $\boldsymbol{R}$ is a vector from the center of the system to its surface, i.e., $|\boldsymbol{R}| = R$. A first-order Taylor expansion of the term on the r.h.s. yields for the i-component of the tidal acceleration

$$a_{{\rm tid},i} = -\sum_{j=1}^{3} \frac{\partial^2 \phi}{\partial r_i \partial r_j} R_j \equiv -\sum_{j=1}^{3} \phi_{,ij} R_j \ ,$$

where we made use of the fact that $\boldsymbol{a} = -\nabla\phi$, and the derivatives are taken at the center of the system. In the final step, we abbreviated the matrix of second partial derivatives of ϕ with $\phi_{,ij}$. This matrix is symmetric, and therefore one can always rotate to a coordinate system in which this matrix is diagonal. We will assume now that the local matter density ρ causing the potential ϕ vanishes; then, from the Poisson equation $\nabla^2\phi = 4\pi G\rho$, we find that the sum of the diagonal elements of $\phi_{,ij}$ is zero. Furthermore, we assume that the tidal field is axially symmetric, with the r_1-axis being the axis of symmetry. In this case, we can write the tidal matrix as $\phi_{,ij} = {\rm diag}(-2t, t, t)$. Writing the radius vector as $\boldsymbol{R} = R(\cos\theta, \sin\theta, 0)$, i.e., restricting it to the r_1-r_2-plane, the tidal acceleration becomes $\boldsymbol{a}_{\rm tid} = tR(2\cos\theta, -\sin\theta, 0)$. The radial component of the tidal acceleration is obtained by projecting $\boldsymbol{a}_{\rm tid}$ along the radial direction,

$$a_{\rm tid,r} = \boldsymbol{a}_{\rm tid} \frac{\boldsymbol{R}}{|\boldsymbol{R}|} = tR\left(2\cos^2\theta - \sin^2\theta\right) = tR\left(3\cos^2\theta - 1\right) \ .$$

The total radial acceleration is then

$$a_{\rm tot,r} = -\frac{GM}{R^2} + a_{\rm tid,r} \ .$$

If this is positive, the net force on a particle is directed outwards, and the particle is stripped from the system. Obviously, $a_{\rm tid,r}$ depends on the position on the surface, here described by θ. Note that the radial component of the tidal acceleration is symmetric under $\theta \to \theta + \pi$, i.e., is the same at opposite points on the sphere. This is in agreement with the observation that the tide gauge has two maxima and two minima at any time on the Earth surface, so that the period of the tidal motion is 12 h, i.e., half a day. Also note that in some regions on the surface, the tidal acceleration is directed inwards, and directed outwards at other points. If there is one point where the total radial acceleration is positive,

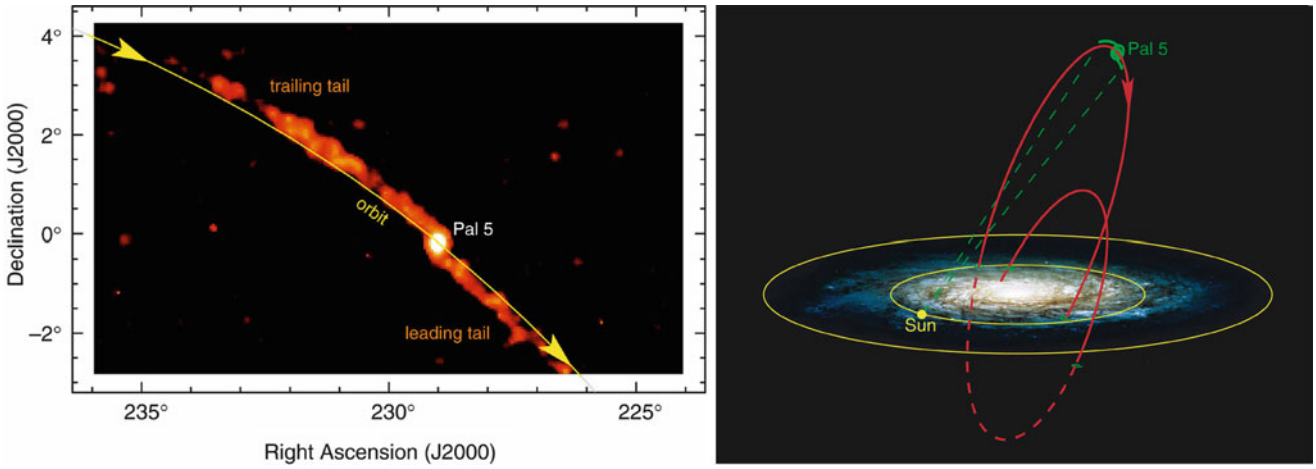


Fig. 2.17 Tidal disruption of the globular cluster Palomar 5. *Left panel*: The white blob shows the globular cluster, from which the two tidal tails emerge, shown in *orange*. These contain more mass than the cluster itself at the current epoch, meaning the cluster has lost more than half its original mass. The tidal tails delineate the cluster's orbit around the Galaxy, which is sketched in the *right panel* as the *red curve*, with the current position of Pal 5 indicated in *green*. Credit: M. Odenkirchen, E. Grebel, Max-Planck-Institut für Astronomie, and the Sloan Digital Sky Survey Collaboration

i.e., directed outwards, the system will lose mass. Assuming $t > 0$, this happens if $2tR > GM/R^2$. In other words, for a system to be stable against tidal stripping, one must have

$$t < \frac{GM}{2R^3} = \frac{2\pi G}{3}\bar{\rho}\,, \tag{2.46}$$

where in the final expression we inserted the mean density $\bar{\rho}$ of the system. Hence, for a given mean density of a system, the tidal gravitational field must not be larger than (2.46) in order for the system to remain stable against tidal stripping.

One application of the foregoing treatment is the disruption of a system in the field of a point mass M_p, given by $\phi(\boldsymbol{r}) = -GM_\mathrm{p}/|\boldsymbol{r}|$. If we choose the system to be located on the r_1-axis, the tidal matrix $\phi_{,ij}$ is diagonal and reads

$$\phi_{,ij} = (GM_\mathrm{p}/r^3)\,\mathrm{diag}(2,-1,-1)\,.$$

Thus, the system is disrupted if

$$\frac{2GM_\mathrm{p}}{r^3} > \frac{GM}{R^3}\,. \tag{2.47}$$

We will return to this example when we consider the tidal disruption of stars in the gravitational field of a black hole.

On its orbit through the Milky Way, a satellite galaxy or a star cluster will experience a tidal force which varies with time. When it gets closer to the center, or to the disk, one expects the tidal field to get stronger than on other parts of the orbit. Depending on its mean density and its orbit, such a system will lose mass in the course of time. This is impressively seen in the globular cluster Pal 5, where the SDSS has found two massive tidal tails of stars that were removed from the cluster due to tidal forces (Fig. 2.17). The 180°-symmetry of the tidal force mentioned before leads to the occurrence of two almost symmetric tidal tails, one moving slightly faster than the cluster (the leading tail), the other slower (trailing tail). The tidally stripped stars form such coherent structures since their velocity dispersion is very small, comparable to the velocity dispersion of the globular cluster. This explains why such tidal streams form a distinct feature for a long time. Since the tidal tails of Pal 5 contain more stellar mass than the remaining cluster, the latter has lived through the best part of its life and will be totally disrupted within its next few orbits around the Galactic center.

As mentioned above, other stellar streams similar to that of Pal 5 have been found, the clearest one being that related to the tidal disruption of Sgr dSph. The corresponding tidal stream is observed to create a full great circle on the sky; a part of it is shown in Fig. 2.18.

As we will discuss later (see Chap. 10), the strong substructure of the stellar halo is expected from our understanding of the evolution of galaxies where galaxies grow in mass through mergers with other galaxies. In this model, the observed substructure are remnants of low-mass galaxies which were accreted onto the Milky Way at some earlier time—in agreement with the discussion above on the possible origin of the thick disk.

2.3.7 The gaseous halo

Besides a stellar component, also gas in various phases is seen outside the disk of the Milky Way. The gas is detected either by its emission or by absorption lines in the spectra of sources located at larger distances. When observing gas, either in emission or absorption, its distance to us is at first unknown, and must be inferred indirectly.

Infalling gas clouds. Neutral hydrogen is observed outside the Galactic disk, in the form of clouds. Most of these clouds, visible in 21 cm line emission, have a negative radial velocity, i.e., they are moving towards us, with velocities of up to

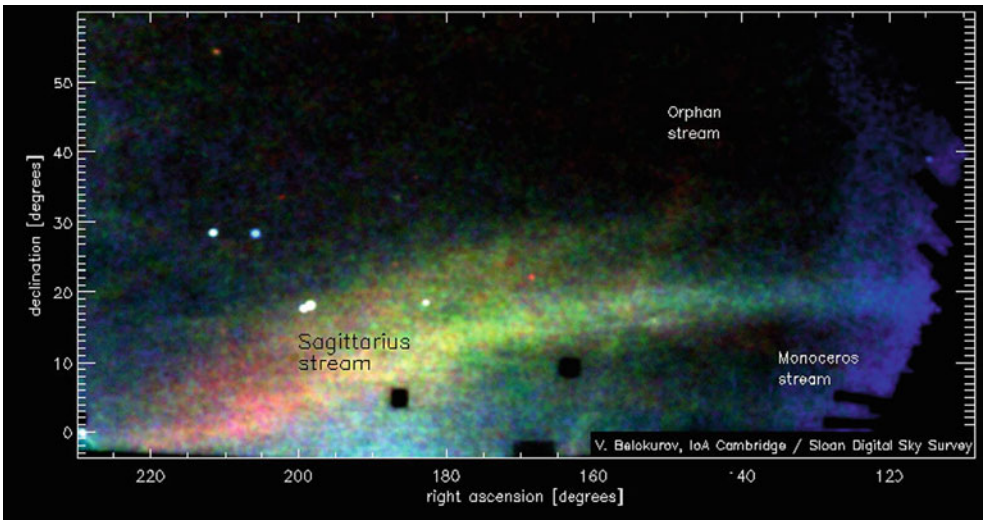


Fig. 2.18 The "Field of Streams", as detected in the SDSS survey. Shown is the two-dimensional distribution of stars, which were color selected by $g - r < 0.4$, and magnitude selected by $19 \leq r \leq 22$. The color selection yields the bluest stars in an old stellar population corresponding to those whose main-sequence lifetime equals the age of the population; hence, they are main sequence turn-off stars. The range in magnitude then corresponds to a corresponding range in distance. The distances are color-coded in this figure, with *blue* corresponding to the nearest stars at $D \sim 10$ kpc, and *red* to the most distant ones at $D \sim 30$ kpc. One sees that the density of stars is far from uniform, but that several almost one-dimensional overdensities are easily identified. The most prominent of these streams, the Sagittarius stream, corresponds to stars which have been tidally stripped from the Sgr dSph. There is a clear distance gradient along the stream visible, with the most distant stars in the lower left of the image. Note that this image covers almost a quarter of the sky. Credit: Vasily Belokurov, SDSS-II Collaboration

$v_r \sim -400$ km/s. Based on their observed velocity, these *high-velocity clouds* (HVCs) cannot be following the general Galactic rotation. In addition, there are clouds with smaller velocities, the intermediate-velocity clouds (IVCs).

These clouds are often organized in big structures on the sky, the largest of which are located close to the Magellanic Clouds (Fig. 2.19). This gas forms the Magellanic Stream, a narrow band of HI emission which follows the Magellanic Clouds along their orbit around the Galaxy (see Fig. 2.20). This gas stream may be the result of a close encounter of the Magellanic Clouds with the Milky Way in the past. The (tidal) gravitational force that the Milky Way had imposed on our neighboring galaxies in such an encounter could strip away part of their interstellar gas. For the Magellanic Stream, the distance can be assumed to coincide with the distance to the Magellanic Clouds.

For the other HVCs, which are not associated with a stellar structure, distances can be estimated through absorption. If we consider a set of stars near to the line-of-sight to a hydrogen cloud, located at different distances from us, then those at distances larger than the gas will show absorption lines caused by the gas (with the same radial velocity, or Doppler shift, as the emission of the gas), and those which are closer will not. Hence, from the interstellar absorption lines of stars in the Galactic halo, the distances to the HVCs can be inferred. These studies became possible after the Sloan Digital Sky Survey, and other imaging surveys, identified a large number of halo stars, so that we now have a pretty good three-dimensional picture of this gas distribution.

Most of the HVCs are at distances between 2 and 15 kpc from us, and within $\sim$ 10 kpc of the Galactic disk. Based on the line width, indicating the thermal velocity of the gas, its temperature is characteristic of a warm neutral medium, $T \sim 10^4$ K, but narrower line components in some HVCs show that cooled gas must be present as well. This neutral hydrogen has a large covering fraction, i.e., more than a third of our sky is covered down to a column density of 2×10^{17} cm^{-2} in neutral hydrogen atoms.

The neutral gas in the HVCs is often associated with optical emission in the Hα line. This emission line is produced in the process of hydrogen recombination, from which one concludes that the hydrogen clouds are partially ionized, most likely due to ionizing radiation from hot stars in the Galactic disk. The total mass contained in the HVCs can be estimated to amount to $\sim 7 \times 10^7 M_\odot$, if it is assumed that their neutral fraction overall is about 50 %. The hydrogen gas associated with the Magellanic Stream contains a mass at least four times this value.

Warm and hot gas. Beside the relatively cold neutral gas seen in the HVCs, there is hotter gas at large distances from the Galactic plane. Gas with temperatures of $T \sim 10^5$ K is observed through absorption lines of highly ionized species in optical and UV spectra of distance sources, like quasars. Indeed, a covering fraction larger than $\sim$ 60 % is found for absorption by doubly ionized silicon (SiIII) and by five times ionized oxygen (OVI). The temperature of the gas can be estimated if several different ions are detected in

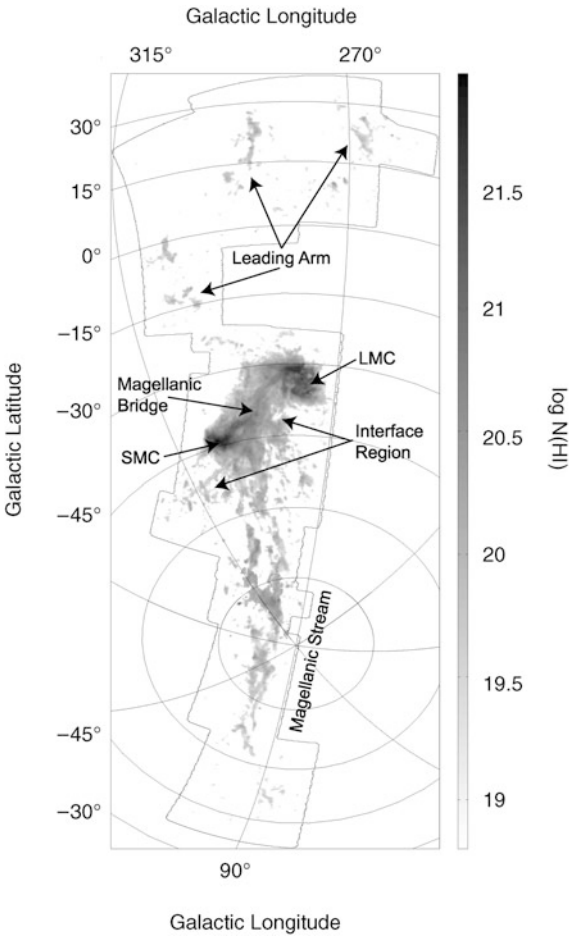


Fig. 2.19 HI-map of a large region in the sky containing the Magellanic Clouds. This map is part of a large survey of HI, observed through its 21 cm line emission, that was performed with the Parkes telescope in Australia, and which maps about a quarter of the Southern sky with a pixel size of $5'$ and a velocity resolution of $\sim 1\,\mathrm{km/s}$. The emission from gas at Galactic velocities was removed in this map. Besides the HI emission by the Magellanic Clouds themselves, gas between them is visible, the Magellanic Bridge and the Magellanic Stream, the latter connected to the Magellanic Clouds by an 'Interface Region'. Gas is also found in the direction of the orbital motion of the Magellanic Clouds around the Milky Way, forming the 'Leading Arm'. Source: C. Brüns et al. 2005, *The Parkes HI Survey of the Magellanic System*, A&A 432, 45, p. 50, Fig. 2. ©ESO. Reproduced with permission

absorption; this then also allows one to determine the total column density of gas. It is estimated that this gas component has a metallicity of $\sim$ 0.2 Solar, and a total mass of $\sim 10^8 M_\odot$.

Hotter gas, with $T \sim 10^6\,\mathrm{K}$, is seen from its X-ray emission, as well as through absorption lines of OVII and O VIII. Most of this gas that we see in emission is believed to be within a few kiloparsecs of the Galactic disk, but there is evidence that some gas extends to larger distances. The presence of this hot gas component is also evidenced by the morphology of some HVCs, which show a head-tail structure (not unlike that of comets), best explained if the hydrogen cloud moves through an ambient medium which compresses its head, and gradually strips off gas from the cloud, which forms the tail.

The gas visible outside the disk constitutes about 10 % of the total interstellar medium in the Milky Way and thus presents a significant reservoir of gas. Some of this gas is believed to have been expelled from the Galactic disk, through outflows generated by supernova explosions, based on theoretical expectations and on the measured high metallicity. This gas cools by adiabatic expansion, and returns to the disk under the influence of gravity; this is thought to be a possible origin of IVCs. Since the flow of this gas resembles that of water in a fountain, this scenario is often called the galactic fountain model. Low-metallicity gas, mainly the HVCs, may be coming from outside the Galaxy and be falling into its gravitational potential for the first time. This would then be a fresh supply of gas, out of which stars will be able to form in the future. Indeed, we believe that the mass of the Milky Way is growing also through this accretion of gas, and this is one of the elements of the models of galaxy evolution that we will discuss in Chap. 10. The inflow of gas is estimated to be a few $M_\odot$ per year, comparable to the star-formation rate in the Milky Way.

2.3.8 The distance to the Galactic center

As already mentioned, our distance from the Galactic center is rather difficult to measure and thus not very precisely known. The general problem with such a measurement is the high extinction in the disk, prohibiting measurements of the distance of individual stars close to the Galactic center. Thus, one has to rely on more indirect methods, and the most important ones will be outlined here.

The visible halo of our Milky Way is populated by globular clusters and also by field stars. They have a spherical, or, more generally, a spheroidal distribution. The center of this distribution is obviously identified with the center of gravity of the Milky Way, around which the halo objects are moving. If one measures the three-dimensional distribution of the halo population, the geometrical center of this distribution should correspond to the Galactic center.

This method can indeed be applied because, due to their extended distribution, halo objects can be observed at relatively large Galactic latitudes where they are not too strongly affected by extinction. As was discussed in Sect. 2.2, the distance determination of globular clusters is possible using photometric methods. On the other hand, one also finds RR Lyrae stars in globular clusters to which the period-luminosity relation can be applied. Therefore, the spatial distribution of the globular clusters can be determined. However, at about 150, the number of known globular clusters is relatively small, resulting in a fairly large statistical error

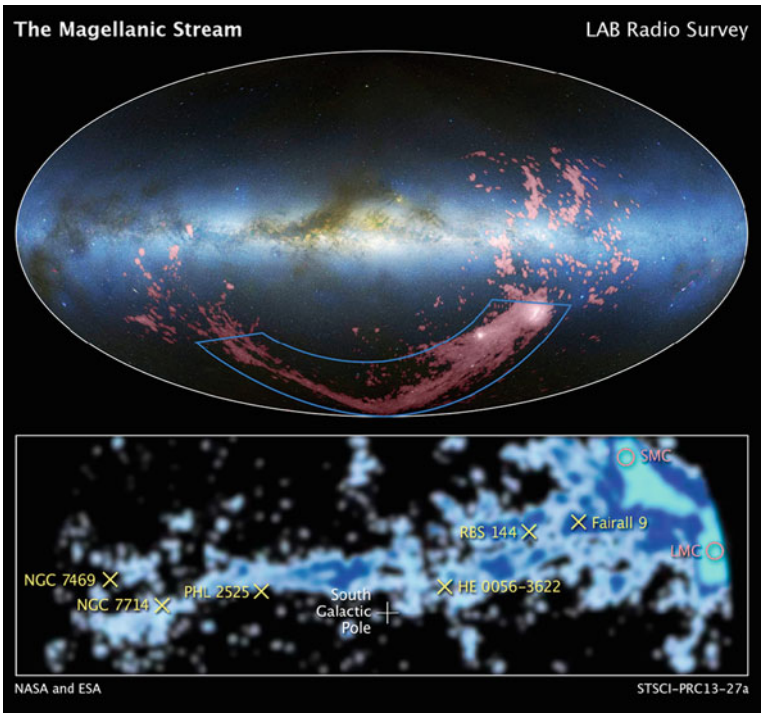


Fig. 2.20 The *image on top* displays the neutral hydrogen distribution belonging to the Magellanic Stream, shown in *pink*, projected onto an optical image of the sky. The Magellanic Clouds are the *two white regions* at the right of the region marked with the *blue box*. The filamentary gas 'above' the Magellanic Clouds is called the 'leading arm', whereas most of the gas between the LMC and SMC is often called the Magellanic Bridge, and the gas connecting this with the Magellanic Stream is called interface region. The *bottom image* is a 21 cm radio map of the that marked region, obtained as part of the Leiden-Argentine-Bonn (LAB) Survey. The crosses mark active galactic nuclei for which UV-spectra were obtained with the Cosmic Origins Spectrograph (COS) onboard HST, to measure the absorption caused by the gas. In particular, the metallicity and chemical composition of the gas was determined. Comparison with the chemical composition of the LMC and SMC shows that the gas of the Magellanic Stream most likely originated from the SMC, from which it was removed by ram-pressure and tidal stripping, though part of the Magellanic Stream was also contributed by the LMC. Credit: David L. Nidever et al., NRAO/AUI/NSF and A. Mellinger, Leiden-Argentine-Bonn (LAB) Survey, Parkes Observatory, Westerbork Observatory, and Arecibo Observatory

for the determination of the common center. Much more numerous are the RR Lyrae field stars in the halo, for which distances can be measured using the period-luminosity relation. The statistical error in determining the center of their distribution is therefore much smaller. On the other hand, this distance to the Galactic center is based only on the calibration of the period-luminosity relation, and any uncertainty in this will propagate into a systematic error on R_0. Effects of the extinction add to this. However, such effects can be minimized by observing the RR Lyrae stars in the NIR, which in addition benefits from the narrower luminosity distribution of RR Lyrae stars in this wavelength regime. These analyses yield a value of $R_0 \approx 8.0\,\mathrm{kpc}$ (see Fig. 2.21).

2.4 Kinematics of the Galaxy

Unlike a solid body, the Galaxy rotates differentially. This means that the angular velocity is a function of the distance R from the Galactic center. Seen from above, i.e., from the NGP, the rotation is clockwise. To describe the velocity field quantitatively we will in the following introduce velocity components in the coordinate system (R, θ, z), as shown in Fig. 2.22. An object following a track $[R(t), \theta(t), z(t)]$ then has the velocity components

$$U := \frac{\mathrm{d}R}{\mathrm{d}t}\,, \quad V := R\frac{\mathrm{d}\theta}{\mathrm{d}t}\,, \quad W := \frac{\mathrm{d}z}{\mathrm{d}t}\,. \tag{2.48}$$

For example, the Sun is not moving on a simple circular orbit around the Galactic center, but currently inwards, $U < 0$, and with $W > 0$, so that it is moving away from the Galactic plane.

In this section we will examine the rotation of the Milky Way. We start with the determination of the velocity components of the Sun. Then we will consider the rotation curve of the Galaxy, which describes the rotational velocity $V(R)$ as a function of the distance R from the Galactic center. We will find the intriguing result that the velocity V does not decline towards large distances, but that it virtually remains constant. Because this result is of extraordinary importance, we will discuss the methods needed to derive it in some detail.

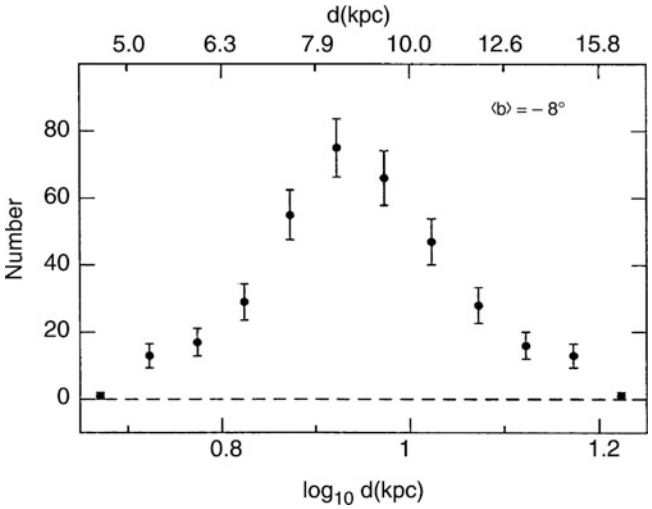


Fig. 2.21 The number of RR Lyrae stars as a function of distance, measured in a direction that closely passes the Galactic center, at $\ell = 0°$ and $b = -8°$. If we assume a spherically symmetric distribution of the RR Lyrae stars, concentrated towards the center, the distance to the Galactic center can be identified with the maximum of this distribution. Source: M. Reid 1993, *The distance to the center of the Galaxy*, ARA&A 31, 345, p. 355. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 31 ©1993 by Annual Reviews www.annualreviews.org

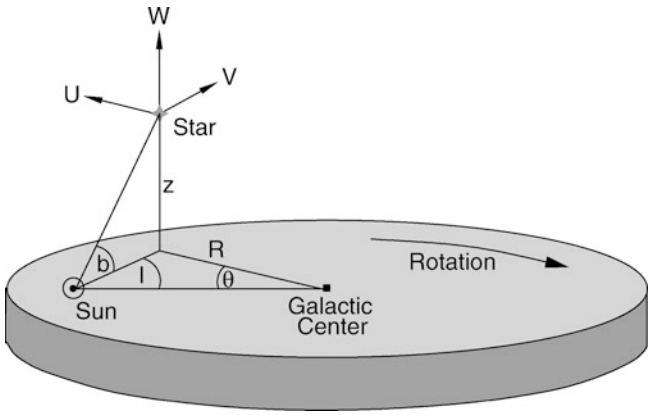


Fig. 2.22 Cylindrical coordinate system (R, θ, z) with the Galactic center at its origin. Note that θ increases in the clockwise direction if the disk is viewed from above. The corresponding velocity components (U, V, W) of a star are indicated. Adopted from B.W. Carroll & D.A. Ostlie 1996, *Introduction to Modern Astrophysics*, Addison-Wesley

2.4.1 Determination of the velocity of the Sun

Local standard of rest. To link local measurements to the Galactic coordinate system (R, θ, z), the *local standard of rest* is defined. It is a fictitious rest-frame in which velocities are measured. For this purpose, we consider a point that is located today at the position of the Sun and that moves along a perfectly circular orbit in the plane of the Galactic disk. The velocity components in the LSR are then by definition,

$$\boxed{U_{\rm LSR} \equiv 0\,, \quad V_{\rm LSR} \equiv V_0\,, \quad W_{\rm LSR} \equiv 0}\,, \tag{2.49}$$

with $V_0 \equiv V(R_0)$ being the orbital velocity at the location of the Sun. Although the LSR changes over time, the time-scale of this change is so large (the orbital period is $\sim 230 \times 10^6$ yr) that this effect is negligible.

Peculiar velocity. The velocity of an object relative to the LSR is called its peculiar velocity. It is denoted by $\boldsymbol{v}$, and its components are given as

$$\boxed{\begin{aligned}\boldsymbol{v} \equiv (u, v, w) &= (U - U_{\rm LSR}, V - V_{\rm LSR}, W - W_{\rm LSR}) \\ &= (U, V - V_0, W)\,.\end{aligned}} \tag{2.50}$$

The velocity of the Sun relative to the LSR is denoted by $\boldsymbol{v}_\odot$. If $\boldsymbol{v}_\odot$ is known, any velocity measured relative to the Sun can be converted into a velocity relative to the LSR: let $\Delta\boldsymbol{v}$ be the velocity of a star relative to the Sun, which is directly measurable using the methods discussed in Sect. 2.2, then the peculiar velocity of this star is

$$\boldsymbol{v} = \boldsymbol{v}_\odot + \Delta\boldsymbol{v}\,. \tag{2.51}$$

Peculiar velocity of the Sun. We consider now an ensemble of stars in the immediate vicinity of the Sun, and assume the Galaxy to be axially symmetric and stationary. Under these assumptions, the number of stars that move outwards to larger radii R equals the number of stars moving inwards. Likewise, as many stars move upwards through the Galactic plane as downwards. If these conditions are not satisfied, the assumption of a stationary distribution would be violated. The mean values of the corresponding peculiar velocity components must therefore vanish,

$$\langle u \rangle = 0\,, \quad \langle w \rangle = 0\,, \tag{2.52}$$

where the brackets denote an average over the ensemble considered. The analog argument is not valid for the v component because the mean value of v depends on the distribution of the orbits: if only circular orbits in the disk existed (with the same orientation as that of the Sun), we would also have $\langle v \rangle = 0$ (this is trivial, since then all stars would have $v = 0$), but this is not the case. From a statistical consideration of the orbits in the framework of stellar dynamics, one deduces that $\langle v \rangle$ is closely linked to the radial velocity dispersion of the stars: the larger it is, the more $\langle v \rangle$ deviates from zero. One finds that

$$\langle v \rangle = -C \left\langle u^2 \right\rangle\,, \tag{2.53}$$

where C is a positive constant that depends on the density distribution and on the local velocity distribution of the stars. The sign in (2.53) follows from noting that a circular orbit has a higher tangential velocity than elliptical orbits, which in addition have a non-zero radial component. Equation (2.53) expresses the fact that the mean rotational velocity of a stellar population around the Galactic center deviates from the corresponding circular orbit velocity, and that the deviation is stronger for a larger radial velocity dispersion. This phenomenon is also known as asymmetric drift. From the mean of (2.51) over the ensemble considered and by using (2.52) and (2.53), one obtains

$$\boxed{\boldsymbol{v}_\odot = \left(-\langle \Delta u \rangle, -C\left\langle u^2 \right\rangle - \langle \Delta v \rangle, -\langle \Delta w \rangle\right)}. \qquad (2.54)$$

One still needs to determine the constant C in order to make use of this relation. This is done by considering different stellar populations and measuring $\langle u^2 \rangle$ and $\langle \Delta v \rangle$ separately for each of them. If these two quantities are then plotted in a diagram (see Fig. 2.23), a linear relation is obtained, as expected from (2.53). The slope C can be determined directly from this diagram. Furthermore, from the intersection with the $\langle \Delta v \rangle$-axis, $v_\odot$ is readily read off. The other velocity components in (2.54) follow by simply averaging, yielding the result:

$$\boxed{\boldsymbol{v}_\odot = (-10, 5, 7)\,\mathrm{km/s}}. \qquad (2.55)$$

Hence, the Sun is currently moving inwards, upwards, and faster than it would on a circular orbit at its location. We have therefore determined $\boldsymbol{v}_\odot$, so we are now able to analyze any measured stellar velocities relative to the LSR. However, we have not yet discussed how V_0, the rotational velocity of the LSR itself, is determined.

Velocity dispersion of stars. The dispersion of the stellar velocities relative to the LSR can now be determined, i.e., the mean square deviation of their velocities from the velocity of the LSR. For young stars (A stars, for example), this dispersion happens to be small. For older K giants it is larger, and is larger still for old, metal-poor red dwarf stars. We observe a very well-defined velocity-metallicity relation which, when combined with the age-metallicity relation, suggests that the oldest stars have the highest peculiar velocities. This effect is observed in all three coordinates and is in agreement with the relation between the age of a stellar population and its scale-height (discussed in Sect. 2.3.1), the latter being linked to the velocity dispersion via σ_z.

Asymmetric drift. If one considers high-velocity stars, only a few are found that have $v > 65\,\mathrm{km/s}$ and which are thus moving much faster around the Galactic center than

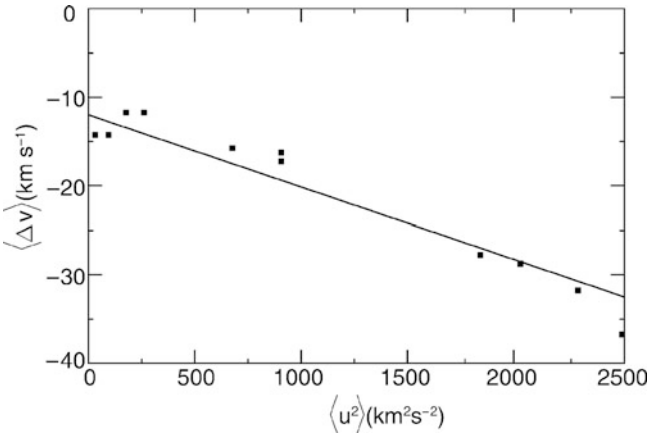


Fig. 2.23 The velocity components $\langle \Delta v \rangle = \langle v \rangle - v_\odot$ are plotted against $\langle u^2 \rangle$ for stars in the Solar neighborhood. Because of the linear relation, $v_\odot$ can be read off from the intersection with the x-axis, and C from the slope. Adopted from B.W. Carroll & D.A. Ostlie 1996, *Introduction to Modern Astrophysics*, Addison-Wesley

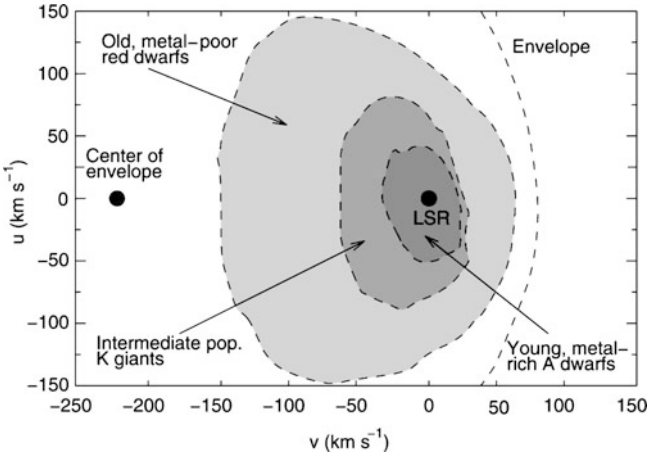


Fig. 2.24 The motion of the Sun around the Galactic center is reflected in the asymmetric drift: while young stars in the Solar vicinity have velocities very similar to the Solar velocity, i.e., small relative velocities, members of other populations (and of other Milky Way components) have different velocities—e.g., for halo objects $v = -220\,\mathrm{km/s}$ on average. Thus, different velocity ellipses show up in a $(u-v)$-diagram. Adopted from B.W. Carroll & D.A. Ostlie 1996, *Introduction to Modern Astrophysics*, Addison-Wesley

the LSR. However, quite a few stars are found that have $v < -250\,\mathrm{km/s}$, so their orbital velocity is opposite to the direction of rotation of the LSR. Plotted in a $(u-v)$-diagram, a distribution is found which is narrowly concentrated around $u = 0\,\mathrm{km/s} = v$ for young stars, as already mentioned above, and which gets increasingly wider for older stars. For the oldest stars, which belong to the halo population, one obtains a circular envelope with its center located at $u = 0\,\mathrm{km/s}$ and $v \approx -220\,\mathrm{km/s}$ (see Fig. 2.24). If we assume that the Galactic halo, to which these high-velocity stars belong, does not rotate (or only very slowly), this asymmetry in the v-distribution can only be caused by the

rotation of the LSR. The center of the envelope then has to be at $-V_0$. This yields the orbital velocity of the LSR

$$\boxed{V_0 \equiv V(R_0) = 220\,\mathrm{km/s}}\,. \tag{2.56}$$

Knowing this velocity, we can then compute the mass of the Galaxy inside the Solar orbit. A circular orbit is characterized by an equilibrium between centrifugal and gravitational acceleration, $V^2/R = GM(<R)/R^2$, so that

$$\boxed{M(<R_0) = \frac{V_0^2 R_0}{G} = 8.8\times 10^{10}\,M_\odot}\,. \tag{2.57}$$

Furthermore, for the orbital period of the LSR, which is similar to that of the Sun, one obtains

$$\boxed{P = \frac{2\pi R_0}{V_0} = 230\times 10^6\,\mathrm{yr}}\,. \tag{2.58}$$

Hence, during the lifetime of the Solar System, estimated to be $\sim 4.6\times 10^9$ yr, it has completed about 20 orbits around the Galactic center.

2.4.2 The rotation curve of the Galaxy

From observations of the velocity of stars or gas around the Galactic center, the rotational velocity V can be determined as a function of the distance R from the Galactic center. In this section, we will describe methods to determine this *rotation curve* and discuss the result.

Decomposition of rotational velocity. We consider an object at distance R from the Galactic center which moves along a circular orbit in the Galactic plane, has a distance D from the Sun, and is located at a Galactic longitude ℓ (see Fig. 2.25). In a Cartesian coordinate system with the Galactic center at the origin, the positional and velocity vectors (we only consider the two components in the Galactic plane because we assume a motion in the plane) are given by

$$\boldsymbol{r} = R\begin{pmatrix}\sin\theta\\ \cos\theta\end{pmatrix}\,,\quad \boldsymbol{V} = \dot{\boldsymbol{r}} = V(R)\begin{pmatrix}\cos\theta\\ -\sin\theta\end{pmatrix}\,,$$

where θ denotes the angle between the Sun and the object as seen from the Galactic center. From the geometry shown in Fig. 2.25 it follows that

$$\boldsymbol{r} = \begin{pmatrix} D\,\sin\ell\\ R_0 - D\,\cos\ell\end{pmatrix}\,.$$

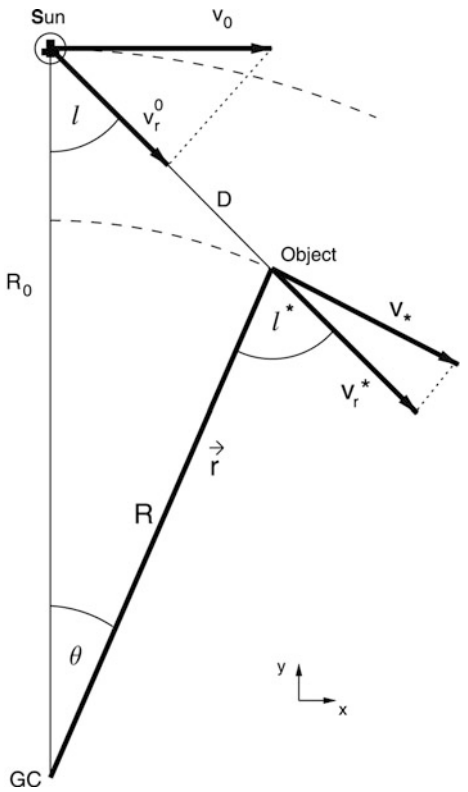


Fig. 2.25 Geometric derivation of the formalism of differential rotation:

$$v_\mathrm{r} = v_\mathrm{r}^* - v_\mathrm{r}^\odot = v_* \sin\ell^* - v_\odot \sin\ell\,,$$
$$v_\mathrm{t} = v_\mathrm{t}^* - v_\mathrm{t}^\odot = v_* \cos\ell^* - v_\odot \cos\ell\,.$$

One has:

$$\frac{\sin\ell}{R} = \frac{\sin(\pi-\ell^*)}{R_0} = \frac{\sin\ell^*}{R_0}\,,$$
$$R\cos\ell^* + D = R_0\cos\ell\,,$$

which implies

$$\begin{aligned} v_\mathrm{r} &= R_0\left(\frac{v_*}{R} - \frac{v_\odot}{R_0}\right)\sin\ell\\ &= (\Omega-\Omega_0)R_0\sin\ell\,,\\ v_\mathrm{t} &= R_0\left(\frac{v_*}{R} - \frac{v_\odot}{R_0}\right)\cos\ell - D\,\frac{v_*}{R}\\ &= (\Omega-\Omega_0)R_0\cos\ell - \Omega D\,. \end{aligned}$$

If we now identify the two expressions for the components of $\boldsymbol{r}$, we obtain

$$\sin\theta = (D/R)\sin\ell\quad,\quad \cos\theta = (R_0/R) - (D/R)\cos\ell\,.$$

If we disregard the difference between the velocities of the Sun and the LSR we get $\boldsymbol{V}_\odot \approx \boldsymbol{V}_\mathrm{LSR} = (V_0, 0)$ in this coordinate system. Thus the relative velocity between the

object and the Sun is, in Cartesian coordinates,

$$\Delta \boldsymbol{V} = \boldsymbol{V} - \boldsymbol{V}_\odot = \begin{pmatrix} V\,(R_0/R) - V\,(D/R)\cos\ell - V_0 \\ -V\,(D/R)\sin\ell \end{pmatrix} .$$

With the angular velocity defined as

$$\Omega(R) = \frac{V(R)}{R} , \tag{2.59}$$

we obtain for the relative velocity

$$\Delta \boldsymbol{V} = \begin{pmatrix} R_0(\Omega - \Omega_0) - \Omega\, D\, \cos\ell \\ -D\,\Omega\, \sin\ell \end{pmatrix} ,$$

where $\Omega_0 = V_0/R_0$ is the angular velocity of the Sun. The radial and tangential velocities of this relative motion then follow by projection of $\Delta \boldsymbol{V}$ along the direction parallel or perpendicular, respectively, to the separation vector,

$$\boxed{v_\mathrm{r} = \Delta \boldsymbol{V} \cdot \begin{pmatrix} \sin\ell \\ -\cos\ell \end{pmatrix} = (\Omega - \Omega_0) R_0 \sin\ell} , \tag{2.60}$$

$$\boxed{v_\mathrm{t} = \Delta \boldsymbol{V} \cdot \begin{pmatrix} \cos\ell \\ \sin\ell \end{pmatrix} = (\Omega - \Omega_0) R_0 \cos\ell - \Omega\, D} . \tag{2.61}$$

A purely geometric derivation of these relations is given in Fig. 2.25.

Rotation curve near R_0, Oort constants. Using (2.60) one can derive the angular velocity by means of measuring v_r, but not the radius R to which it corresponds. Therefore, by measuring the radial velocity alone $\Omega(R)$ cannot be determined. If one measures v_r and, in addition, the proper motion $\mu = v_\mathrm{t}/D$ of stars, then Ω and D can be determined from the equations above, and from D and ℓ one obtains $R = \sqrt{R_0^2 + D^2 - 2R_0 D\cos\ell}$. The effects of extinction prohibits the use of this method for large distances D, since we have considered objects in the Galactic disk. For small distances $D \ll R_0$, which implies $|R - R_0| \ll R_0$, we can make a local approximation by evaluating the expressions above only up to first order in $(R - R_0)/R_0$. In this linear approximation we get

$$\Omega - \Omega_0 \approx \left(\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right)_{|R_0} (R - R_0) , \tag{2.62}$$

where the derivative has to be evaluated at $R = R_0$. Hence

$$v_\mathrm{r} = (R - R_0)\left(\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right)_{|R_0} R_0 \sin\ell ,$$

and furthermore, with (2.59),

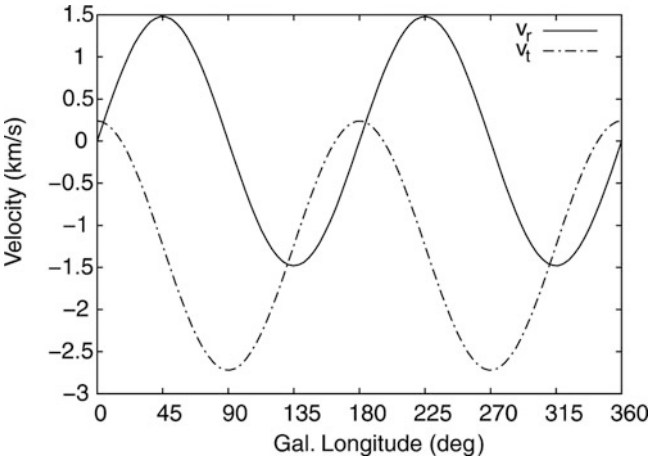


Fig. 2.26 The radial velocity v_r of stars at a fixed distance D is proportional to $\sin 2\ell$; the tangential velocity v_t is a linear function of $\cos 2\ell$. From the amplitude of the oscillating curves and from the mean value of v_t the Oort constants A and B can be derived, respectively [see (2.65)]

$$R_0\left(\frac{\mathrm{d}\Omega}{\mathrm{d}R}\right)_{|R_0} = \frac{R_0}{R}\left[\left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{|R_0} - \frac{V}{R}\right] \approx \left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{|R_0} - \frac{V_0}{R_0} ,$$

in zeroth order in $(R - R_0)/R_0$. Combining the last two equations yields

$$v_\mathrm{r} = \left[\left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{|R_0} - \frac{V_0}{R_0}\right](R - R_0)\sin\ell ; \tag{2.63}$$

in analogy to this, we obtain for the tangential velocity

$$v_\mathrm{t} = \left[\left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{|R_0} - \frac{V_0}{R_0}\right](R - R_0)\cos\ell - \Omega_0\, D . \tag{2.64}$$

For $|R - R_0| \ll R_0$ it follows that $R_0 - R \approx D\cos\ell$; if we insert this into (2.63) and (2.64) we get

$$\boxed{v_\mathrm{r} \approx A\, D\, \sin 2\ell \;\;,\;\; v_\mathrm{t} \approx A\, D\, \cos 2\ell + B\, D} , \tag{2.65}$$

where A and B are the *Oort constants*

$$\boxed{\begin{aligned} A &:= -\frac{1}{2}\left[\left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{|R_0} - \frac{V_0}{R_0}\right] , \\ B &:= -\frac{1}{2}\left[\left(\frac{\mathrm{d}V}{\mathrm{d}R}\right)_{|R_0} + \frac{V_0}{R_0}\right] . \end{aligned}} \tag{2.66}$$

The radial and tangential velocity fields relative to the Sun show a sine curve with period π, where v_t and v_r are phase-shifted by $\pi/4$. This behavior of the velocity field in the Solar neighborhood is indeed observed (see Fig. 2.26). By fitting

the data for $v_r(\ell)$ and $v_t(\ell)$ for stars of equal distance D one can determine A and B, and thus

$$\boxed{\Omega_0 = \frac{V_0}{R_0} = A - B \quad , \quad \left(\frac{dV}{dR}\right)_{|R_0} = -(A+B)} \,. \tag{2.67}$$

The Oort constants thus yield the angular velocity of the Solar orbit and its derivative, and therefore the local kinematical information. If our Galaxy was rotating rigidly so that Ω was independent of the radius, $A = 0$ would follow. But the Milky Way rotates differentially, i.e., the angular velocity depends on the radius. Measurements yield the following values for A and B,

$$\boxed{\begin{aligned} A &= (14.8 \pm 0.8)\ \mathrm{km\,s^{-1}\,kpc^{-1}} \,, \\ B &= (-12.4 \pm 0.6)\ \mathrm{km\,s^{-1}\,kpc^{-1}} \,. \end{aligned}} \tag{2.68}$$

Galactic rotation curve for $R < R_0$; tangent point method. To measure the rotation curve for radii that are significantly smaller than R_0, one has to turn to large wavelengths due to extinction in the disk. Usually the 21 cm emission line of neutral hydrogen is used, which can be observed over large distances, or the emission of CO in molecular gas. These gas components are found throughout the disk and are strongly concentrated towards the plane. Furthermore, the radial velocity can easily be measured from the Doppler effect. However, since the distance to a hydrogen cloud cannot be determined directly, a method is needed to link the measured radial velocities to the distance of the gas from the Galactic center. For this purpose the *tangent point method* is used.

Consider a line-of-sight at fixed Galactic longitude ℓ, with $\cos\ell > 0$ (thus 'inwards'). The radial velocity v_r along this line-of-sight for objects moving on circular orbits is a function of the distance D, according to (2.60). If $\Omega(R)$ is a monotonically decreasing function, v_r attains a maximum where the line-of-sight is tangent to the local orbit, and thus its distance R from the Galactic center attains the minimum value $R_{\min}$. This is the case at

$$D = R_0 \cos\ell \quad , \quad R_{\min} = R_0 \sin\ell \tag{2.69}$$

(see Fig. 2.27). The maximum radial velocity there, according to (2.60), is

$$v_{r,\max} = [\Omega(R_{\min}) - \Omega_0]\ R_0 \sin\ell = V(R_{\min}) - V_0 \sin\ell \,, \tag{2.70}$$

so that from the measured value of $v_{r,\max}$ as a function of direction ℓ, the rotation curve inside R_0 can be determined,

$$\boxed{V(R) = \left(\frac{R}{R_0}\right) V_0 + v_{r,\max}(\sin\ell = R/R_0)} \,. \tag{2.71}$$

In the optical regime of the spectrum this method can only be applied locally, i.e., for small D, due to extinction. This is the case if one observes in a direction nearly tangential to the orbit of the Sun, i.e., if $0 < \pi/2 - \ell \ll 1$ or $0 < \ell - 3\pi/2 \ll 1$, or $|\sin\ell| \approx 1$, so that $R_0 - R_{\min} \ll R_0$. In this case we get, to first order in $(R_0 - R_{\min})$, using (2.69),

$$\begin{aligned} V(R_{\min}) &\approx V_0 + \left(\frac{dV}{dR}\right)_{|R_0} (R_{\min} - R_0) \\ &= V_0 - \left(\frac{dV}{dR}\right)_{|R_0} R_0\,(1 - \sin\ell) \,, \end{aligned} \tag{2.72}$$

so that with (2.70)

$$\boxed{\begin{aligned} v_{r,\max} &= \left[V_0 - \left(\frac{dV}{dR}\right)_{|R_0} R_0\right](1 - \sin\ell) \\ &= 2\,A\,R_0\,(1 - \sin\ell) \,, \end{aligned}} \tag{2.73}$$

where (2.66) was used in the last step. This relation can also be used for determining the Oort constant A.

To determine $V(R)$ for smaller R by employing the tangent point method, we have to observe in wavelength regimes in which the Galactic plane is transparent, using radio emission lines of gas. In Fig. 2.27, a typical intensity profile of the 21 cm line along a line-of-sight is sketched; according to the Doppler effect this can be converted directly into a velocity profile using $v_r = (\lambda - \lambda_0)/\lambda_0$. It consists of several maxima that originate in individual gas clouds. The radial velocity of each cloud is defined by its distance R from the Galactic center (if the gas follows the Galactic rotation), so that the largest radial velocity will occur for gas closest to the tangent point, which will be identified with $v_{r,\max}(\ell)$. Figure 2.28 shows the observed intensity profile of the ^{12}CO line as a function of the Galactic longitude, from which the rotation curve for $R < R_0$ can be read off.

> With the tangent point method, applied to the 21 cm line of neutral hydrogen or to radio emission lines of molecular gas, the rotation curve of the Galaxy inside the Solar orbit, i.e., for $R < R_0$, can be measured.

Rotation curve for $R > R_0$. The tangent point method cannot be applied for $R > R_0$ because for lines-of-sight at

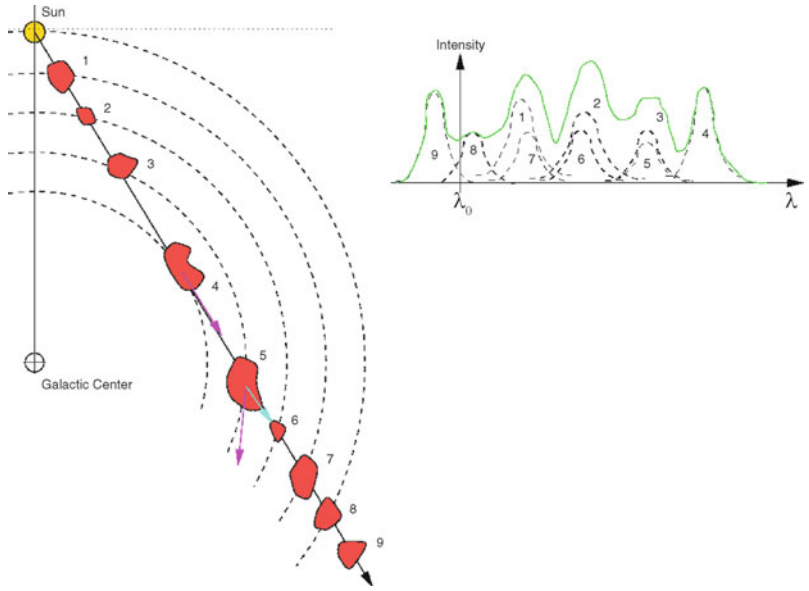


Fig. 2.27 The ISM is optically thin for 21 cm radiation, and thus we receive the 21 cm emission of HI regions from everywhere in the Galaxy. Due to the motion of an HI cloud relative to us, the wavelength is shifted. This can be used to measure the radial velocity of the cloud. With the assumption that the gas is moving on a circular orbit around the Galactic center, one expects that for the cloud in the tangent point (cloud 4), the full velocity is projected along the line-of-sight so that this cloud will therefore have the largest radial velocity. If the distance of the Sun to the Galactic center is known, the velocity of a cloud and its distance from the Galactic center can then be determined. Adopted from B.W. Carroll & D.A. Ostlie 1996, *Introduction to Modern Astrophysics*, Addison-Wesley

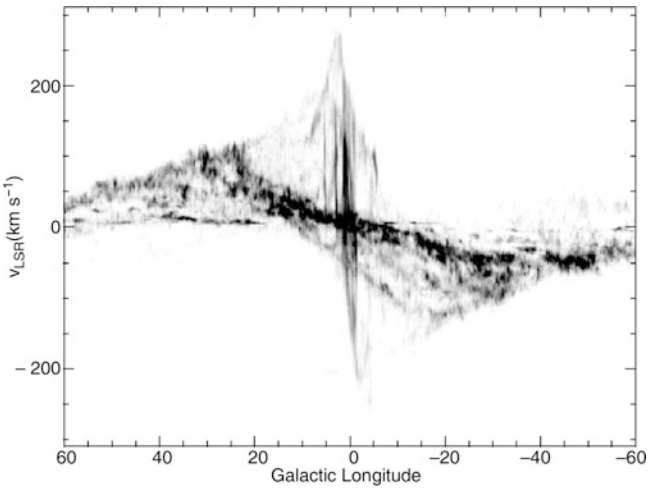


Fig. 2.28 ^{12}CO emission of molecular gas in the Galactic disk. For each ℓ, the intensity of the emission in the $\ell - v_r$ plane is plotted, integrated over the range $-2° \leq b \leq 2°$ (i.e., very close to the middle of the Galactic plane). Since v_r depends on the distance along each line-of-sight, characterized by ℓ, this diagram contains information on the rotation curve of the Galaxy as well as on the spatial distribution of the gas. The maximum velocity at each ℓ is rather well defined and forms the basis for the tangent point method. Source: P. Englmaier & O. Gerhard 1999, *Gas dynamics and large-scale morphology of the Milky Way galaxy*, MNRAS 304, 512, p. 514, Fig. 1. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

$\pi/2 < \ell < 3\pi/2$, the radial velocity v_r attains no maximum. In this case, the line-of-sight is nowhere parallel to a circular orbit.

Measuring $V(R)$ for $R > R_0$ requires measuring v_r for objects whose distance can be determined directly, e.g., Cepheids, for which the period-luminosity relation (Sect. 2.2.7) is used, or O- and B-stars in HII-regions. With ℓ and D known, R can then be calculated, and with (2.60) we obtain $\Omega(R)$ or $V(R)$, respectively. Any object with known D and v_r thus contributes one data point to the Galactic rotation curve. Since the distance estimates of individual objects are always affected by uncertainties, the rotation curve for large values of R is less accurately known than that inside the Solar circle. Recent measurements of blue horizontal-branch stars within the outer halo of the Milky Way by SDSS yielded an estimate of the rotation curve out to $r \sim 60$ kpc. The situation will improve dramatically once the results from Gaia will become available: Gaia will measure distances via trigonometric parallaxes, and proper motions of many star outside the Solar circle.

It turns out that the rotation curve for $R > R_0$ does not decline outwards (see Fig. 2.29) as we would expect from the distribution of visible matter in the Milky Way. Both the stellar density and the gas density of the Galaxy decline exponentially for large R—e.g., see (2.35). This steep radial decline of the visible matter density should imply that $M(R)$, the mass inside R, is nearly constant for $R \gtrsim R_0$, from which a velocity profile like $V \propto R^{-1/2}$ would follow, according to Kepler's law. However, this is not the case: $V(R)$ is virtually constant for $R > R_0$, indicating that $M(R) \propto R$. In fact, a small decrease to about 180 km/s at

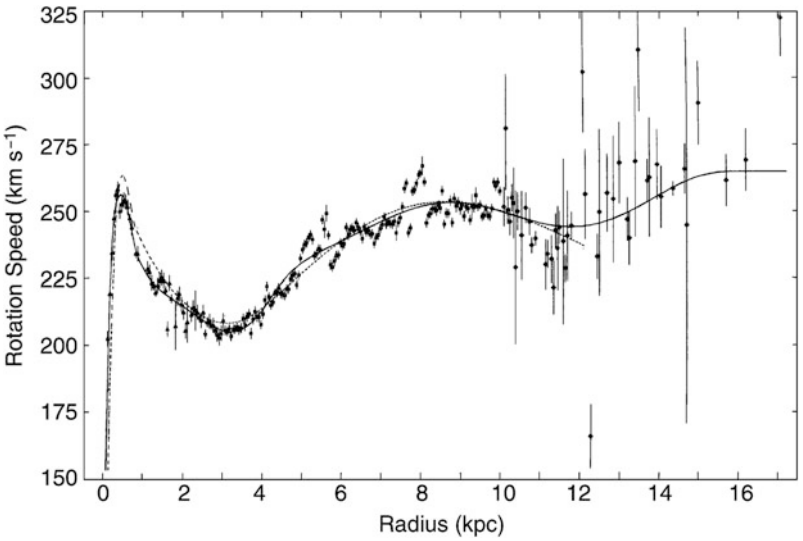


Fig. 2.29 Rotation curve of the Milky Way. Inside the "Solar circle", that is at $R < R_0$, the radial velocity is determined quite accurately using the tangent point method; the measurements outside have larger uncertainties. Source: D. Clemens 1985, *Massachusetts-Stony Brook Galactic plane CO survey—The Galactic disk rotation curve* ApJ 295, 422, p. 429, Fig. 3. ©AAS. Reproduced with permission

$R = 60$ kpc was estimated, corresponding to a total mass of $(4.0 \pm 0.7) \times 10^{11} M_\odot$ enclosed within the inner 60 kpc, but this decrease is *much* smaller than expected from Keplerian rotation. In order to get an almost constant rotational velocity of the Galaxy, much more matter has to be present than we observe in gas and stars.

> The Milky Way contains, besides stars and gas, an additional component of matter that dominates the mass at $R \gtrsim R_0$. Its presence is known only by its gravitational effect, since it has not been observed directly yet, neither in emission nor in absorption. Hence, it is called *dark matter*.

In Sect. 3.3.4 we will see that this is a common phenomenon. The rotation curves of spiral galaxies are flat up to the maximum radius at which they can be measured; *spiral galaxies contain dark matter.* A better way of phrasing is would be to say that the visible galaxy is embedded in a *dark matter halo*, since the total mass of the Milky Way (and other spiral galaxies) is dominated by dark matter.

2.4.3 The gravitational potential of the Galaxy

We have little direct indications about the spatial extent of the dark matter halo, and thus its total mass, because at large radii R there are not many luminous objects whose orbit we can use to measure the rotation curve out there. From the motion of satellite galaxies, such as the Magellanic Clouds, one gets mass estimates at larger distances, but with less accuracy. For example, the mass inside of 100 kpc is estimated to be $(7 \pm 2.5) \times 10^{11} M_\odot$ from such satellite motions. Furthermore, it is largely unknown whether this halo is approximately spherical or deviates significantly from sphericity, being either oblate or prolate. The stellar streams that we discussed in Sect. 2.3.6 above can in principle be used to constrain the axis ratio of the total matter distribution out to large radii—if the gravitational potential of the Milky Way was spherical, the streams would lie in a single orbital plane, so that deviations from it can be used to probe the axis ratio of the potential. However, currently the results from such studies are burdened with uncertainties, and different results are obtained from different studies.

The nature of dark matter is thus far unknown; in principle, we can distinguish two totally different kinds of dark matter candidates:

- *Astrophysical dark matter*, consisting of compact objects—e.g., faint stars like white dwarfs, brown dwarfs, black holes, etc. Such objects were assigned the name MACHOs, which stands for 'MAssive Compact Halo Objects'.
- *Particle physics dark matter*, consisting of elementary particles which thus far have escaped detection in laboratories.

Although the origin of astrophysical dark matter would be difficult to understand (not least because of the baryon abundance in the Universe—see Sect. 4.4.5—and because of the metal abundance in the ISM), a direct distinction between the two alternatives through observation would be of great interest. In the following section we will describe a method which is able to probe whether the dark matter in our Galaxy consists of MACHOs.

2.5 The Galactic microlensing effect: The quest for compact dark matter

In 1986, Bohdan Paczyński proposed to test the possible presence of MACHOs by performing microlensing experi-

ments. As we will soon see, this was a daring idea at that time, but since then such experiments have been carried out. In this section we will mainly summarize and discuss the results of these searches for MACHOs. We will start with a description of the microlensing effect and then proceed with its specific application to the search for MACHOs.

2.5.1 The gravitational lensing effect I

Einstein's deflection angle. *Light, just like massive particles, is deflected in a gravitational field.* This is one of the specific predictions by Einstein's theory of gravity, General Relativity. Quantitatively it predicts that a light beam which passes a point mass M at a distance ξ is deflected by an angle $\hat{\alpha}$, which amounts to

$$\boxed{\hat{\alpha} = \frac{4\,G\,M}{c^2\,\xi}}\,. \tag{2.74}$$

The deflection law (2.74) is valid as long as $\hat{\alpha} \ll 1$, which is the case for weak gravitational fields. If we now set $M = M_\odot$, $\xi = R_\odot$ in the foregoing equation, we obtain

$$\boxed{\hat{\alpha}_\odot \approx 1''.74}$$

for the light deflection at the limb of the Sun. This deflection of light was measured during a Solar eclipse in 1919 from the shift of the apparent positions of stars close to the shaded Solar disk. Its agreement with the value predicted by Einstein made him world-famous over night, because this was the first real and challenging test of General Relativity. Although the precision of the measured value back then was only $\sim 30\,\%$, it was sufficient to confirm Einstein's theory. By now the law (2.74) has been measured in the Solar System with a relative precision of about 5×10^{-4}, and Einstein's prediction has been confirmed.

Not long after the discovery of gravitational light deflection at the Sun, the following scenario was considered. If the deflection was sufficiently strong, light from a very distant source could be visible at two positions in the sky: one light ray could pass a mass concentration, located between us and the source, 'to the right', and the second one 'to the left', as sketched in Fig. 2.30. The astrophysical consequence of this gravitational light deflection is also called the *gravitational lens effect*. We will discuss various aspects of the lens effect in the course of this book, and we will review its astrophysical applications.

The Sun is not able to cause multiple images of distant sources. The maximum deflection angle $\hat{\alpha}_\odot$ is much smaller than the angular radius of the Sun, so that two beams of light that pass the Sun to the left and to the right cannot converge by light deflection at the position of the Earth. Given its

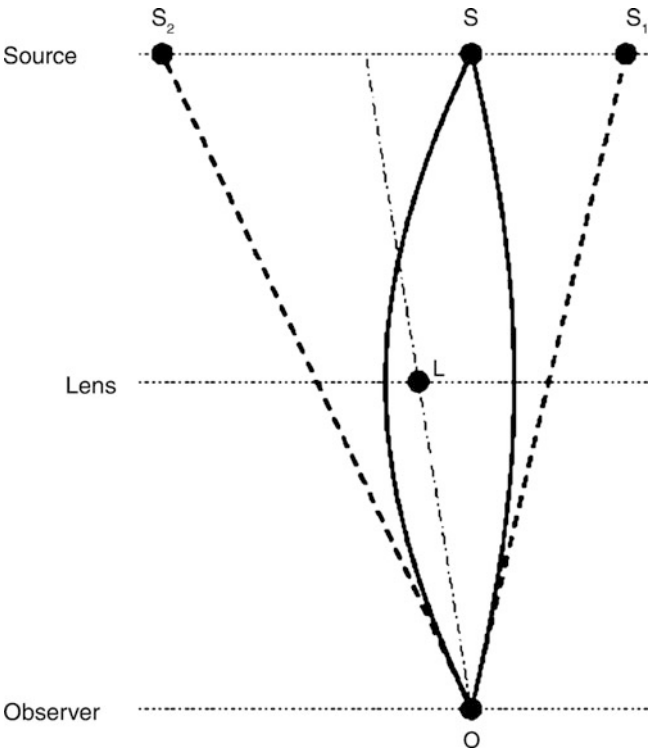


Fig. 2.30 Sketch of a gravitational lens system. If a sufficiently massive mass concentration is located between us and a distant source, it may happen that we observe this source at two different positions on the sphere. Source: J. Wambsganss 1998, *Gravitational Lensing in Astronomy*, Living Review in Relativity 1, 12, Fig. 2. ©Max Planck Society and the author; Living Reviews in Relativity, published by the Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Germany

radius, the Sun is too close to produce multiple images, since its angular radius is (far) larger than the deflection angle $\hat{\alpha}_\odot$. However, the light deflection by more distant stars (or other massive celestial bodies) can produce multiple images of sources located behind them.

Lens geometry. The geometry of a gravitational lens system is depicted in Fig. 2.31. We consider light rays from a source at distance D_s from us that pass a mass concentration (called a lens or deflector) at a separation $\boldsymbol{\xi}$. The deflector is at a distance D_d from us. In Fig. 2.31 $\boldsymbol{\eta}$ denotes the true, two-dimensional position of the source in the source plane, and $\boldsymbol{\beta}$ is the true angular position of the source, that is the angular position at which it would be observed in the absence of light deflection,

$$\boldsymbol{\beta} = \frac{\boldsymbol{\eta}}{D_\mathrm{s}}\,. \tag{2.75}$$

The position of the light ray in the lens plane is denoted by $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$ is the corresponding angular position,

$$\boldsymbol{\theta} = \frac{\boldsymbol{\xi}}{D_\mathrm{d}}\,. \tag{2.76}$$

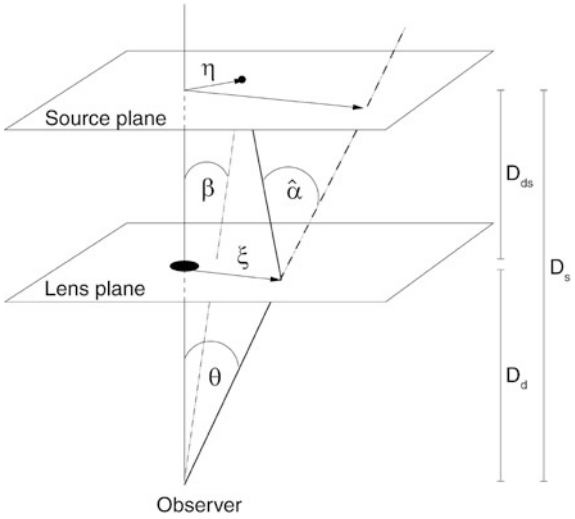


Fig. 2.31 Geometry of a gravitational lens system. Consider a source to be located at a distance D_s from us and a mass concentration at distance D_d. An optical axis is defined that connects the observer and the center of the mass concentration; its extension will intersect the so-called source plane, a plane perpendicular to the optical axis at the distance of the source. Accordingly, the lens plane is the plane perpendicular to the line-of-sight to the mass concentration at distance D_d from us. The intersections of the optical axis and the planes are chosen as the origins of the respective coordinate systems. Let the source be at the location $\boldsymbol{\eta}$ in the source plane; a light beam that encloses an angle $\boldsymbol{\theta}$ to the optical axis intersects the lens plane at the point $\boldsymbol{\xi}$ and is deflected by an angle $\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi})$. All these quantities are two-dimensional vectors. The condition that the source is observed in the direction $\boldsymbol{\theta}$ is given by the lens equation (2.77) which follows from the theorem of intersecting lines. Adapted from: P. Schneider, J. Ehlers & E.E. Falco 1992, *Gravitational Lenses*, Springer-Verlag

Hence, $\boldsymbol{\theta}$ is the observed position of the source on the sphere relative to the position of the 'center of the lens' which we have chosen as the origin of the coordinate system, $\boldsymbol{\xi} = 0$. Like the position vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, the angles $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are two-dimensional vectors, corresponding to the two directions on the sky. D_ds is the distance of the source plane from the lens plane. As long as the relevant distances are much smaller than the 'radius of the Universe' c/H_0, which is certainly the case within our Galaxy and in the Local Group, we have $D_\mathrm{ds} = D_\mathrm{s} - D_\mathrm{d}$. However, this relation is no longer valid for cosmologically distant sources and lenses; we will come back to this issue in Sect. 4.3.3.

Lens equation. From Fig. 2.31 we can deduce the condition that a light ray from the source will reach us from the direction $\boldsymbol{\theta}$ (or $\boldsymbol{\xi}$),

$$\boldsymbol{\eta} = \frac{D_\mathrm{s}}{D_\mathrm{d}}\boldsymbol{\xi} - D_\mathrm{ds}\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) \,, \tag{2.77}$$

or, after dividing by D_s and using (2.75) and (2.76):

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \frac{D_\mathrm{ds}}{D_\mathrm{s}}\,\hat{\boldsymbol{\alpha}}(D_\mathrm{d}\boldsymbol{\theta}) \,. \tag{2.78}$$

Due to the factor multiplying the deflection angle in (2.78), it is convenient to define the *reduced deflection angle*

$$\boxed{\boldsymbol{\alpha}(\boldsymbol{\theta}) := \frac{D_\mathrm{ds}}{D_\mathrm{s}}\,\hat{\boldsymbol{\alpha}}(D_\mathrm{d}\boldsymbol{\theta})} \,, \tag{2.79}$$

so that the lens equation (2.78) attains the simple form

$$\boxed{\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})} \,. \tag{2.80}$$

Multiple images of a source occur if the lens equation (2.80) has multiple solutions $\boldsymbol{\theta}_i$ for a (true) source position $\boldsymbol{\beta}$—in this case, the source is observed at the positions $\boldsymbol{\theta}_i$ on the sphere.

The deflection angle $\boldsymbol{\alpha}(\boldsymbol{\theta})$ depends on the mass distribution of the deflector. We will discuss the deflection angle for an arbitrary density distribution of a lens in Sect. 3.11. Here we will first concentrate on point masses, which is—in most cases—a good approximation for the lensing effect by stars.

For a point mass, we get—see (2.74)—

$$|\boldsymbol{\alpha}|\,(\boldsymbol{\theta}) = \frac{D_\mathrm{ds}}{D_\mathrm{s}}\,\frac{4\,G\,M}{c^2\,D_\mathrm{d}\,|\boldsymbol{\theta}|} \,,$$

or, if we account for the direction of the deflection (the deflection angle always points towards the point mass),

$$\boxed{\boldsymbol{\alpha}(\boldsymbol{\theta}) = \frac{4\,G\,M}{c^2}\,\frac{D_\mathrm{ds}}{D_\mathrm{s}\,D_\mathrm{d}}\,\frac{\boldsymbol{\theta}}{|\boldsymbol{\theta}|^2}} \,. \tag{2.81}$$

Explicit solution of the lens equation for a point mass. The lens equation for a point mass is simple enough to be solved analytically which means that for each source position $\boldsymbol{\beta}$ the respective image positions $\boldsymbol{\theta}_i$ can be determined. In (2.81), the left-hand side is an angle, whereas $\boldsymbol{\theta}/|\boldsymbol{\theta}|^2$ is an inverse of an angle. Hence, the prefactor of this term must be the square of an angle, which is called the *Einstein angle* of the lens,

$$\boxed{\theta_\mathrm{E} := \sqrt{\frac{4\,G\,M}{c^2}\,\frac{D_\mathrm{ds}}{D_\mathrm{s}\,D_\mathrm{d}}}} \,; \tag{2.82}$$

thus the lens equation (2.80) for the point-mass lens with a deflection angle (2.81) can be written as

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \theta_\mathrm{E}^2\,\frac{\boldsymbol{\theta}}{|\boldsymbol{\theta}|^2} \,.$$

Obviously, θ_E is a characteristic angle in this equation, so that for practical reasons we will use the scaling

$$\boldsymbol{y} := \frac{\boldsymbol{\beta}}{\theta_\mathrm{E}} \quad ; \quad \boldsymbol{x} := \frac{\boldsymbol{\theta}}{\theta_\mathrm{E}} \,,$$

and the lens equation simplifies to

$$\boldsymbol{y} = \boldsymbol{x} - \frac{\boldsymbol{x}}{|\boldsymbol{x}|^2} \,. \tag{2.83}$$

After multiplication with $\boldsymbol{x}$, this becomes a quadratic equation, whose solutions are

$$\boldsymbol{x} = \frac{1}{2}\left(|\boldsymbol{y}| \pm \sqrt{4 + |\boldsymbol{y}|^2}\right) \frac{\boldsymbol{y}}{|\boldsymbol{y}|} \,. \tag{2.84}$$

From this solution of the lens equation one can immediately draw a number of conclusions:

- For each source position $\boldsymbol{y} \neq \boldsymbol{0}$, the lens equation for a point-mass lens has two solutions—any source is (formally, at least) imaged twice. The reason for this is the divergence of the deflection angle for $\theta \to 0$. This divergence does not occur in reality because of the finite geometric extent of the lens (e.g., the radius of the star), as the solutions are of course physically relevant only if $\xi = D_\mathrm{d}\theta_\mathrm{E}|\boldsymbol{x}|$ is larger than the radius of the star. We need to point out again that we explicitly exclude the case of strong gravitational fields such as the light deflection near a black hole or a neutron star, for which the equation for the deflection angle has to be modified, since there the gravitational field is no longer weak.
- The two images $\boldsymbol{x}_i$ are collinear with the lens and the source. In other words, the observer, lens, and source define a plane, and light rays from the source that reach the observer are located in this plane as well. One of the two images is located on the same side of the lens as the source ($\boldsymbol{x} \cdot \boldsymbol{y} > 0$), the second image is located on the other side—as is already indicated in Fig. 2.30.
- If $\boldsymbol{y} = 0$, so that the source is positioned exactly behind the lens, the full circle $|\boldsymbol{x}| = 1$, or $|\boldsymbol{\theta}| = \theta_\mathrm{E}$, is a solution of the lens equation (2.83)—the source is seen as a circular image. In this case, the source, lens, and observer no longer define a plane, and the problem becomes axially symmetric. Such a circular image is called an *Einstein ring*. Ring-shaped images have indeed been observed, as we will discuss in Sect. 3.11.3.
- The angular diameter of this ring is then $2\theta_\mathrm{E}$. From the solution (2.84), one can easily see that the separation between the two images is about $\Delta x = |\boldsymbol{x}_1 - \boldsymbol{x}_2| \gtrsim 2$ (as long as $|\boldsymbol{y}| \lesssim 1$), hence

$$\Delta\theta \gtrsim 2\theta_\mathrm{E} \,;$$

the Einstein angle thus specifies the characteristic image separation. Situations with $|\boldsymbol{y}| \gg 1$, and hence angular separations significantly larger than $2\theta_\mathrm{E}$, are astrophysically of only minor relevance, as will be shown below.

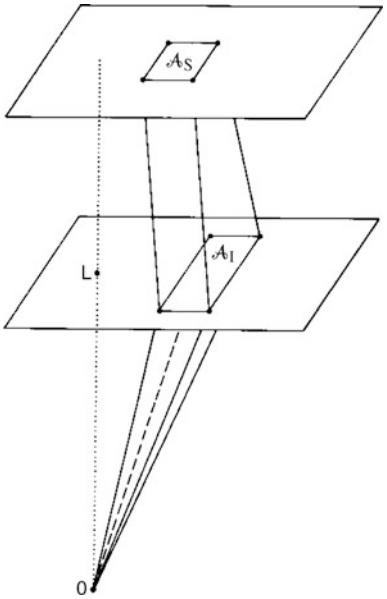


Fig. 2.32 Light beams are deflected differentially, leading to changes of the shape and the cross-sectional area of the beam. As a consequence, the observed solid angle subtended by the source, as seen by the observer, is modified by gravitational light deflection. In the example shown, the observed solid angle $\mathcal{A}_\mathrm{I}/D_\mathrm{d}^2$ is larger than the one subtended by the undeflected source, $\mathcal{A}_\mathrm{S}/D_\mathrm{s}^2$—the image of the source is thus magnified. Source: P. Schneider, J. Ehlers & E.E. Falco 1992, *Gravitational Lenses*, Springer-Verlag

Magnification—the principle. Light beams are not only deflected as a whole, but they are also subject to differential deflection. For instance, those rays of a light beam (also called light bundle) that are closer to the lens are deflected more than rays at the other side of the beam. The differential deflection is an effect of the tidal component of the deflection angle; this is sketched in Fig. 2.32. By differential deflection, the solid angle which the image of the source subtends on the sky changes. Let ω_s be the solid angle the source would subtend if no lens was present, and ω the observed solid angle of the image of the source in the presence of a deflector. Since gravitational light deflection is not linked to emission or absorption of radiation, the surface brightness (or specific intensity) is preserved. The flux of a source is given as the product of surface brightness and solid angle. Since the former of the two factors is unchanged by light deflection, but the solid angle changes, the observed flux of the source is modified. If S_0 is the flux of the unlensed source and S the flux of an image of the source, then

$$\boxed{\mu := \frac{S}{S_0} = \frac{\omega}{\omega_\mathrm{s}}} \tag{2.85}$$

describes the change in flux that is caused by a magnification (or a diminution) of the image of a source. Obviously, the magnification is a purely geometrical effect.

Magnification for 'small' sources. For sources and images that are much smaller than the characteristic scale of the lens, the magnification μ is given by the differential area distortion of the lens mapping (2.80),

$$\boxed{\mu = \left|\det\left(\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}}\right)\right|^{-1} \equiv \left|\det\left(\frac{\partial \beta_i}{\partial \theta_j}\right)\right|^{-1}} . \qquad (2.86)$$

Hence for small sources, the ratio of solid angles of the lensed image and the unlensed source is described by the determinant of the local Jacobi matrix.[14]

The magnification can therefore be calculated for each individual image of the source, and the total magnification of a source, given by the ratio of the sum of the fluxes of the individual images and the flux of the unlensed source, is the sum of the magnifications for the individual images.

Magnification for the point-mass lens. For a point-mass lens, the magnifications for the two images (2.84) are

$$\mu_\pm = \frac{1}{4}\left(\frac{y}{\sqrt{y^2+4}} + \frac{\sqrt{y^2+4}}{y} \pm 2\right) . \qquad (2.87)$$

From this it follows that for the '+'-image $\mu_+ > 1$ for all source positions $y = |\boldsymbol{y}|$, whereas the '−'-image can have magnification either larger or less than unity, depending on y. The magnification of the two images is illustrated in Fig. 2.33, while Fig. 2.34 shows the magnification for several different source positions y. For $y \gg 1$, one has $\mu_+ \gtrsim 1$ and $\mu_- \sim 0$, from which we draw the following conclusion: if the source and lens are not sufficiently well aligned, the secondary image is strongly demagnified and the primary image has magnification very close to unity. For this reason, situations with $y \gg 1$ are of little relevance since then essentially only one image is observed which has about the same flux as the unlensed source.

For $y \to 0$, the two magnifications diverge, $\mu_\pm \to \infty$. The reason for this is purely geometric: in this case, out of a 0-dimensional point source a one-dimensional image, the Einstein ring, is formed. This divergence is not physical, of course, since infinite magnifications do not occur in reality. The magnifications remain finite even for $y = 0$, for two reasons. First, real sources have a finite extent, and for these the magnification is finite. Second, even if one had a point source, wave effects of the light (interference) would lead to a finite value of μ. The total magnification of a point source by a point-mass lens follows from the sum of the magnifications (2.87),

$$\boxed{\mu(y) = \mu_+ + \mu_- = \frac{y^2+2}{y\sqrt{y^2+4}}} . \qquad (2.88)$$

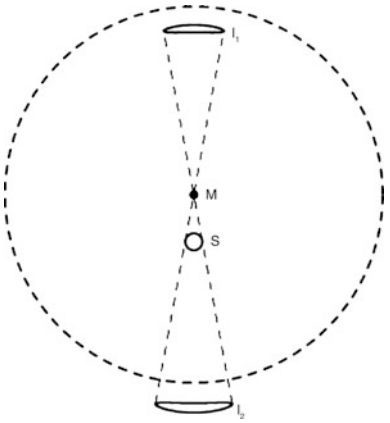


Fig. 2.33 Illustration of the lens mapping by a point mass M. The unlensed source S and the two images I_1 and I_2 of the lensed source are shown. We see that the two images have a solid angle different from the unlensed source, and they also have a different shape. The *dashed circle* shows the Einstein radius of the lens. Source: B. Paczyński 1996, *Gravitational Microlensing in the Local Group* ARA&A 34, 419, p. 424. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 34 ©1996 by Annual Reviews www.annualreviews.org

2.5.2 Galactic microlensing effect

After these theoretical considerations we will now return to the starting point of our discussion, employing the lensing effect as a potential diagnostic for dark matter in our Milky Way, if this dark matter were to consist of compact mass concentrations, e.g., very faint stars.

Image splitting. Considering a star in our Galaxy as the lens, (2.82) yields the Einstein angle

$$\boxed{\theta_\mathrm{E} = 0.902\,\mathrm{mas}\left(\frac{M}{M_\odot}\right)^{1/2}\left(\frac{D_\mathrm{d}}{10\,\mathrm{kpc}}\right)^{-1/2}\left(1-\frac{D_\mathrm{d}}{D_\mathrm{s}}\right)^{1/2}} . \qquad (2.89)$$

Since the angular separation $\Delta\theta$ of the two images is about $2\theta_\mathrm{E}$, the typical image splittings are about a milliarcsecond (mas) for lens systems including Galactic stars; such angular separations are as yet not observable with optical telescopes. This insight made Einstein believe in 1936, after he conducted a detailed quantitative analysis of gravitational

[14] The determinant in (2.86) is a generalization of the derivative in one spatial dimension to higher dimensional mappings. Consider a scalar mapping $y = y(x)$; through this mapping, a 'small' interval Δx is mapped onto a small interval Δy, where $\Delta y \approx (\mathrm{d}y/\mathrm{d}x)\,\Delta x$. The Jacobian determinant occurring in (2.86) generalizes this result to a two-dimensional mapping from the lens plane to the source plane.

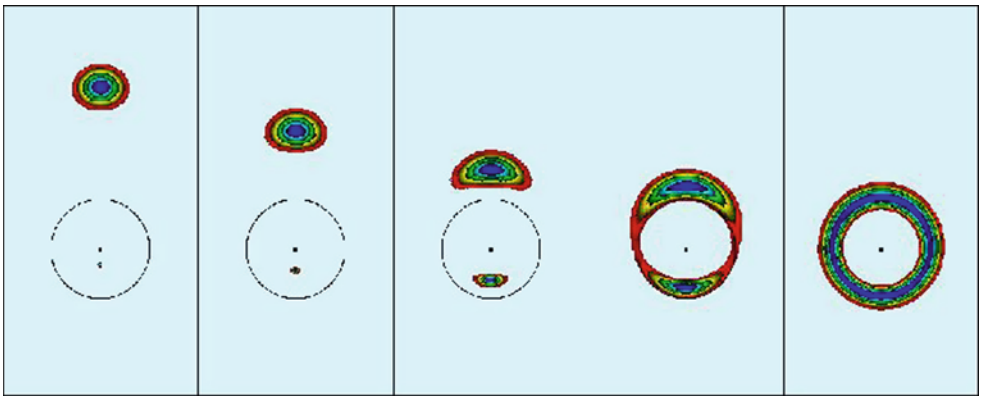


Fig. 2.34 Image of a circular source with a radial brightness profile—indicated by colors—for different relative positions of the lens and source. y decreases from *left to right*; in the rightmost figure $y = 0$ and an Einstein ring is formed. Source: J. Wambsganss 1998, *Gravitational Lensing in Astronomy*, Living Review in Relativity 1, 12, Fig. 20. ©Max Planck Society and the author; Living Reviews in Relativity, published by the Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Germany

lensing by point masses, that the lens effect will not be observable.[15]

Magnification. Bohdan Paczyński pointed out in 1986 that, although image splitting was unobservable, the magnification by the lens should nevertheless be measurable. To do this, we have to realize that the absolute magnification is observable only if the unlensed flux of the source is known—which is not the case, of course (for nearly all sources). However, the magnification, and therefore also the observed flux, changes with time by the relative motion of source, lens, and ourselves. Therefore, the flux is a function of time, caused by the time-dependent magnification.

Characteristic time-scale of the variation. Let v be a typical transverse velocity of the lens, then its angular velocity (or proper motion) is

$$\dot{\theta} = \frac{v}{D_\mathrm{d}} = 4.22\,\mathrm{mas\,yr^{-1}} \left(\frac{v}{200\,\mathrm{km/s}}\right)\left(\frac{D_\mathrm{d}}{10\,\mathrm{kpc}}\right)^{-1}, \tag{2.90}$$

if we consider the source and the observer to be at rest. The characteristic time-scale of the variability is then given by

$$\boxed{t_\mathrm{E} := \frac{\theta_\mathrm{E}}{\dot{\theta}} = 0.214\,\mathrm{yr}\left(\frac{M}{M_\odot}\right)^{1/2}\left(\frac{D_\mathrm{d}}{10\,\mathrm{kpc}}\right)^{1/2} \times \left(1-\frac{D_\mathrm{d}}{D_\mathrm{s}}\right)^{1/2}\left(\frac{v}{200\,\mathrm{km/s}}\right)^{-1}.} \tag{2.91}$$

[15]The expression 'microlens' has its origin in the angular scale (2.89) that was discussed in the context of the lens effect on quasars by stars at cosmological distances, for which one obtains image splittings of about one microarcsecond; see Sect. 5.4.1.

This time-scale is of the order of a month for lenses with $M \sim M_\odot$ and typical Galactic velocities. In the general case that source, lens, and observer are all moving, v has to be considered as an effective velocity. Alternatively, the motion of the source in the source plane can be considered.

The fact that t_E comes out to be a month for characteristic values of distances and velocities in our Galaxy is a fortunate coincidence, since it implies that these variations are in fact observable. If the time-scale was a factor ten times larger, the characteristic light curve would extend over several observing periods and include the annual gaps where the sources are not visible, making the detection of events much more difficult. If t_E was of order several years, the variability time-scale would be longer than the life-time of most projects in astrophysics. Conversely, it t_E was considerable shorter than a day, the variations would be difficult to record.

Light curves. In most cases, the relative motion can be considered linear, so that the position of the source in the source plane can be written as

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \dot{\boldsymbol{\beta}}(t - t_0)\,.$$

Using the scaled position $\boldsymbol{y} = \boldsymbol{\beta}/\theta_\mathrm{E}$, for $y = |\boldsymbol{y}|$ we obtain

$$\boxed{y(t) = \sqrt{p^2 + \left(\frac{t - t_\mathrm{max}}{t_\mathrm{E}}\right)^2}}, \tag{2.92}$$

where $p = y_\mathrm{min}$ specifies the minimum distance from the optical axis, and t_max is the time at which $y = p$ attains this minimum value, thus when the magnification $\mu = \mu(p) = \mu_\mathrm{max}$ is maximized. From this, and using (2.88), one obtains the light curve

$$S(t) = S_0\,\mu(y(t)) = S_0\,\frac{y^2(t)+2}{y(t)\,\sqrt{y^2(t)+4}}\,. \qquad (2.93)$$

Examples for such light curves are shown in Fig. 2.35. They depend on only four parameters: the flux of the unlensed source S_0, the time of maximum magnification $t_{\max}$, the smallest distance of the source from the optical axis p, and the characteristic time scale $t_{\rm E}$. All these values are directly observable in a light curve. One obtains $t_{\max}$ from the time of the maximum of the light curve, S_0 is the flux that is measured for very large and small times, $S_0 = S(t \to \pm\infty)$, or $S_0 \approx S(t)$ for $|t - t_{\max}| \gg t_{\rm E}$. Furthermore, p follows from the maximum magnification $\mu_{\max} = S_{\max}/S_0$ by inversion of (2.88), and $t_{\rm E}$ from the width of the light curve.

Only $t_{\rm E}$ contains information of astrophysical relevance, because the time of the maximum, the unlensed flux of the source, and the minimum separation p provide no information about the lens. Since $t_{\rm E} \propto \sqrt{M\,D_{\rm d}}/v$, this time scale contains the combined information on the lens mass, the distances to the lens and the source, and the transverse velocity: *Only the combination* $t_{\rm E} \propto \sqrt{M\,D_{\rm d}}/v$ *can be derived from the light curve, but not mass, distance, or velocity individually*.

Paczyński's idea can be expressed as follows: if the halo of our Milky Way consists (partially) of compact objects, a distant compact source should, from time to time, be lensed by one of these MACHOs and thus show characteristic changes in flux, corresponding to a light curve similar to those in Fig. 2.35. The number density of MACHOs is proportional to the probability or abundance of lensing events, and the characteristic mass of the MACHOs is proportional to the square of the typical variation time scale $t_{\rm E}$. All one has to do is measure the light curves of a sufficiently large number of background sources and extract all lens events from those light curves to obtain information on the population of potential MACHOs in the halo. A given halo model predicts the spatial density distribution and the distribution of velocities of the MACHOs and can therefore be compared to the observations in a statistical way. However, one faces the problem that the abundance of such lensing events is very small.

Probability of a lensing event. In practice, a system of a foreground lens and a background source is considered a lensing event if $p \le 1$, or $\beta_{\min} \le \theta_{\rm E}$, and hence $\mu_{\max} \ge 3/\sqrt{5} \approx 1.34$, i.e., if the relative trajectory of the source passes within the Einstein circle of the lens.

If the dark halo of the Milky Way consisted solely of MACHOs, the probability that a very distant source is lensed (in the sense of $|\boldsymbol{\beta}| \le \theta_{\rm E}$) would be $\sim 10^{-7}$, where the exact value depends on the direction to the source. At any one time,

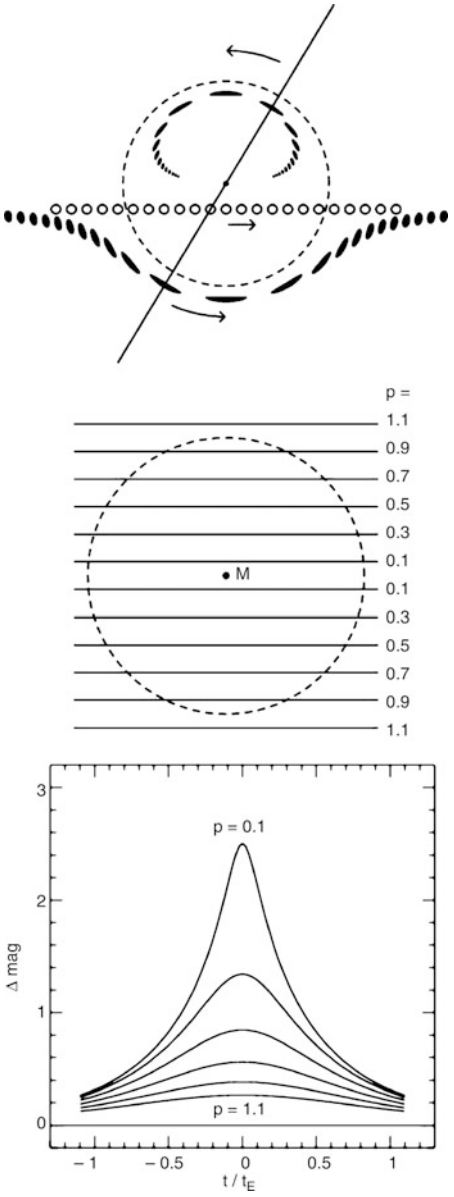


Fig. 2.35 Illustration of a Galactic microlensing event: In the *upper panel* a source (depicted by the open circles) moves behind a point-mass lens; for each source position two images of the source are formed, which are indicated by the black ellipses. Note that Fig. 2.33 shows the imaging properties for one of the source positions shown here. The identification of the corresponding image pair with the source position follows from the fact that, in projection, the source, the lens, and the two images are located on a *straight line*, which is indicated for one source position; this property follows from the collinearity of source and images mentioned in the text. The *dashed circle* represents the Einstein ring. In the *middle panel*, different trajectories of the source are shown, each characterized by the smallest projected separation p to the lens. The light curves resulting from these relative motions, which can be calculated using (2.93), are then shown in the *bottom panel* for different values of p. The smaller p is, the larger the maximum magnification will be, here measured in magnitudes. Source: B. Paczyński 1996, *Gravitational Microlensing in the Local Group* ARA&A 34, 419, p. 425, 426, 427. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 34 ©1996 by Annual Reviews www.annualreviews.org

one of $\sim 10^7$ distant sources would be located within the Einstein radius of a MACHO in our halo. The immediate consequence of this is that the light curves of millions of sources have to be monitored to detect this effect. Furthermore, these sources have to be located within a relatively small region on the sphere to keep the total solid angle that has to be photometrically monitored relatively small. This condition is needed to limit the required observing time, so that many such sources should be present within the field-of-view of the camera used. The stars of the Magellanic Clouds are well suited for such an experiment: they are close together on the sphere, but can still be resolved into individual stars.

Problems, and their solution. From this observational strategy, a large number of problems arise immediately; they were discussed in Paczyński's original paper. First, the photometry of so many sources over many epochs produces a huge amount of data that need to be handled; they have to be stored and reduced. Second, one has the problem of 'crowding': the stars in the Magellanic Clouds are densely packed on the sky, which renders the photometry of individual stars difficult. Third, stars also show intrinsic variability—about 1 % of all stars are variable. This intrinsic variability has to be distinguished from that due to the lens effect. Due to the small probability of the latter, selecting the lensing events is comparable to searching for a needle in a haystack. Finally, it should be mentioned that one has to ensure that the experiment is indeed sensitive enough to detect lensing events. A 'calibration experiment' would therefore be desirable.

Faced with these problems, it seemed daring to seriously think about the realization of such an observing program. However, a fortunate event helped, in the magnificent time of the easing of tension between the US and the Soviet Union, and their respective allies, at the end of the 1980s. Physicists and astrophysicists, that had been partly occupied with issues concerning national security, then saw an opportunity to meet new challenges. In addition, scientists in national laboratories had much better access to sufficient computing power and storage capacity than those in other research institutes, attenuating some of the aforementioned problems. While the expected data volume was still a major problem in 1986, it could be handled a few years later. Also, wide-field cameras were constructed, with which large areas of the sky could be observed simultaneously. Software was developed specialized to the photometry of objects in crowded fields, so that light curves could be measured even if individual stars in the image were no longer cleanly separated.

To distinguish between lensing events and intrinsic variability of stars, we note that the microlensing light curves have a characteristic shape that is described by only four parameters. The light curves should be symmetric and achromatic because gravitational light deflection is independent of the frequency of the radiation. Furthermore, due to the small lensing probability, any source should experience at most one microlensing event and show a constant flux before and after, whereas intrinsic variations of stars are often periodic and in nearly all cases chromatic.

And finally a control experiment could be performed: the lensing probability in the direction of the Galactic bulge is known, or at least, we can obtain a lower limit for it from the observed density of stars in the disk. If a microlens experiment is carried out in the direction of the Galactic bulge, we *have to* find lensing events if the experiment is sufficiently sensitive.

2.5.3 Surveys and results

In the early 1990s, two collaborations (MACHO and EROS) began the search for microlensing events towards the Magellanic Clouds. Another group (OGLE) started searching in an area of the Galactic bulge. Fields in the respective survey regions were observed regularly, typically once every night if weather conditions permitted. From the photometry of the stars in the fields, light curves for many millions of stars were generated and then checked for microlensing events.

First detections. In 1993, all three groups reported their first results. The MACHO collaboration found one event in the Large Magellanic Cloud (LMC), the EROS group two events, and the OGLE group observed one event in the Galactic bulge. The light curve of the first MACHO event is plotted in Fig. 2.36. It was observed in two different filters, and the fit to the data, which corresponds to a standard light curve (2.93), is the same for both filters, proving that the event is achromatic. Together with the quality of the fit to the data, this is very strong evidence for the microlensing nature of the event.

Statistical results. After 1993, all three aforementioned teams proceeded with their observations and analysis (Fig. 2.37), and more groups have begun the search for microlensing events, choosing various lines of sight. The most important results from these experiments can be summarized as follows:

About 20 events have been found in the direction of the Magellanic Clouds, and some ten thousand in the direction of the bulge. The statistical analysis of the data revealed the lensing probability towards the bulge to be higher than originally expected. This can be explained by the fact that *our Galaxy features a bar* (see Chap. 3). This bar was also observed in IR maps such as those made by the COBE satellite. The events in the direction of the bulge are dominated by lenses that are part of the bulge themselves, and their column

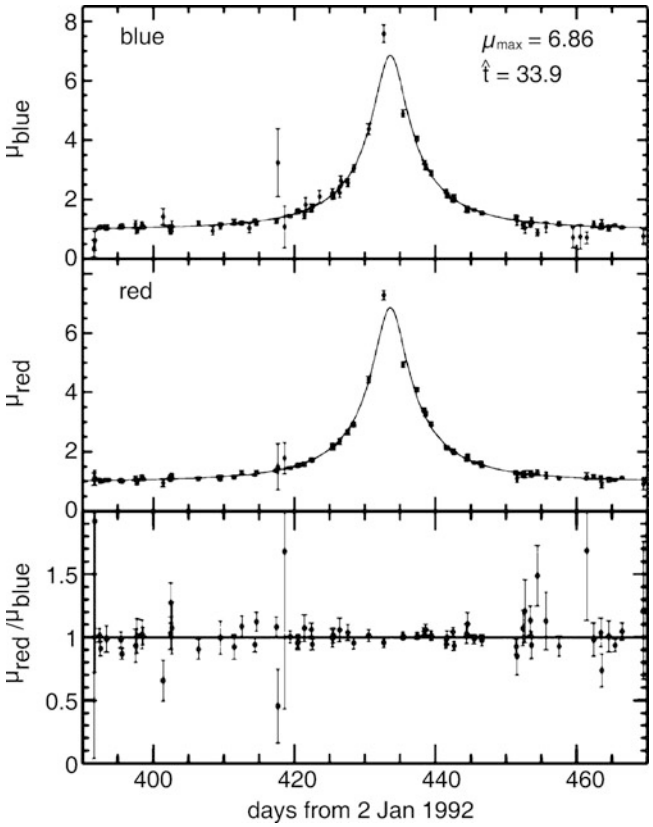


Fig. 2.36 Light curve of the first observed microlensing event in the LMC, in two broad-band filters. The *solid curve* is the best-fitting microlensing light curve as described by (2.93), with $\mu_{max} = 6.86$. The ratio of the magnifications in both filters is displayed at the bottom, and it is compatible with 1. Some of the data points deviate significantly from the curve; this means that either the errors in the measurements were underestimated, or this event is more complicated than one described by a point-mass lens—see Sect. 2.5.4. Source: C. Alcock et al. 1993, *Possible gravitational microlensing of a star in the Large Magellanic Cloud*, Nature 365, 621

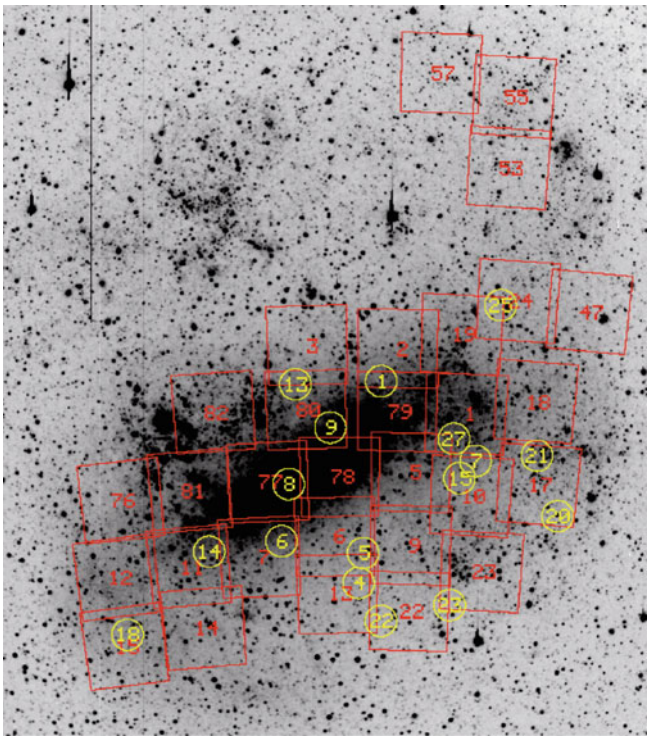


Fig. 2.37 In this $8° \times 8°$ image of the LMC, 30 fields are marked in *red* which the MACHO group has searched for microlensing events during the $\sim$ 5.5 yr of their experiment; images were taken in two filters to test for achromaticity. The positions of 17 microlensing events are marked by *yellow circles*; these have been subject to statistical analysis. Source: C. Alcock et al. 2000, *The MACHO Project: Microlensing Results from 5.7 Years of Large Magellanic Cloud Observations*, ApJ 542, 281, p. 284, Fig. 1. ©AAS. Reproduced with permission

density is increased by the bar-like shape of the bulge. On the other hand, the lensing probability in the direction of the Magellanic Clouds is much *smaller* than expected for the case where the dark halo consists solely of MACHOs. Based on the analysis of the MACHO collaboration, the observed statistics of lensing events towards the Magellanic Clouds is best explained if about 20 % of the halo mass consists of MACHOs, with a characteristic mass of about $M \sim 0.5 M_\odot$ (see Fig. 2.38).

Interpretation and discussion. This latter result is not easy to interpret and came as a real surprise. If a result compatible with $\sim$ 100 % had been found, it would have been obvious to conclude that the dark matter in our Milky Way consists of compact objects. Otherwise, if very few lensing events had been found, it would have been clear that MACHOs do not contribute significantly to the dark matter. But a value of 20 % does not immediately allow any unambiguous interpretation. Taken at face value, the result from the MACHO group would imply that the total mass of MACHOs in the Milky Way halo is about the same as that in the stellar disk.

Furthermore, the estimated mass scale is hard to understand: what could be the nature of MACHOs with $M = 0.5 M_\odot$? Normal stars can be excluded, because they would be far too luminous to escape direct observations. White dwarfs are also unsuitable candidates, because to produce such a large number of white dwarfs as a final stage of stellar evolution, the total star formation in our Milky Way, integrated over its lifetime, needs to be significantly larger than normally assumed. In this case, many more massive stars would also have formed, which would then have released the metals they produced into the ISM, both by stellar winds and in supernova explosions. In such a scenario, the metal content of the ISM would therefore be distinctly higher than is actually observed. The only possibility of escaping this argument is with the hypothesis that the mass function of newly formed stars (the initial mass function, IMF) was different in the early phase of the Milky Way compared to that observed today. The IMF that needs to be assumed in

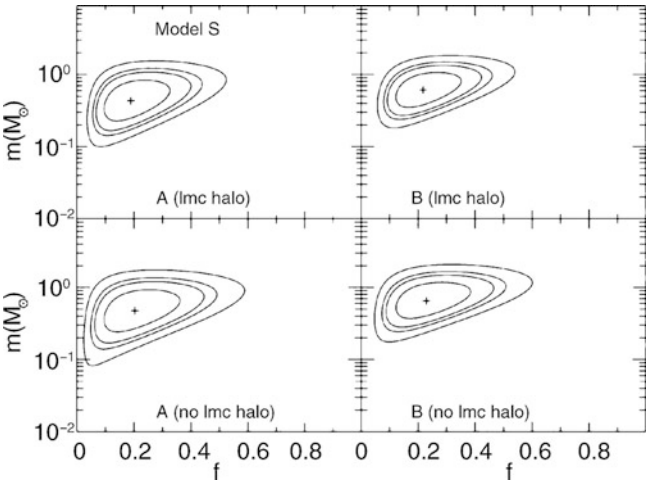


Fig. 2.38 Probability contours for a specific halo model as a function of the characteristic MACHO mass M (here denoted by m) and the mass fraction f of MACHOs in the halo. The halo of the LMC was either taken into account as an additional source for microlenses (lmc halo) or not (no lmc halo), and two different selection criteria (A, B) for the statistically complete microlensing sample were employed. In all cases, $M \sim 0.5 M_\odot$ and $f \sim 0.2$ are the best-fit values. Source: C. Alcock et al. 2000, *The MACHO Project: Microlensing Results from 5.7 Years of Large Magellanic Cloud Observations*, ApJ 542, 281, p. 304, Fig. 12. ©AAS. Reproduced with permission

this case is such that for each star of intermediate mass which evolves into a white dwarf, far fewer high-mass stars, mainly responsible for the metal enrichment of the ISM, must have formed in the past compared to today. However, we lack a plausible physical model for such a scenario, and it is in conflict with the star-formation history that we observe in the high-redshift Universe (see Chap. 9).

Neutron stars can be excluded as well, because they are too massive (typically $> 1 M_\odot$); in addition, they are formed in supernova explosions, implying that the aforementioned metallicity problem would be even greater for neutron stars. Would stellar-mass black holes be an alternative? The answer to this question depends on how they are formed. They could not originate in SN explosions, again because of the metallicity problem. If they had formed in a very early phase of the Universe (they are then called primordial black holes), this would be an imaginable, though perhaps quite exotic, alternative.

However, we have strong indications that the interpretation of the MACHO results is not as straightforward as described above. Some doubts have been raised as to whether all events reported as being due to microlensing are in fact caused by this effect. In fact, one of the microlensing source stars identified by the MACHO group showed another bump 7 years after the first event. Given the extremely small likelihood of two microlensing events happening to a single source this is almost certainly a star with unusual variability. There are good arguments to attribute two events to stars in the thick disk.

As argued previously, by means of t_E we only measure a combination of lens mass, transverse velocity, and distance. The result given in Fig. 2.38 is therefore based on the statistical analysis of the lensing events in the framework of a halo model that describes the shape and the radial density profile of the halo. However, microlensing events have been observed for which more than just t_E can be determined—e.g., events in which the lens is a binary star, or those for which t_E is larger than a few months. In this case, the orbit of the Earth around the Sun, which is not a linear motion, has a noticeable effect, causing deviations from the standard light curve. Such parallax events have indeed been observed.[16] Three events are known in the direction of the Magellanic Clouds in which more than just t_E could be measured. In all three cases the lenses are most likely located in the Magellanic Clouds themselves (an effect called self-lensing) and not in the halo of the Milky Way. If for those three cases, where the degeneracy between lens mass, distance, and transverse velocity can be broken, the respective lenses are not MACHOs in the Galactic halo, we might then suspect that in most of the other microlensing events the lens is not a MACHO either. Therefore, it is currently unclear how to interpret the results of the MACHO survey. In particular, it is unclear to what extent self-lensing contributes to the results. Furthermore, the quantitative results depend on the halo model.

The EROS collaboration used an observation strategy which was sightly different from that of the MACHO group, by observing a number of fields in very short time intervals. Since the duration of a lensing event depends on the mass of the lens as $\Delta t \propto M^{1/2}$—see (2.91)—they were also able to probe very small MACHO masses. The absence of lensing events of very short duration then allowed them to derive limits for the mass fraction of such low-mass MACHOs, as is shown in Fig. 2.39. In particular, neither the EROS nor the OGLE group could reproduce the relatively large event rate found by the MACHO group; indeed, the EROS and OGLE results do not require any unknown component of compact objects in our Milky Way, and OGLE derived an upper bound of $\sim 2\,\%$ of the dark matter in our Milky Way which could be in the form of compact objects.

We have to emphasize that the microlensing surveys have been enormously successful experiments because they accomplished exactly what was expected at the beginning of the observations. They measured the lensing probability in the direction of the Magellanic Clouds and the Galactic bulge, excluded the possibility that a major fraction of the dark matter is in compact objects, and revealed the structure of the Galactic bulge.

[16] These parallax events in addition prove that the Earth is in fact orbiting around the Sun—even though this is not really a new insight. . ..

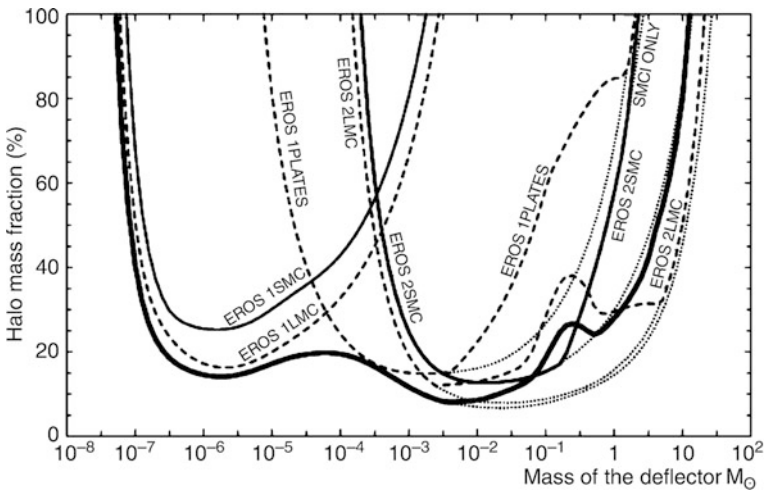


Fig. 2.39 From observations by the EROS collaboration, a large mass range for MACHO candidates can be excluded. The maximum allowed fraction of the halo mass contained in MACHOs is plotted as a function of the MACHO mass M, as an upper limit with 95 % confidence. A standard model for the mass distribution in the Galactic halo was assumed which describes the rotation curve of the Milky Way quite well. The *various curves* show different phases of the EROS experiment. They are plotted separately for observations in the directions of the LMC and the SMC. The experiment EROS 1 searched for microlensing events on short time-scales but did not find any; this yields the upper limits at small masses. Upper limits at larger masses were obtained by the EROS 2 experiment. The *thick solid curve* represents the upper limit derived from combining the individual experiments. If not a single MACHO event had been found the upper limit would have been described by the dotted line. Source: C. Afonso et al. 2003, *Limits on Galactic dark matter with 5 years of EROS SMC data*, A&A 400, 951, p. 955, Fig. 3. ©ESO. Reproduced with permission

The microlensing surveys did not constrain the density of compact objects with masses $\gtrsim 10 M_\odot$, since the variability time-scale from such high-mass lenses becomes comparable to the survey duration. Whereas such high-mass MACHOs are physically even less plausible than $\sim 0.5\, M_\odot$ candidates, it still would be good to be able to rule them out. This can be done by studying wide binary systems in the stellar halo. If the dark matter in our Galaxy would be present in form of high-mass MACHOs, these would affect the binary population, in particular by disrupting wide binaries. From considering the separation distribution of halo binaries, it can be excluded that high-mass compact objects constitute the dark matter in the Galactic halo.

2.5.4 Variations and extensions

Besides the search for MACHOs, microlensing surveys have yielded other important results and will continue to do so in the future. For instance, the distribution of stars in the Galaxy can be measured by analyzing the lensing probability as a function of direction. A huge number of variable stars have been newly discovered and accurately monitored; the extensive and publicly accessible databases of the surveys form an invaluable resource for stellar astrophysics. Proper motions of several million stars have been determined, based on $\sim$ 20 yr of microlensing surveys. Furthermore, globular clusters in the LMC have been identified from these photometric observations.

For some lensing events, the radius and the surface structure of distant stars can be measured with very high precision. This is possible because the magnification μ depends on the position of the source. Situations can occur, for example where a binary star acts as a lens (see Fig. 2.40), in which the dependence of the magnification on the position in the source plane is very sensitive. Since the source—the star—is in motion relative to the line-of-sight between Earth and the lens, its different regions are subject to different magnification, depending on the time-dependent source position. A detailed analysis of the light curve of such events then enables us to reconstruct the light distribution on the surface of the star. The light curve of one such event is shown in Fig. 2.41. For these lensing events the source can no longer be assumed to be a point source. Rather, the details of the light curve are determined by its light distribution. Therefore, another length-scale appears in the system, the radius of the star. This length-scale shows up in the corresponding microlensing light curve, as can be seen in Fig. 2.41, by the time-scale which characterizes the width of the peaks in the light curve—it is directly related to the ratio of the stellar radius and the transverse velocity of the lens. With this new scale, the degeneracy between M, v, and D_d is partially broken, so that these special events provide more information than the 'classical' ones.

In fact, the light curve in Fig. 2.36 is probably not caused by a single lens star, but instead by additional slight disturbances from a companion star. This would explain the deviation of the observed light curve from a simple model

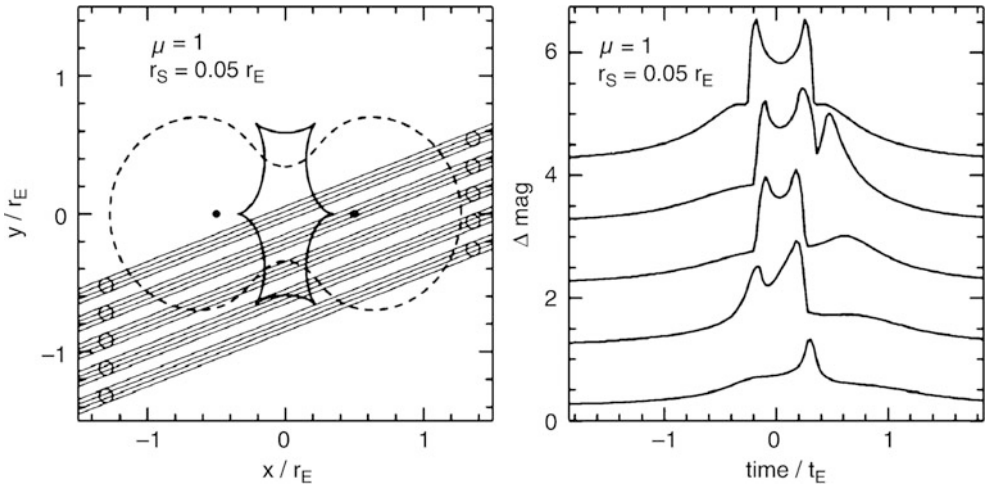


Fig. 2.40 If a binary star acts as a lens, significantly more complicated light curves can be generated. In the *left-hand panel* tracks are plotted for five different relative motions of a background source; the *dashed curve* is the so-called *critical curve*, formally defined by $\det(\partial\boldsymbol{\beta}/\partial\boldsymbol{\theta}) = 0$, and the *solid line* is the corresponding image of the critical curve in the source plane, called a *caustic*. Light curves corresponding to these five tracks are plotted in the *right-hand panel*. If the source crosses the caustic, the magnification μ becomes very large—formally infinite if the source was point-like. Since it has a finite extent, μ has to be finite as well; from the maximum μ during caustic crossing, the radius of the source can be determined, and sometimes even the variation of the surface brightness across the stellar disk, an effect known as limb darkening. Source: B. Paczyński 1996, *Gravitational Microlensing in the Local Group* ARA&A 34, 419, p. 435, 434. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 34 ©1996 by Annual Reviews www.annualreviews.org

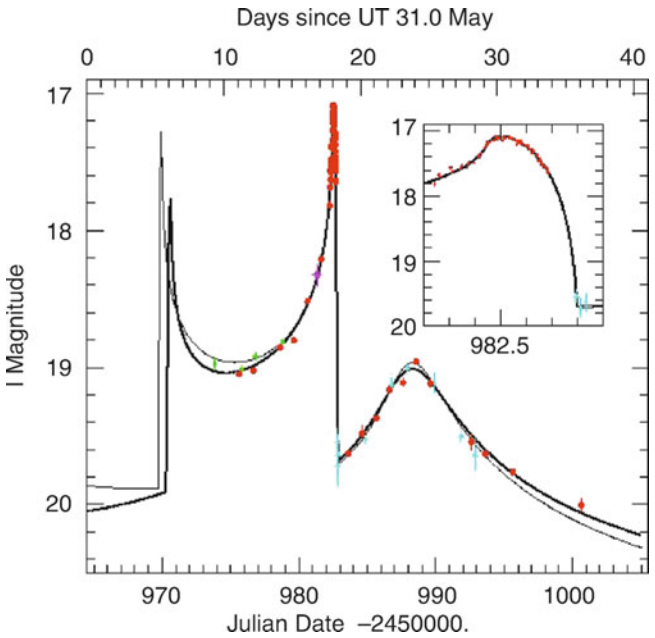


Fig. 2.41 Light curve of an event in which the lens was a binary star. Note the qualitative similarity of this light curve with the second one from the top in Fig. 2.40. The MACHO group discovered this 'binary event'. Members of the PLANET collaboration obtained this data using four different telescopes (in Chile, Tasmania, and South Africa). The second caustic crossing is highly resolved (displayed in the *small diagram*) and allows us to draw conclusions about the size and the brightness distribution of the source star. The two curves show the fits of a binary lens to the data. Source: M.D. Albrow et al. 1999, *The Relative Lens-Source Proper Motion in MACHO 98-SMC-1*, ApJ 512, 672, p. 674, Fig. 2. ©AAS. Reproduced with permission

light curve. However, the sampling in time of this particular light curve is not sufficient to determine the parameters of the binary system.

By now, detailed light curves with very good time coverage have been measured, which was made possible with an alarm system. The data from those groups searching for microlensing events are analyzed immediately after observations, and potential candidates for interesting events are published on the Internet. Other groups (such as the PLANET collaboration, for example) then follow-up these systems with very good time coverage by using several telescopes spread over a large range in geographical longitude. This makes around-the-clock observations of the events possible. Using this method, light curves of extremely high quality have been measured, and events in which the lens is a binary with a very large mass ratio have been detected—so large that the lighter of the two masses is not a star, but a planet. Indeed, more than a dozen extrasolar planets have been found by microlensing surveys. Whereas this number at first sight is not so impressive, given that many more extrasolar planets were discovered by other methods, the selection function in microlensing surveys is quite different. In contrast to the radial velocity method (where the periodic change of the radial velocity of the parent star, caused by its motion around the center of mass of the star-planet system, is measured), microlensing has detected lower-mass planets and planets at larger separation from the host star.

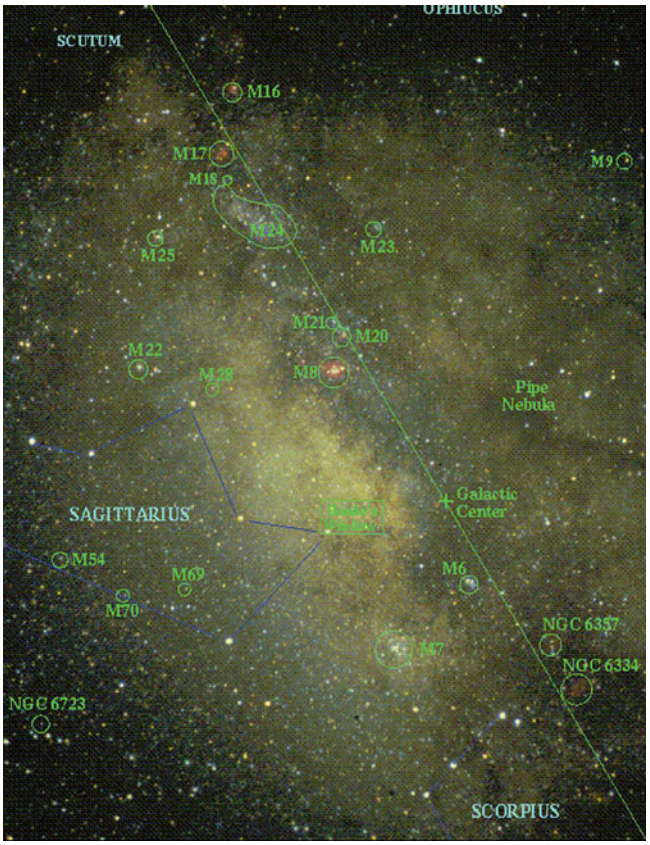


Fig. 2.42 Optical image in the direction of the Galactic center. The size of the image is $\sim 10° \times 15°$. Marked are some Messier objects: gas nebulae such as M8, M16, M17, M20; open star clusters such as M6, M7, M18, M21, M23, M24, and M25; globular clusters such as M9, M22, M28, M54, M69, and M70. Also marked is the Galactic center, as well as the Galactic plane, which is indicated by a *line*. Baade's Window can be easily recognized, a direction in which the extinction is significantly lower than in nearby directions, so that a clear increase in stellar density is visible there. This is the reason why the microlensing observations towards the Galactic center were preferably done in Baade's Window. Credit: W. Keel (U. Alabama, Tuscaloosa), Cerro Tololo, Chile

2.6 The Galactic center

The Galactic center (GC, see Fig. 2.42) is not observable at optical wavelengths, because the extinction in the V band is ~ 28 mag. Our information about the GC has been obtained from radio-, IR-, and X-ray radiation, although even in the K-band, the extinction is still ~ 3 mag. Since the GC is nearby, and thus serves as a prototype of the central regions of galaxies, its observation is of great interest for our understanding of the processes taking place in the centers of galaxies.

2.6.1 Where is the Galactic center?

The question of where the center of our Milky Way is located is by no means trivial, because the term 'center' is in fact not well-defined. Is it the center of mass of the Galaxy, or the point around which the stars and the gas are orbiting? And how could we pinpoint this 'center' accurately? Fortunately, the center can nevertheless be localized because, as we will see below, a distinct source exists that is readily identified as the Galactic center.

Radio observations in the direction of the GC show a relatively complex structure, as is displayed in Fig. 2.43. A central disk of HI gas exists at radii from several 100 pc up to about 1 kpc. Its rotational velocity yields an estimate of the enclosed mass $M(R)$ for $R \gtrsim 100$ pc. Furthermore, radio filaments are observed which extend perpendicularly to the Galactic plane, and a large number of supernova remnants are seen. Within about 2 kpc from the center, roughly $3 \times 10^7 M_\odot$ of atomic hydrogen is found. Optical images show regions close to the GC towards which the extinction is significantly lower. The best known of these is Baade's Window—most of the microlensing surveys towards the bulge are conducted in this region. It is the brightest region in Fig. 2.42, not because the stellar density is highest there, but the obscuration is smallest. In addition, a fairly large number of globular clusters and gas nebulae are observed towards the central region. X-ray images (Fig. 2.44) show numerous X-ray binaries, as well as diffuse emission by hot gas.

The innermost 8 pc contain the radio source Sgr A (Sagittarius A), which itself consists of different components:

- A circumnuclear molecular ring, shaped like a torus, which extends between radii of $2\,\text{pc} \lesssim R \lesssim 8\,\text{pc}$ and is inclined by about 20° relative to the Galactic disk. The rotational velocity of this ring is about ~ 110 km/s, nearly independent of R. This ring has a sharp inner boundary; this cannot be the result of an equilibrium flow, because internal turbulent motions would quickly (on a time scale of $\sim 10^5$ yr) blur this boundary. Probably, it is evidence of an energetic event that occurred in the Galactic center within the past $\sim 10^5$ yr. This interpretation is also supported by other observations, e.g., by a clumpiness in density and temperature.
- Sgr A East, a non-thermal (synchrotron) source of shell-like structure. Presumably this is a supernova remnant (SNR), with an age between 100 and 5000 years.
- Sgr A West is located about 1′.5 away from Sgr A East. It is a thermal source, an unusual HII region with a spiral-like structure.
- Sgr A* is a compact radio source close to the center of Sgr A West. Recent observations with mm-VLBI show that its extent is smaller than about 1 AU. The radio luminosity is $L_{\text{rad}} \sim 2 \times 10^{34}$ erg/s. Except for the emission in the mm and cm domain, Sgr A* is a weak source. Since other galaxies often have a compact radio source in their center, Sgr A* is an excellent candidate for being the center of our Milky Way.

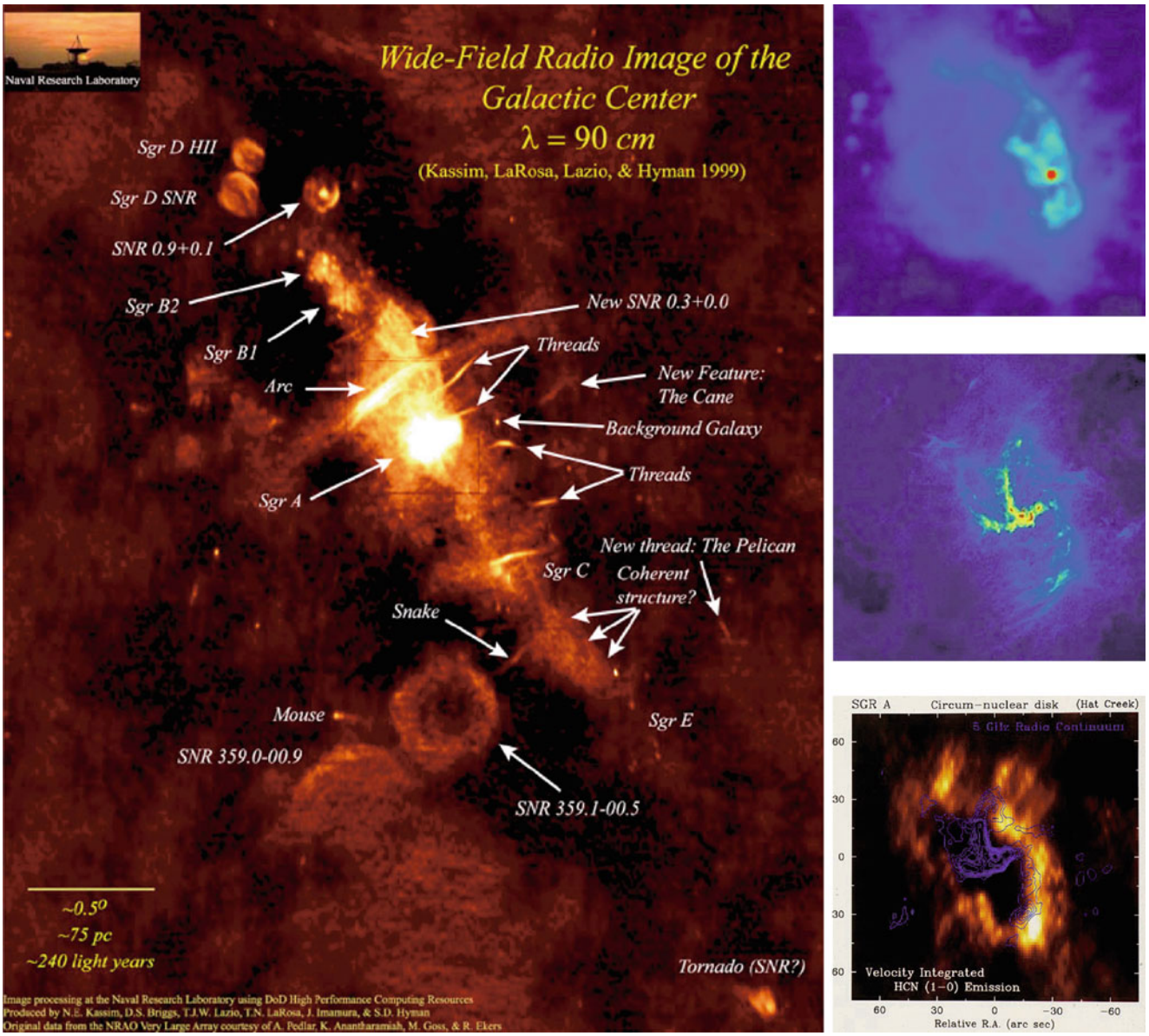


Fig. 2.43 *Left*: A VLA wide-field image of the region around the Galactic center, with a large number of sources identified. *Upper right*: a 20 cm continuum VLA image of Sgr A East. *Center right*: Sgr A West, as seen in a 6-cm continuum VLA image, where the red dot marks Sgr A*. *Lower right*: the circumnuclear ring in HCN line emission. Source: *Left*: N.E. Kassim, from T.N. LaRosa et al. 2000, *A Wide-Field 90 Centimeter VLA Image of the Galactic Center Region*, AJ 119, 207, P. 209, Fig. 1. ©AAS. Reproduced with permission. Credit: Produced by the U.S. Naval Research Laboratory by Dr. N.E. Kassim and collaborators from data obtained with the National Radio Astronomy's Very Large Array Telescope, a facility of the National Science Foundation operated under the cooperative agreement with associated Universities, Inc. Basic research in radio astronomy is supported by the U.S. Office of Naval Research. Upper right: from R.L. Plante et al. 1995, *The magnetic fields in the galactic center: Detection of H1 Zeeman splitting*, ApJ 445, L113, Fig. 1. ©AAS. Reproduced with permission. Center right: Image courtesy of NRAO/AUI, National Radio Astronomy Observatory. Lower right: Image courtesy of Leo Blitz and Hat Creek Observatory

Through observations of stars which contain a radio maser[17] source, the astrometry of the GC in the radio domain was matched to that in the IR, i.e., the position of Sgr A* is also known in the IR.[18] The uncertainty in the relative

[17]Masers are regions of stimulated non-thermal emission which show a very high surface brightness. The maser phenomenon is similar to that of lasers, except that the former radiate in the microwave regime of the spectrum. Masers are sometimes found in the atmospheres of active stars.

[18]One problem in the combined analysis of data taken in different wavelength bands is that astrometry in each individual wavelength band can be performed with a very high precision—e.g., individually in the radio and the IR band—however, the relative astrometry between these bands is less well known. To stack maps of different wavelengths

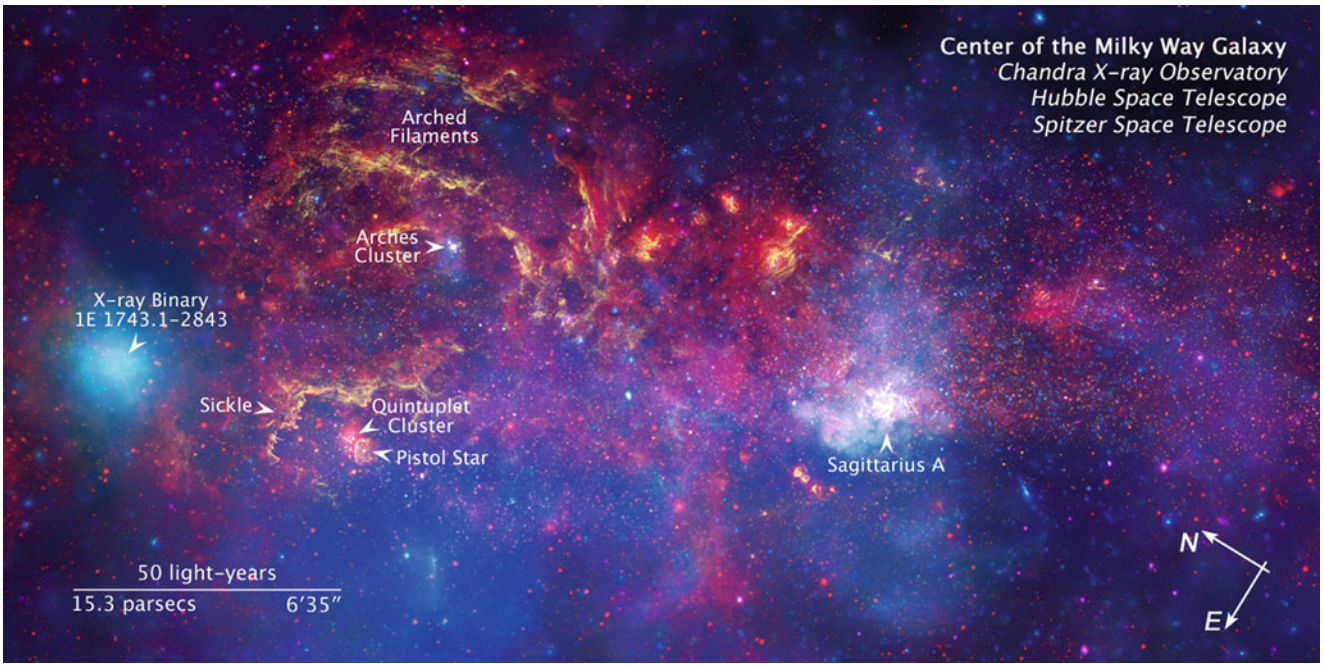


Fig. 2.44 A composite image of the Galactic center: X-ray emission as observed by Chandra is shown in *blue*, mid-infrared emission (Spitzer) shown in *red*, and near-IR radiation (HST) in *yellow-brown*. The long side of the image is 32′.5, corresponding to $\sim$ 75 pc at the distance of the Galactic center. The Galactic center, in which a supermassive black hole is suspected to reside, is the bright white region to the right of the center of this image. The X-ray image contains hundreds of white dwarfs, neutron stars, and black holes that radiate in the X-ray regime due to accretion phenomena (accreting X-ray binaries). The diffuse X-ray emission originates in diffuse hot gas with a temperature of about $T \sim 10^7$ K. Credit: NASA, ESA, CXC, SSC, and STScI

positions between radio and IR observations is only $\sim$ 30 mas—at a presumed distance of the GC of 8 kpc, 1 arcsec corresponds to 0.0388 pc, or about 8000 AU.

2.6.2 The central star cluster

Density distribution. Observations in the K-band ($\lambda \sim 2\mu$m) show a compact star cluster that is centered on Sgr A*. Its density behaves like $\propto r^{-1.8}$ within the distance range 0.1 pc $\lesssim r \lesssim$ 1 pc. The number density of stars in its inner region is so large that close stellar encounters may be common. It can be estimated that a star has a close encounter about every $\sim 10^6$ yr. Thus, it is expected that the distribution of the stars is 'thermalized', which means that the local velocity distribution of the stars is the same everywhere, i.e., it is close to a Maxwellian distribution with a constant velocity dispersion. For such an isothermal distribution we expect a density profile $n \propto r^{-2}$, which is in good agreement with the observation. Most of the stars in the nuclear star cluster have an age $\gtrsim$ 1 Gyr; they are late-type giant stars.

precisely 'on top of each other', knowledge of exact relative astrometry is essential. This can be gained if a population of compact sources exists that is observable in both wavelength domains and for which accurate positions can be measured.

In addition, young O and B stars are found in the central parsec. From their spectroscopic observations, it was inferred that almost all of these hot, young stars reside in one of two rotating thick disks. These disks are strongly inclined to the Galactic plane, one rotates 'clockwise' around the GC, the other 'counterclockwise'. These two disks have a clearly defined inner edge at about 1″, corresponding to 0.04 pc, and a surface mass density $\propto r^{-2}$. The age of these early-type stars is 6 ± 2 Myr, i.e., of the same order as the time between two strong encounters.

Another observational result yields a striking and interesting discrepancy with respect to the idea of an isothermal distribution. Instead of the expected constant dispersion σ of the radial velocities of the stars, a strong radial dependence is observed: σ increases towards smaller r. For example, one finds $\sigma \sim 55$ km/s at $r = 5$ pc, but $\sigma \sim 180$ km/s at $r = 0.15$ pc. This discrepancy indicates that the gravitational potential in which the stars are moving is generated not only by themselves. According to the virial theorem, the strong increase of σ for small r implies the presence of a central mass concentration in the star cluster.

The origin of very massive stars near the GC. One of the unsolved problems is the presence of these massive stars close to the Galactic center. One finds that most of the innermost stars are main-sequence B-stars. Their small lifetime of

$\sim 10^8$ yr probably implies that these stars were born close to the Galactic center. This, however, is very difficult to understand. Both the strong tidal gravitational field of the central black hole (see below) and the presumably strong magnetic field in this region will prevent the 'standard' star-formation picture of a collapsing molecular cloud: the former effect tends to disrupt such a cloud while the latter stabilizes it against gravitational contraction. In order to form the early-type stars found in the inner parsec of the Galaxy, the gas clouds need to be considerably denser than currently observed, but may have been at some earlier time during a phase of strong gas infall. Several other solutions to this problem have been suggested. Perhaps the most plausible is a scenario in which the stars are born at larger distances from the Galactic center and then brought there by dynamical processes, involving strong gravitational scattering events. If a stellar binary has an orbit which brings it close to the central region, the strong tidal gravitational field can disrupt the binary, with one of its star being brought into a gravitationally bound orbit around the black hole, and the other being expelled from the central region.

Proper motions. Since the middle of the 1990s, proper motions of stars in this star cluster have also been measured, using the methods of speckle interferometry and adaptive optics. These produce images at diffraction-limited angular resolution, about $\sim 0''.15$ in the K-band at the ESO/NTT (3.5 m) and about $\sim 0''.05$ at 10 m-class telescopes. Proper motions are currently known for about 6000 stars within ~ 1 pc of Sgr A*, of which some 700 additionally have radial velocity measurements, so that their three-dimensional velocity vector is known. The radial and tangential velocity dispersions resulting from these measurements are in good mutual agreement. Thus, it can be concluded that a basically isotropic distribution of the stellar orbits exists, simplifying the study of the dynamics of this stellar cluster.

2.6.3 A black hole in the center of the Milky Way

The S-star cluster. Whereas the distribution of young A-stars in the nuclear disks shows a sharp cut-off at around $1''$, there is a distribution of stars within $\sim 1''$ of Sgr A* which is composed mainly of B-stars; these are known as the S-star cluster. Some stars of this cluster have a proper motion well in excess of 1000 km/s, up to $\sim$ 10000 km/s. Combining the velocity dispersions in radial and tangential directions reveals them to be increasing according to the Kepler law for the presence of a point mass, $\sigma \propto r^{-1/2}$ down to $r \sim 0.01$ pc.

By now, the *acceleration* of some stars in the star cluster has also been measured, i.e., the change of proper motion with time, which is a direct measure of the gravitational force. From these measurements Sgr A* indeed emerges as the focus of the orbits and thus as the center of mass. For $\sim$ 25 members of the S-star cluster, the information from proper motion and radial velocity measurements allowed the reconstruction of orbits; these are shown in the left-hand panel of Fig. 2.45. For one of these stars, S2, observations between 1992 and 2008 have covered a full orbit around Sgr A*, with an orbital period of 15.8 yr, as shown in the right panels of Fig. 2.45. Its velocity exceeded 5000 km/s when it was closest to Sgr A*. The minimum separation of this star from Sgr A* then was only 6×10^{-4} pc, or about 100 AU. In 2012, a new S-star with a period of only 11.5 yr was discovered.

From the observed kinematics of the stars, the enclosed mass $M(r)$ can be calculated, see Fig. 2.46. The corresponding analysis yields that $M(r)$ is basically constant over the range $0.01\,\mathrm{pc} \lesssim r \lesssim 0.5\,\mathrm{pc}$. This exciting result clearly indicates the presence of a point mass. The precise value of this mass is a bit uncertain, mainly due to the uncertainty in the distance of the Galactic center from us. A characteristic value obtained from recent analysis yields a distance to the Galactic center of $R_0 \approx 8.3$ kpc, and a black hole mass of

$$\boxed{M = (4.3 \pm 0.4) \times 10^6 M_\odot}\,, \qquad (2.94)$$

which is slightly larger than the estimate based on the data shown in Fig. 2.46. For radii above $\sim$ 1 pc, the mass of the star cluster dominates; it nearly follows an isothermal density distribution with a core radius of $\sim$ 0.34 pc and a central density of $3.6 \times 10^6 M_\odot/\mathrm{pc}^3$. This result is also compatible with the kinematics of the gas in the center of the Galaxy. However, stars are much better kinematic indicators because gas can be affected by magnetic fields, viscosity, and various other processes besides gravity.

> The kinematics of stars in the central star cluster of the Galaxy shows that our Milky Way contains a mass concentration in which $\sim 4 \times 10^6 M_\odot$ are concentrated within a region smaller than 0.01 pc. This is almost certainly a black hole in the center of our Galaxy, at the position of the compact radio source Sgr A*.

Why a black hole? We have interpreted the central mass concentration as a black hole; this requires some further explanation:

- The energy for the central activity in quasars, radio galaxies, and other AGNs is produced by accretion of gas onto a supermassive black hole (SMBH); we will discuss this in more detail in Sect. 5.3. Thus we know that at least a sub-class of galaxies contains a central SMBH.

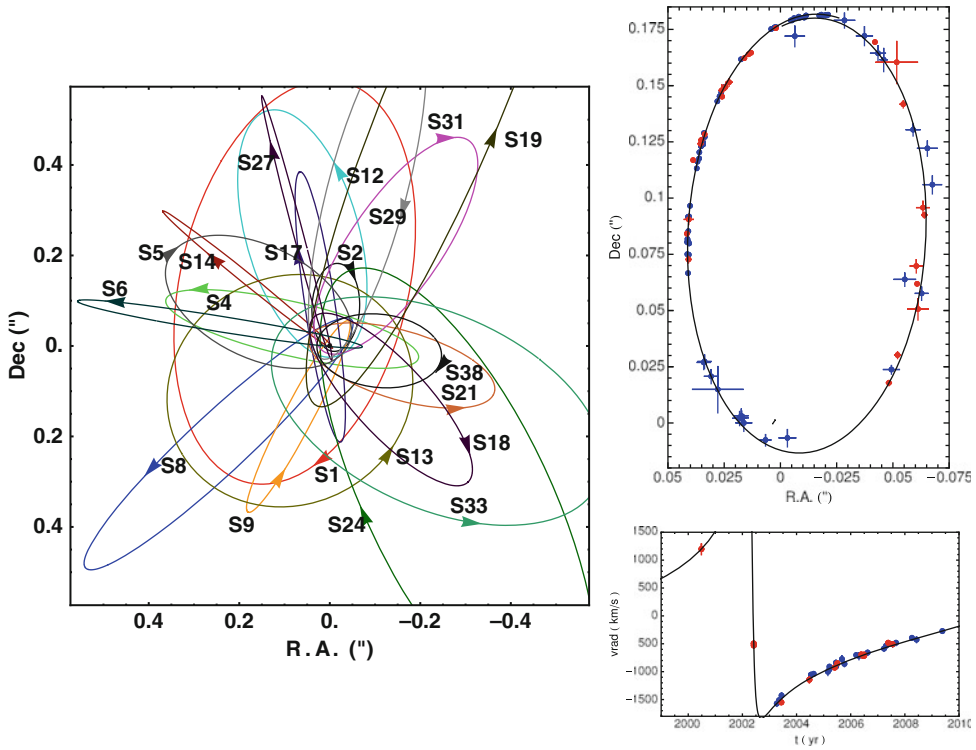


Fig. 2.45 The *left figure* shows the orbits of about two dozen stars in the central arcsecond around Sgr A*, as determined from their measured proper motions and radial velocity. For one of the stars, denoted by S2, a full orbit has been observed, as shown in the *upper right panel*. The data shown here were obtained between 1992 and 2008, using data taken with the NTT and the VLT (*blue points*) and the Keck telescopes (*red points*). The orbital time is 15.8 yr, and the orbit has a strong eccentricity. The *lower right panel* shows the radial velocity measurements of S2. In both of the *right panels*, the best fitting model for the orbital motion is plotted as a *curve*. Source: *Left*: S. Gillessen et al. 2009, *Monitoring Stellar Orbits Around the Massive Black Hole in the Galactic Center*, ApJ 692, 1075, p. 1096, Fig. 16. ©AAS. Reproduced with permission. *Right*: S. Gillessen et al. 2009, *The Orbit of the Star S2 Around SGR A* from Very Large Telescope and Keck Data*, ApJ 707, L114, p. L115, L116, Figs. 2 & 3. ©AAS. Reproduced with permission

Furthermore, we will see in Sect. 3.8 that many 'normal' galaxies, especially ellipticals, harbor a black hole in their center. The presence of a black hole in the center of our own Galaxy would therefore not be something unusual.

- To bring the radial mass profile $M(r)$, as inferred from the stellar kinematics, into accordance with an extended mass distribution, its density distribution must be very strongly concentrated, with a density profile steeper than $\propto r^{-4}$; otherwise the mass profile $M(r)$ would not be as flat as observed and shown in Fig. 2.46. Hence, this hypothetical mass distribution must be vastly different from the expected isothermal distribution which has a mass profile $\propto r^{-2}$, as discussed in Sect. 2.6.2. However, observations of the stellar distribution provide no indication of an inwardly increasing density of the star cluster with such a steep profile.
- Even if such an ultra-dense star cluster (with a central density of $\gtrsim 4 \times 10^{12} M_\odot/\mathrm{pc}^3$) was present it could not be stable, but instead would dissolve within $\sim 10^7$ yr through frequent stellar collisions.
- Sgr A* itself has a proper motion of less than 20 km/s. It is therefore the dynamical center of the Milky Way. Due to the large velocities of its surrounding stars, one would derive a mass of $M \gg 10^3 M_\odot$ for the radio source, assuming equipartition of energy (see also Sect. 2.6.4). Together with the tight upper bounds for its extent, a lower limit for the density of $10^{18} M_\odot/\mathrm{pc}^3$ can then be obtained.

We have to emphasize at this point that the gravitational effect of the black hole on the motion of stars and gas is constrained to the innermost region of the Milky Way. As one can see from Fig. 2.46, the gravitational field of the SMBH dominates the rotation curve of the Galaxy only for $R \lesssim 2$ pc—this is the very reason why the detection of the SMBH is so difficult. At larger radii, the presence of the SMBH is of no relevance for the rotation curve of the Milky Way.

2.6.4 The proper motion of Sgr A*

From a series of VLBI observations of the position of Sgr A*, covering 8 years, the proper motion of this compact radio source was measured with very high precision. To do this, the

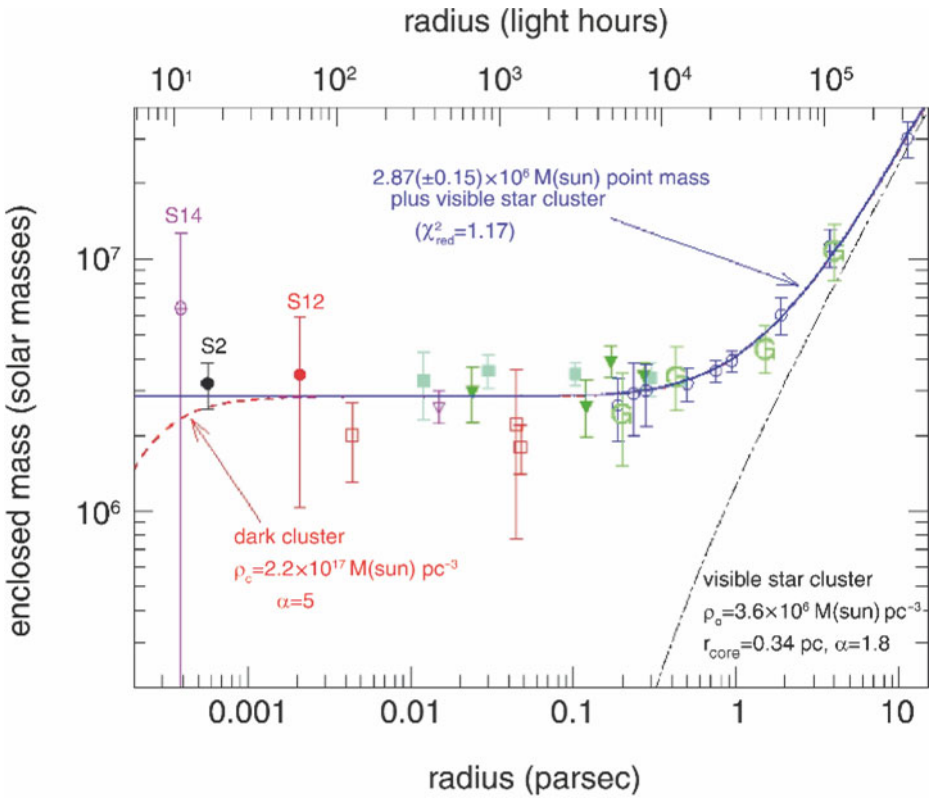


Fig. 2.46 Determination of the mass $M(r)$ within a radius r from Sgr A*, as measured by the radial velocities and proper motions of stars in the central cluster. Mass estimates obtained from individual stars (S14, S2, S12) are given by the points with error bars for small r. The other data points were derived from the kinematic analysis of the observed proper motions of the stars, where different methods have been applied. As can be seen, these methods produce results that are mutually compatible, so that the shape of the mass profile plotted here can be regarded to be robust, whereas the normalization depends on R_0 which was assumed to be 8 kpc for this figure. The *solid curve* is the best-fit model, representing a point mass of $2.9 \times 10^6 M_\odot$ plus a star cluster with a central density of $3.6 \times 10^6 M_\odot/\mathrm{pc}^3$ (the mass profile of this star cluster is indicated by the *dash-dotted curve*). The *dashed curve* shows the mass profile of a hypothetical cluster with a very steep profile, $n \propto r^{-5}$, and a central density of $2.2 \times 10^{17} M_\odot\,\mathrm{pc}^{-3}$. Source: R. Schödel et al. 2003, *Stellar Dynamics in the Central Arcsecond of Our Galaxy*, ApJ 596, 1015, p. 1027, Fig. 11. ©AAS. Reproduced with permission

position of Sgr A* was determined relative to two compact extragalactic radio sources. Due to their large distances these are not expected to show any proper motion, and the VLBI measurements show that their separation vector is indeed constant over time. The position of Sgr A* over the observing period is plotted in Fig. 2.47.

From the plot, we can conclude that the observed proper motion of Sgr A* is essentially parallel to the Galactic plane. The proper motion perpendicular to the Galactic plane is about 0.2 mas/yr, compared to the proper motion in the Galactic plane of 6.4 mas/yr. If $R_0 = (8.0 \pm 0.5)$ kpc is assumed for the distance to the GC, this proper motion translates into an orbital velocity of (241 ± 15) km/s, where the uncertainty is dominated by the exact value of R_0 (the error in the measurement alone would yield an uncertainty of only 1 km/s). This proper motion is easily explained by the Solar orbital motion around the GC, i.e., this measurement contains no hint of a non-zero velocity of the radio source Sgr A* itself. In fact, the small deviation of the proper motion from the orientation of the Galactic plane can be explained by the peculiar velocity of the Sun relative to the LSR (see Sect. 2.4.1). If this is taken into account, a velocity perpendicular to the Galactic disk of $v_\perp = (-0.4 \pm 0.9)$ km/s is obtained for Sgr A*. The component of the orbital velocity within the disk has a much larger uncertainty because we know neither R_0 nor the rotational velocity V_0 of the LSR very precisely. The small upper limit for $v_\perp$ suggests, however, that the motion in the disk should also be very small. Under the (therefore plausible) assumption that Sgr A* has no peculiar velocity, the ratio R_0/V_0 can be determined from these measurements with an as yet unmatched precision.

What also makes this observation so impressive is that from it we can directly derive a lower limit for the mass of Sgr A*. Since this radio source is surrounded by $\sim 10^6$ stars within a sphere of radius ~ 1 pc, the net acceleration towards the center is not vanishing, even in the case of a statistically isotropic distribution of stars. Rather, due to the discrete nature of the mass distribution, a stochastic force exists that changes with time because of the orbital motion of the stars. The radio source is accelerated by this force,

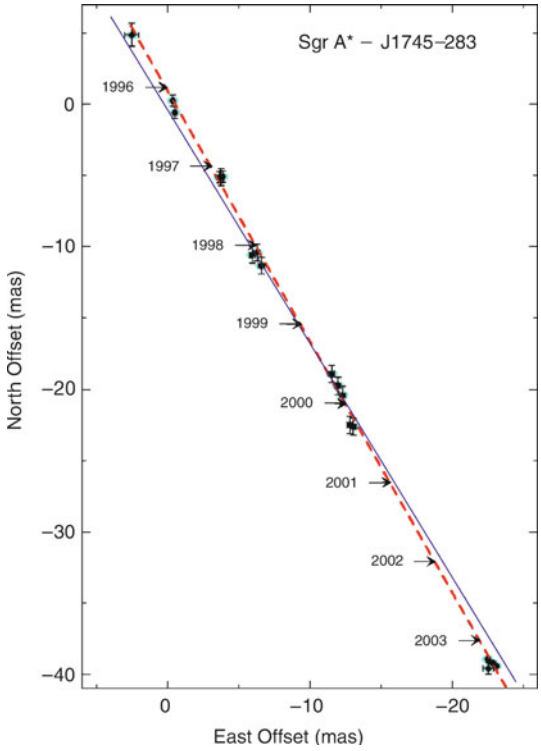


Fig. 2.47 The position of Sgr A* at different epochs, relative to the position in 1996. To a very good approximation the motion is linear, as indicated by the dashed best-fit *straight line*. In comparison, the *solid line* shows the orientation of the Galactic plane. Source: M. Reid & A. Brunthaler 2004, *The Proper Motion of Sagittarius A*. II. The Mass of Sagittarius A**, ApJ 616, 872, p. 875, Fig. 1. ©AAS. Reproduced with permission

causing a motion of Sgr A* which becomes larger the smaller the mass of the source. The very strong limits to the velocity of Sgr A* enable us to derive a lower limit for its mass of $0.4 \times 10^6 M_\odot$. This mass limit is significantly lower than the mass of the SMBH that was derived from the stellar orbits, but it is the mass of the radio source itself. Although we have excellent reasons to assume that Sgr A* coincides with the SMBH, the upper limit on the peculiar velocity of Sgr A* is the first proof for a large mass of the radio source itself.

2.6.5 Flares from the Galactic center

Observation of flares. In 2000, the X-ray satellite Chandra discovered a powerful X-ray flare from Sgr A*. This event lasted for about 3 h, and the X-ray flux increased by a factor of 40 during this period. XMM-Newton confirmed the existence of X-ray flares, recording one where the luminosity increased by a factor of $\sim$ 200. Most of the flares seen, however, have a much smaller peak amplitude, of a few to ten times the quiescent flux of the source, and the typical flare duration is $\sim$ 30 min. During the flares, variability of the X-flux on time-scales of several minutes is observed. Combining the flare duration with the short time-scale of variability of a few minutes indicates that the emission must originate from a very small source, not larger than $\sim 10^{13}$ cm in size.

Monitoring Sgr A* at longer wavelengths, variability was found as well. Figure 2.48 shows the simultaneous lightcurves of Sgr A* during one night in May 2009. The source flared in X-rays, with two flares close in time. These flares are also seen at the near-IR, sub-mm and mm wavelengths, nearly simultaneously. Flares are seen more frequently in the NIR than in X-rays, occurring several times per day, where X-ray flares occur about once per day. Simultaneous observations, such as those in Fig. 2.48, indicate that every X-ray flare is accompanied by a flare in the NIR; the converse is not true, however. It thus seems that the flares in the different wavelength regimes have a common origin. From a set of such simultaneous observing campaigns, it was found that there is a time lag between the variations at different wavelengths. Typically, NIR flares occur $\sim$ 2 h earlier than those seen at (sub-)mm wavelengths, and they are narrower, whereas the X-ray and NIR flares are essentially simultaneous.

There was some discussion about a possible quasi-periodicity of the NIR light curves, but the observational evidence for this is not unambiguous. Nevertheless, polarization observation of Sgr A* may provide support for a model in which the variability is caused by a source moving around the central black hole. Anticipating our discussion about AGN in Chap. 5, the model assumes that there is a 'hot spot' on the surface of an accretion disk, whereby relativistic effects modulate the received flux from this source component as it orbits around the black hole.

X-ray echos. With a mass of $M_\bullet \approx 4 \times 10^6\, M_\odot$, the central black hole in the Milky Way could in principle power a rather luminous active galactic nucleus, such as is observed in other active galaxies, e.g., Seyfert galaxies. This, however, is obviously not the case—the luminosity of Sgr A* is many orders of magnitudes smaller than the nucleus in Seyfert galaxies with similar mass central black holes (see Chap. 5). The reason for the inactivity of our Galactic center is therefore not the black hole mass, but the absence of matter which is accreted onto it. The fact that the Galactic center region emits thermal radiation in the X-rays shows the presence of gas. But this gas cannot flow to the central black hole, presumably because of its high temperature and associated high pressure. This line of argument is supported by the fact that the central X-ray source is resolved, and hence much more extended than the Schwarzschild radius of the black hole, where the bulk of the energy generation by accretion takes place (see Sect. 5.3.2 for more details). However, the variability of Sgr A* may be seen as an indication that the accretion rate can change in the course of time.

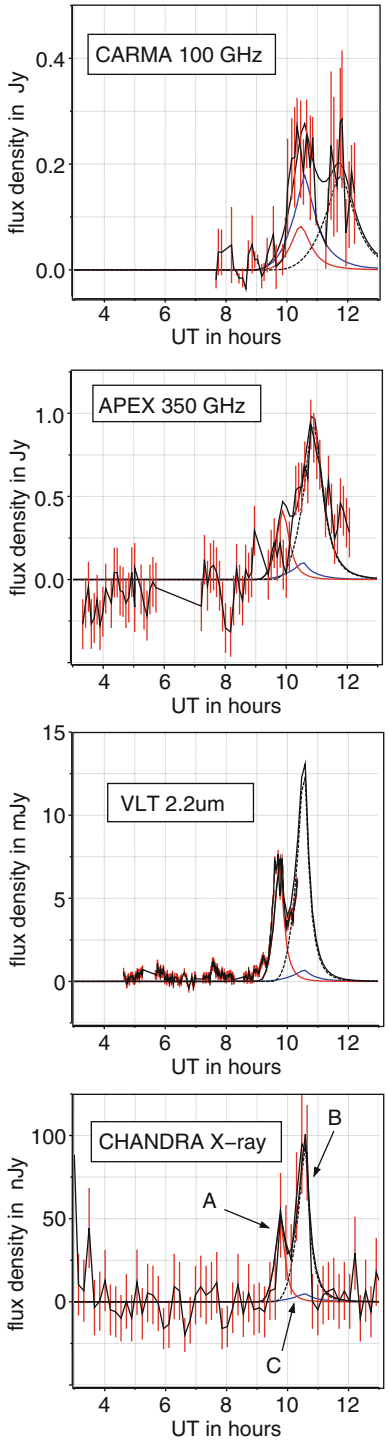


Fig. 2.48 Variability of Sgr A* is shown here in simultaneous observations at four different wavelengths, carried out in May 2009. The *red bars* in each panel are the error bars of the observed flux, from which the quiescent flux level was subtracted, with their central values connected with a *thin line*. The *thick solid curve* corresponds to a model for the flare emission across the wavebands, whereas the other three curves (*dashed*, *blue* and *red*) are individual components of this model. One sees that the first flare occurs at all wavelength, whereas the second, main flare, was not covered by the near-IR observations. Source: A. Eckart et al. 2012, *Millimeter to X-ray flares from Sagittarius A**, A&A 537, A52, Fig. 1. ©ESO. Reproduced with permission

Maybe there have been times when the luminosity of Sgr A* was considerably larger than it is currently. Indeed, there are some indications for this being the case. Photons emitted at earlier times than the ones we observe now from Sgr A* may still reach us today, if they were scattered by electrons, or if these photons have exited gas that, as a consequence, emits radiation. In both cases, the total light-travel time from the source to us would be larger, since the geometric light path is longer. We may therefore see the evidence of past activity as a light echo of radiation, which reaches us from slightly different directions.

There is now strong evidence for such a light echo. Hard X-ray radiation can lead to the removal of a strongly bound electron in iron, which subsequently emits a fluorescence line at 6.4 keV. The distribution of this iron line radiation in a region close to Sgr A* is shown in Fig. 2.49. This region contains a large number of molecular clouds, i.e., high-density neutral gas. The images in Fig. 2.49 show the variation of this line flux over a time period of about 5 year. We see that the spatial distribution of this line flux changes over this time-scale, with the flux increasing to the left part of this region as time progresses. The apparent velocity, with which the peak of the line emission moves across the region, is considerably larger than the speed of light—it shows superluminal motion. This evidence has recently been further strengthened with Chandra observations of the same region showing variations on even shorter time-scales. This high velocity, however, is not necessarily a violation of Einstein's Special Relativity. In fact, this phenomenon can be easily understood in the framework of a reflection model: Suppose there is a screen of scattering material between us and a source. The further away a point in the screen is from the line connecting us and the source, the larger is the geometrical path of a ray which propagates to this point in the screen, and is scattered there towards our direction. The scattered radiation from a flare in the source will thus appear at different points in the screen as time progresses, and the speed with which the point changes in the screen can exceed the speed of light, without violating relativity; this will be shown explicitly in Problem 2.6. The argument is the same, independent of whether the light is scattered, or if a fluorescence line is excited. In fact, the material responsible for the light echo does not need to be located between us and the source, it can also be located behind the source.

In fact, this phenomenon has not only be seen in the region shown in Fig. 2.49. The massive molecular cloud Sgr B2 also shows the prominent fluorescence line of iron, as well as X-ray continuum emission. The line and continuum flux decreased by a factor $\sim$ 2 over a time-scale of $\sim$ 10 yr—whereas the extent of the molecular cloud is much larger than ten light-years. Furthermore, there is no strong X-ray source known close to Sgr B2 which would be able to power the fluorescence line.

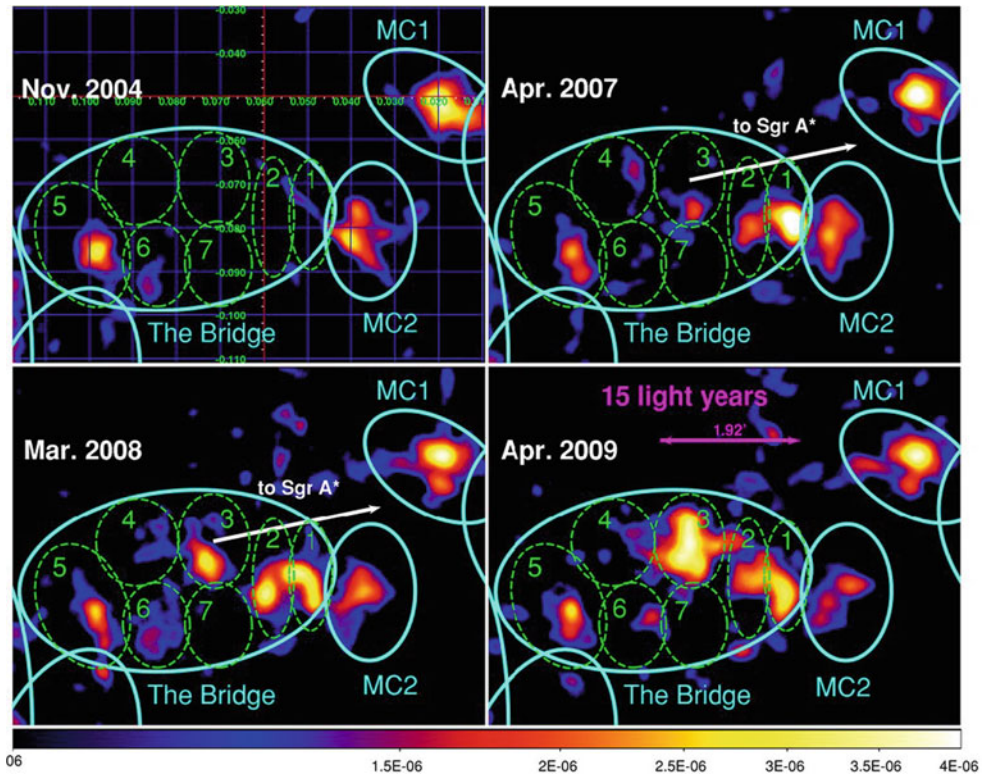


Fig. 2.49 The flux distribution of the 6.4 keV iron line in the region of molecular clouds near the Galactic center, at four different epochs. These XMM observations show that the flux distribution is changing on time-scales of a few years. However, the size of the region is much larger than a few light years—see the scale bar in the *lower right panel*. Thus, it seems that the variations are propagating through this region with a velocity larger than the speed of light. The explanation for this phenomenon is the occurrence of a light echo. Sgr A* is located in the direction indicated by the *white arrow* in the *upper right panel*, at a projected distance of about 40 light-years from the molecular cloud MC2. Source: G. Ponti et al. 2010, *Discovery of a Superluminal Fe K Echo at the Galactic Center: The Glorious Past of Sgr A* Preserved by Molecular Clouds*, ApJ 714, 732, p. 742, Fig. 10. ©AAS. Reproduced with permission

These observations are compatible with a model in which Sgr A* had a strong flare some 100 yr ago, and what we see are the light echos of this flare. The luminosity of the flare must have exceeded 2×10^{39} erg/s in the X-ray regime, and it must have faded rather quickly, in order to generate such short-term variations of the echo. The location of the flare must be located in a region close to Sgr A*, though one cannot conclude with certainty that Sgr A* was the exact location—there are several compact stellar remnants in its immediate vicinity which may have caused such a flare. Nevertheless, the requested luminosity is higher than that one usually assigns to compact stellar-mass objects, and Sgr A* as the putative source of the flare appears quite likely. Hence, the light echo phenomenon gives us an opportunity to look back in time.

The Fermi bubbles. Another potential hint for an increased nuclear activity of the Galactic center was found with the Fermi satellite. It discovered two large structures in gamma-rays above and below the Galactic center, extending up to a Galactic latitude of $|b| \lesssim 50°$, i.e., a spatial scale of ~ 8 kpc from the center of the Milky Way (see Fig. 2.50). Emission from these regions is seen in the energy range between 1 and 100 GeV, with a hard energy spectrum, much harder than the diffuse gamma-ray emission from the Milky Way. The two 'Fermi bubbles' are associated with an enhanced microwave emission, seen by the WMAP and Planck satellites (the so-called microwave 'haze'), and appear to have well-defined edges, which are also seen in X-rays. Furthermore, almost spatially coincident giant radio lobes with strong linear polarization were detected.

The origin of the Fermi bubbles is currently strongly debated in the literature. One possibility is strongly enhanced activity of Sgr A* in the past, that drove out a strong flow of energetic plasma—similar to AGNs—and whose remnant we still see. Alternatively, the Galactic center region is a site of active star formation, which may be the origin of a massive outflow of magnetized plasma.

2.6.6 Hypervelocity stars in the Galaxy

Discovery. In 2005, a Galactic star was discovered which travels with a velocity of at least 700 km/s relative to the

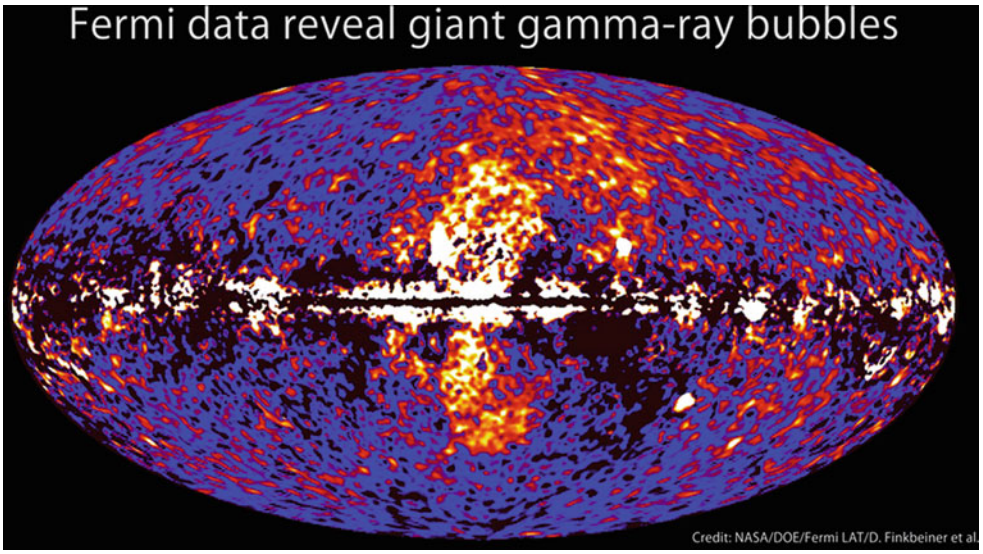


Fig. 2.50 Gamma-ray map of the sky in the energy range between 1 and 10 GeV. The Fermi-bubbles show up above and below the Galactic center, extending up to $\sim 50^\circ$ from the disk. Credit: NASA/DOE/Fermi LAT/D. Finkbeiner et al.

Galactic rest frame. This B-star has a distance of 110 kpc from the Galactic center, and its actual space velocity depends on its transverse motion which has not yet been measured, due to the large distance of the object from us. However, since the distance of the star is far larger than the separation between the Sun and the Galactic center, so that the directions Galactic center–star and Sun–star are nearly the same, the measured radial velocity from the Sun is very close to the radial velocity relative to the Galactic center.

The velocity of this star is so large that it greatly exceeds the escape velocity from the Galaxy; hence, this star is gravitationally unbound to the Milky Way. Within 4 years after this first discovery, about 15 more such *hypervelocity stars* were discovered, all of them early-type stars (O- or B-stars) with Galactic rest-frame velocities in excess of the escape velocity at their respective distance from the Galactic center. Hence, they will all escape the gravitational potential of the Galaxy. Furthermore, a larger number of stars have been detected whose velocity in the Galactic frame exceeds $\sim 300\,\mathrm{km/s}$ but is most likely not large enough to let them escape from the gravitational field of the Galaxy—i.e., these stars are on bound orbits. In a sample of eight of them, all were found to move away from the Galactic center. This indicates that their lifetime is considerable smaller than their orbital time scale (because otherwise, if they could survive for half an orbital period, one would expect to find also approaching stars), yielding an upper bound on their lifetime of 2 Gyr. Therefore, these stars are most likely on the main sequence.

Acceleration of hypervelocity stars. The fact that the hypervelocity stars are gravitationally unbound to the Milky Way implies that they must have been accelerated very recently, i.e., less than a crossing time through the Galaxy ago. In addition, since they are early-type stars, they must have been accelerated within the lifetime of such stars. The acceleration mechanism must be of gravitational origin and is related to the *dynamical instability* of N-body systems, with $N > 2$. A pair of objects will orbit in their joint gravitational field, either on bound orbits (ellipses) or unbound ones (gravitational scattering on hyperbolic orbits); in the former case, the system is stable and the two masses will orbit around each other literally forever. If more than two masses are involved this is no longer the case—such a system is inherently unstable. Consider three masses, initially bound to each other, orbiting around their center-of-mass. In general, their orbits will not be ellipses but are more complicated; in particular, they are not periodic. Such a system is, mathematically speaking, chaotic. A chaotic system is characterized by the property that the state of a system at time t depends very sensitively on the initial conditions set at time $t_\mathrm{i} < t$. Whereas for a dynamically stable system the positions and velocities of the masses at time t are changed only a little if their initial conditions are slightly varied (e.g., by giving one of the masses a slightly larger velocity), in a chaotic, dynamically unstable system even tiny changes in the initial conditions can lead to completely different states at later times. Any N-body system with $N > 2$ is dynamically unstable.

Back to our three-body system. The three masses may orbit around each other for an extended period of time, but their gravitational interaction may then change the state of the system suddenly, in that one of the three masses attains a sufficiently high velocity relative to the other two and may escape to infinity, whereas the other two masses form a binary system. What was a bound system initially may become an unbound system later on. This behavior may appear unphysical at first sight—where does the energy come from to eject one of the stars? Is this process violating energy conservation?

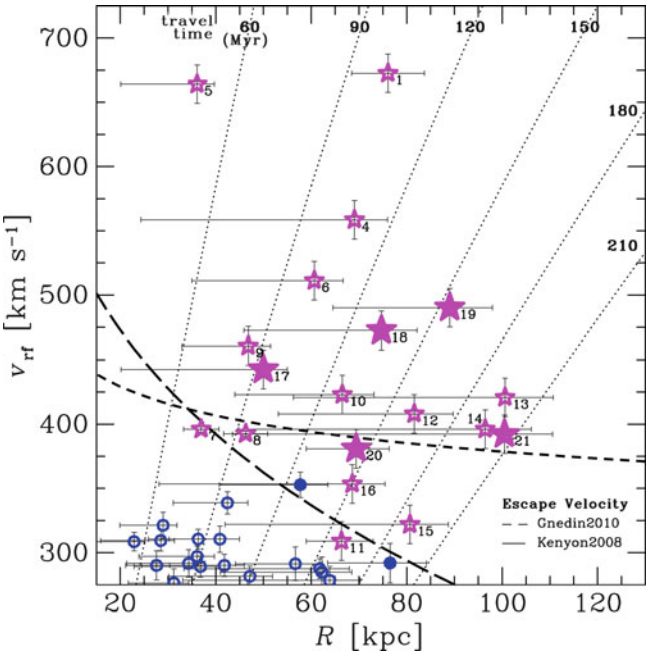


Fig. 2.51 The minimum velocity in the Galactic rest frame is plotted against the distance from the Galactic center, for a total of 37 stars. The *star symbols* show hypervelocity stars, whereas *circles* are stars which are possibly on gravitationally bound orbits in the Galaxy. The *long-* and *short-dashed curves* indicate the escape velocity from the Milky Way, as a function of distance, according to two different models for the total mass distribution in the Galaxy. The *dotted curves* indicate constant travel time of stars from the Galactic center to a given distance with current space velocity, labeled by this time in units of 10^6 yr. The distances are estimated assuming that the stars are on the main sequence, whereas the error bars indicate the plausible range of distances if these stars were on the blue horizontal branch. Source: W.R. Brown, M.J. Geller & S.J. Kenyon 2012, *MMT Hypervelocity Star Survey. II. Five New Unbound Stars*, ApJ 751, 55, p. 5, Fig. 3. ©AAS. Reproduced with permission

Of course not! The trick lies in the properties of gravity: a binary has *negative* binding energy, and the more negative, the tighter the binary orbit is. By three-body interactions, the orbit of two masses can become tighter (one says that the binary 'hardens'), and the corresponding excess energy is transferred to the third mass which may then become gravitationally unbound. In fact, a single binary of compact stars can in principle take up all the binding energy of a star cluster and 'evaporate' all other stars.

This discussion then leads to the explanation of hypervelocity stars. The characteristic escape velocity of the 'third mass' will be the orbital velocity of the three-body system before the escape. The only place in our Milky Way where orbital velocities are as high as that observed for the hypervelocity stars is the Galactic center. In fact, the travel time of a star with current velocity of $\sim$ 600 km/s from the Galactic center to Galacto-centric distances of $\sim$ 80 kpc is of order 10^8 yr (see Fig. 2.51), slightly shorter than the main-sequence lifetime of a B-star. Furthermore, most of the bright stars in the central 1″ of the Galactic center region are B-stars. Therefore, the immediate environment of the central black hole is the natural origin for these hypervelocity stars. Indeed, long before their discovery the existence of such stars was predicted. When a binary system gets close to the black hole, this three-body interaction can lead to the ejection of one of the two stars into an unbound orbit, whereas the other star gets bound to the black hole. This is considered the most plausible explanation for the presence of young stars (like the B-stars of the S-star cluster) near to the black hole. Thus, the existence of hypervelocity stars can be considered as an additional piece of evidence for the presence of a central black hole in our Galaxy.

For one of the hypervelocity stars, the time to travel from the Galactic center to its current position is estimated to be much longer than its main sequence lifetime, by a factor of $\sim$ 3. Given that it is located just 16° away from the Large Magellanic Cloud, it was suggested that it had been ejected from there. However, for this star a proper motion was measured with HST, and its direction is fully compatible with coming from the Galactic center, ruling out an LMC origin. Therefore, that star is not a main sequence star, but most likely a so-called blue straggler.

The acceleration of hypervelocity stars near the Galactic center may not be the only possible mechanism. Another suggested origin can be related to the possible existence of intermediate black holes with $M_\bullet \sim 10^3 M_\odot$, either at the center of dense star clusters or as freely propagating in the Milky Way, and may be the relics of earlier accretion events of low-mass galaxies.

Hypervelocity stars are not the only fastly moving stars in the Milky Way, but there is a different population of runaway stars. These stars are created through supernova explosions in binaries. Let us consider a binary, in which the heavier star (the primary) undergoes a supernova explosion, possibly leaving behind a neutron star. During the explosion, the star expels the largest fraction of its mass, on a time-scale that is short compared to the orbital period of the binary, due to the high expansion velocity. Thus, almost instantaneously, the system is transformed into one where the primary star has lost most of its mass. Given that the velocity of the secondary star did not change through this process, thus being the orbital velocity corresponding to the original binary, this velocity is now far larger than the orbital velocity of the new binary. Therefore, the system of secondary and the neutron star are no longer gravitationally bound, and they will both separate, with a velocity similar to the original orbital velocity. For close binaries, this can also exceed 100 km/s, and is the origin of the high space velocities observed for pulsars. However, these runaway stars can hardly be confused with hypervelocity stars, since they are rare and are produced near the Galactic disk.

2.7 Problems

2.1. Angular size of the Moon. The diameter of the Moon is 3476 km, and its mean distance from Earth is about 385 000 km. Calculate the angular diameter of the Moon as seen on the sky. What fraction of the full sky does the Moon cover?

2.2. Helium abundance from stellar evolution. Assume that the baryonic matter M of a galaxy, such as the Milky Way, consisted purely of hydrogen when it was formed. In this case, all heavier elements must have formed from nuclear fusion in the interior of its stellar population. Assume further that the total luminosity L of the galaxy is caused by burning hydrogen into helium, and let this luminosity be constant over the total lifetime of the galaxy, here assumed to be 10^{10} yr, with a correspondingly constant baryonic mass-to-light ratio of $M/L = 3M_\odot/L_\odot$. What is the mass fraction in helium that would be generated by the nuclear fusion process? Would this fraction be large enough to explain the observed helium abundance of $\sim 27\,\%$?

2.3. Flat rotation curve. We saw that the rotation curve of the Milky Way is flat, $V(R) \approx$ const. Assume a spherically-symmetric density distribution $\rho(r)$. Determine the functional form of $\rho(r)$ which yields a flat rotation curve.

2.4. The Sun as a gravitational lens. What is the minimum distance a Solar-like star needs to have from us in order to produce multiple images of very distant sources, and how large would the achievable image splitting be? Make use of the fact that the angular diameter of the Sun is $32'$ on average.

2.5. Kepler rotation around the Galactic center black hole. We have mentioned that the Galactic center hosts a star cluster with a characteristic velocity dispersion of $\sim 55\,\mathrm{km/s}$ at $r \gtrsim 4\,\mathrm{pc}$. How does this velocity compare with the circular velocity of an object around the central SMBH? Make use of the fact that $\sqrt{GM_\odot/c^2} = 1.495\,\mathrm{km}$, the so-called gravitational radius corresponding to a Solar mass.

2.6. Superluminal motion through scattering. Assume that there is a (infinitely thin) sheet of scattering material between us and the Galactic center (GC). Let that screen be perpendicular to the line-of-sight to the GC, and have a distance D from the GC, so that our distance to this screen is $D_{\mathrm{sc}} = R_0 - D$. A light flash at the GC will be seen in scattered light as a ring whose radius changes in time. Calculate the radius $R(t)$ of this ring, and determine its apparent velocity $\mathrm{d}R/\mathrm{d}t$. Can that be larger than the velocity of light? Assume that the opening angle of the ring, as seen both by the GC and by us, is small, so that $R/D \ll 1$, $R/D_{\mathrm{sc}} \ll 1$. Furthermore, assume that the screen is close to the Galactic center, so that $D \ll R_0$. Can you get a similar effect from a scattering screen behind the Galactic center?

The world of galaxies

3

The insight that our Milky Way is just one of many galaxies in the Universe is less than 100 years old, despite the fact that many had already been known for a long time. The catalog by Charles Messier (1730–1817), for instance, lists 103 diffuse objects. Among them M31, the Andromeda galaxy, is listed as the 31st entry in the Messier catalog. Later, this catalogue was extended to 110 objects. John Dreyer (1852–1926) published the *New General Catalog (NGC)* that contains nearly 8000 objects, most of them galaxies. Spiral structure in some of the nebulae was discovered in 1845 by William Parsons, and in 1912, Vesto Slipher found that the spiral nebulae are rotating, using spectroscopic analysis. But the nature of these extended sources, then called nebulae, was still unknown at that time; it was unclear whether they are part of our Milky Way or outside it.

The nature of the nebulae. The year 1920 saw a public debate (the Great Debate) between Harlow Shapley and Heber Curtis. Shapley believed that the nebulae are part of our Milky Way, whereas Curtis was convinced that the nebulae must be objects located outside the Galaxy. The arguments which the two opponents brought forward were partly based on assumptions which later turned out to be invalid, as well as on incorrect data. Much of the controversy can be traced back to the fact that at that time it was not known that dust in the Galactic disk leads to an extinction of distant objects. We will not go into the details of their arguments which were partially linked to the assumed size of the Milky Way since, only a few years later, the question of the nature of the nebulae was resolved.

In 1925, Edwin Hubble discovered Cepheids in Andromeda (M31). Using the period-luminosity relation for these pulsating stars (see Sect. 2.2.7) he derived a distance of 285 kpc. This value is a factor of $\sim$ 3 smaller than the distance of M31 known today, but it provided clear evidence that M31, and thus also other spiral nebulae, must be extragalactic. This then immediately implied that they consist of innumerable stars, like our Milky Way. Hubble's results were considered conclusive by his contemporaries and marked the beginning of extragalactic astronomy. It is not coincidental that at this time George Hale began to arrange the funding for an ambitious project. In 1928 he obtained six million dollars for the construction of the 5 m telescope on Mt. Palomar which was completed in 1949.

Outline of this chapter. This chapter is about galaxies. We will confine the consideration here to 'normal' galaxies in the local Universe; galaxies at large distances, some of which are in a very early evolutionary state, will be discussed in Chap. 9, and active galaxies, like quasars for example, will be discussed later in Chap. 5. In Sect. 3.1, a classification scheme of galaxies that was introduced by Edwin Hubble will be described; most of the luminous galaxies in the local Universe find their place on this *Hubble sequence of galaxies*. The properties of the two main types of galaxies, elliptical and spiral galaxies, are then described in more detail in the following two sections. In Sect. 3.4, we will show that the parameters describing elliptical and spiral galaxies, such as mass, luminosity and size, have a quite regular distribution; the various galaxy properties are strongly mutually related, giving rise to so-called scaling relations.

We will then turn in Sect. 3.5 to investigating the stellar population of galaxies, in particular related to the question of whether the emitted spectral energy distribution of a galaxy can be understood as a sum of the emission of its stars, and how the spectrum of galaxies is related to the properties of the stellar population. The insights gained from that consideration allow us to understand and interpret the finding that the colors of galaxies fall mainly into two groups—they are either red or blue. As we shall see in Sect. 3.6, this offers an alternative classification scheme of galaxies which is independent of their morphology; this obviously comes in handy if one wants to classify galaxies at large distances for which morphological information is much more difficult to obtain, due to their small angular sizes on the sky. We will also see how this new classification fits together with the Hubble sequence.

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_3,

Fig. 3.1 Galaxies occur in different shapes and sizes, and often they are grouped together in groups or clusters. This cluster, the Hercules cluster (also called Abell 2151), lies at a redshift of $z = 0.037$ and contains numerous galaxies of different types and luminosities. The galaxies differ in their morphology, as well as in their colors—spiral galaxies are considerably bluer than elliptical galaxies. In the center of the image, an interacting pair of spiral galaxies (known as NGC 6050/IC 1179, or together as Arp 272) is visible. Credit: Canada-France-Hawaii Telescope/Coelum, Image by Jean-Charles Cuillandre (CFHT) & Giovanni Anselmi (Coelum)

After a short section on the chemical evolution of galaxies, we will describe in Sect. 3.8 evidence for the existence of supermassive black holes in the center of galaxies, with masses ranging up to $10^9 M_\odot$, and for a tight relation between the black hole mass and properties of the stellar component of the galaxies. We then turn to the question on how distances of galaxies can be measured directly, i.e., without employing the Hubble law (1.6). These distance determinations are required in order to calibrate the Hubble law, i.e., to determine the Hubble constant H_0.

The distribution of galaxies in luminosity will be studied in Sect. 3.10; we will see that there exists a characteristic luminosity L^* of galaxies, such that most of the stars in the current Universe are hosted by galaxies whose luminosity varies in a rather narrow interval around L^*. We will see towards the end of this book that the occurrence of this characteristic luminosity (or stellar mass) scale is one of the smoking guns for understanding the cosmic evolution of galaxies. In Sect. 3.11 we describe the gravitational lensing effects caused by massive galaxies, and study some of its applications.

3.1 Classification

Galaxies are observed to have a variety of properties (see Fig. 3.1)—shapes, luminosities, colors, metallicities, etc. The classification of objects depends on the type of observation according to which this classification is made. This is also the case for galaxies. Historically, optical photometry was the method used to observe galaxies. Thus, the morphological classification defined by Hubble is still the best-known today. Besides morphological criteria, color indices, spectroscopic parameters (based on emission or absorption lines),

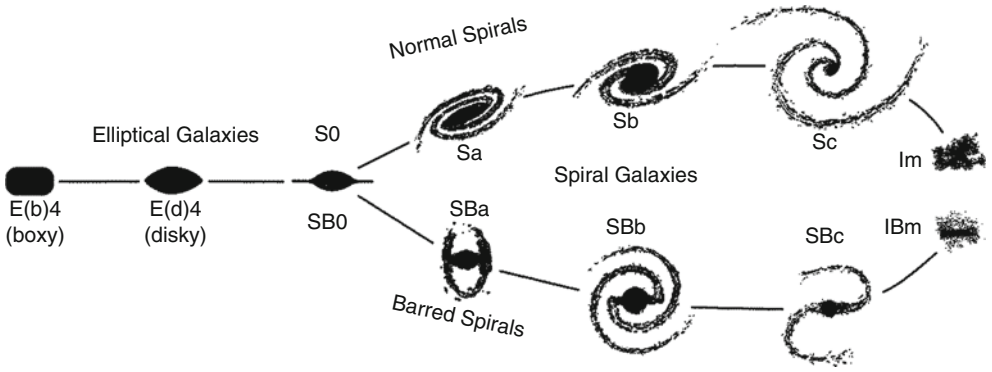


Fig. 3.2 Hubble's 'tuning fork' for galaxy classification. Adapted from: J. Kormendy & R. Bender 1996, *A Proposed Revision of the Hubble Sequence for Elliptical Galaxies*, ApJ 464, L119, Fig. 1. ©AAS. Reproduced with permission

the broad-band spectral distribution (galaxies with/without radio- and/or X-ray emission, or emission in the infrared), as well as other features may also be used.

We start with the Hubble sequence of galaxies, before briefly mention in Sect. 3.1.2 other types of galaxies which do not fit into the Hubble sequence, and outline an alternative classification scheme in Sect. 3.1.3.

3.1.1 Morphological classification: The Hubble sequence

Figure 3.2 shows the classification scheme defined by Hubble. According to this, three main types of galaxies exist (see also Fig. 3.3 for examples):

- *Elliptical galaxies* (E's) are galaxies that have nearly elliptical isophotes[1] without any clearly defined structure. They are subdivided according to their ellipticity $\epsilon \equiv 1 - b/a$, where a and b denote the semi-major and the semi-minor axes, respectively. Ellipticals are found over a relatively broad range in ellipticity, $0 \leq \epsilon \lesssim 0.7$. The notation E$n$ is commonly used to classify the ellipticals with respect to ϵ, with $n = 10\epsilon$; i.e., an E4 galaxy has an axis ratio of $b/a = 0.6$, and E0's have circular isophotes.
- *Spiral galaxies* consist of a disk with spiral arm structure and a central bulge. They are divided into two subclasses: *normal spirals* (S's) and *barred spirals* (SB's). In each of these subclasses, a sequence is defined that is ordered according to the brightness ratio of bulge and disk, and that is denoted by a, ab, b, bc, c, cd, d. Objects along this sequence are often referred to as being either an *early-type* or a *late-type*; hence, an Sa galaxy is an early-type spiral, and an SBc galaxy is a late-type barred spiral. We stress explicitly that this nomenclature is not a statement of the evolutionary stage of the objects but is merely a nomenclature of purely historical origin.
- *Irregular galaxies* (Irr's) are galaxies with only weak (Irr I) or no (Irr II) regular structure. The classification of Irr's is often refined. In particular, the sequence of spirals is extended to the classes Sdm, Sm, Im, and Ir (m stands for Magellanic; the Large Magellanic Cloud is of type SBm).
- *S0 galaxies* are a transition between ellipticals and spirals which are also called lenticulars as they are lentil-shaped galaxies. They contain a bulge and a large enveloping region of relatively unstructured brightness which often appears like a disk without spiral arms. *Ellipticals and S0 galaxies are referred to as early-type galaxies, spirals as late-type galaxies. As before, these names are only historical and are not meant to describe an evolutionary track!*

Obviously, the morphological classification is at least partially affected by projection effects. If, for instance, the spatial shape of an elliptical galaxy is a triaxial ellipsoid, then the observed ellipticity ϵ will depend on its orientation with respect to the line-of-sight. Also, it will be difficult to identify a bar in a spiral that is observed from its side ('edge-on').

Besides the aforementioned main types of galaxy morphologies, others exist which do not fit into the Hubble scheme. Many of these are presumably caused by interaction between galaxies (see below). Furthermore, we observe galaxies with radiation characteristics that differ significantly from the spectral behavior of 'normal' galaxies. These galaxies will be discussed next.

3.1.2 Other types of galaxies

The light from 'normal' galaxies is emitted mainly by stars. Therefore, the spectral distribution of the radiation from such galaxies is in principle a superposition of the spectra of their stellar population. The spectrum of stars is, to a first approximation, described by a Planck function (see Appendix A)

[1] Isophotes are contours along which the surface brightness of a source is constant. If the light profile of a galaxy is elliptical, then its isophotes are ellipses.

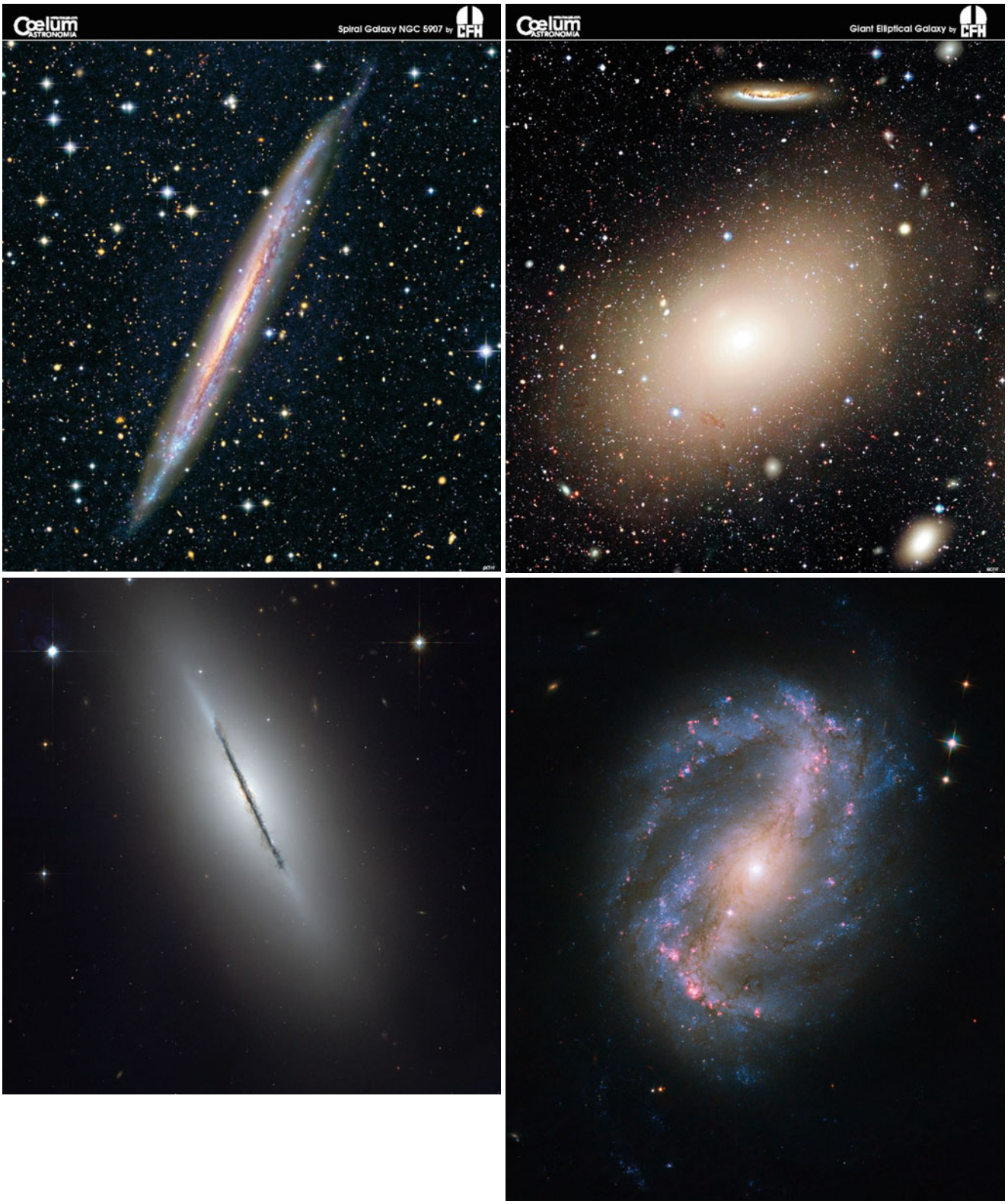


Fig. 3.3 Four galaxies at different locations on the Hubble sequence. NGC 5907 (*top left*) is a large edge-on spiral galaxy whose dust layer inside the stellar disk is seen due to its reddening effect. In contrast, NGC 5866 is an edge-on S0 (lenticular) galaxy (*bottom left*) though a thin disk is visible like in the edge-on spiral galaxy, the morphology is clearly distinct. The *top right* image shows the giant elliptical galaxy M86 located in the Virgo cluster of galaxies, whereas the *bottom right* panel displays the barred spiral galaxy NGC 6217. Credits: *Top right* and *left*: Canada-France-Hawaii Telescope/Coelum, Image by Jean-Charles Cuillandre (CFHT) & Giovanni Anselmi (Coelum). *Bottom left*: NASA, ESA, and The Hubble Heritage Team (STScI/AURA), W. Keel (University of Alabama, Tuscaloosa). *Bottom right*: NASA, ESA, and the Hubble SM4 ERO Team

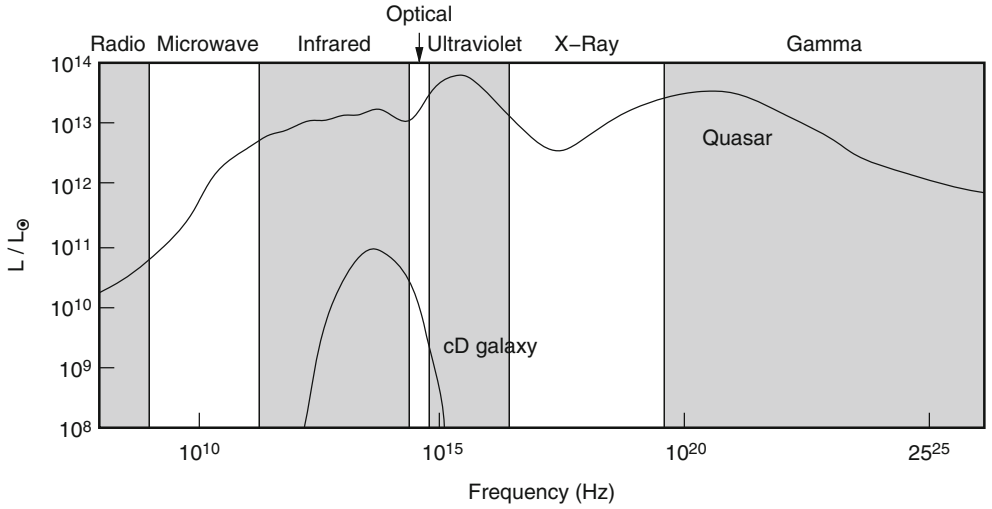


Fig. 3.4 The spectrum of a quasar (3C273) in comparison to that of an elliptical galaxy, in terms of the ratio $\nu L_\nu / L_\odot$. While the radiation from the elliptical is concentrated in a narrow range spanning less than two decades in frequency, the emission from the quasar is observed over the full range of the electromagnetic spectrum, and the energy per logarithmic frequency interval is roughly constant. This demonstrates that the light from the quasar cannot be interpreted as a superposition of stellar spectra, but instead has to be generated by completely different sources and by different radiation mechanisms

that depends only on the star's surface temperature. A typical stellar population covers a temperature range from a few thousand Kelvin up to a few tens of thousand Kelvin. Since the Planck function has a well-localized maximum and from there steeply declines to both sides, most of the energy of such 'normal' galaxies is emitted in a relatively narrow frequency interval that is located in the optical and NIR sections of the spectrum.

In addition to these, other galaxies exist whose spectral distribution cannot be described by a superposition of stellar spectra. One example is the class of active galaxies which generate a significant fraction of their luminosity from gravitational energy that is released in the infall of matter onto a supermassive black hole, as was mentioned in Sect. 1.2.4. The activity of such objects can be recognized in various ways. For example, some of them are very luminous in the radio and/or in the X-ray portion of the spectrum (see Fig. 3.4), or they show strong emission lines with a width of several thousand km/s if the line width is interpreted as due to Doppler broadening, i.e., $\Delta\lambda/\lambda = \Delta v/c$. In many cases, by far the largest fraction of luminosity is produced in a very small central region: the active galactic nucleus (AGN) that gave this class of galaxies its name. In quasars, the central luminosity can be of the order of $\sim 10^{13} L_\odot$, about a thousand times as luminous as the total luminosity of our Milky Way. We will discuss active galaxies, their phenomena, and their physical properties in detail in Chap. 5.

Another type of galaxy also has spectral properties that differ significantly from those of 'normal' galaxies, namely the starburst galaxies. Normal spiral galaxies like our Milky Way form new stars at a star-formation rate of $\sim 3 M_\odot/\mathrm{yr}$ which can be derived, for instance, from the Balmer lines of hydrogen generated in the HII regions around young, hot stars. By contrast, elliptical galaxies show only marginal star formation or none at all. However, there are galaxies which have a much higher star-formation rate, reaching values of $100 M_\odot/\mathrm{yr}$ and more. If many young stars are formed we would expect these starburst galaxies to radiate strongly in the blue or in the UV part of the spectrum, corresponding to the maximum of the Planck function for the most massive and most luminous stars (see Fig. 3.5). This expectation is not fully met though: star formation takes place in the interior of dense molecular clouds which often also contain large amounts of dust. If the major part of star formation is hidden from our direct view by layers of absorbing dust, these galaxies will not be very prominent in blue light. However, the strong radiation from the young, luminous stars heats the dust; the absorbed stellar light is then re-emitted in the form of thermal dust emission in the infrared region of the electromagnetic spectrum—these galaxies can thus be extremely luminous in the IR. They are called ultra-luminous infrared galaxies (ULIRGs). We will describe the phenomena of starburst galaxies in more detail in Sect. 9.3.1. Of special interest is the discovery that the star-formation rate of galaxies seems to be closely related to interactions between galaxies—many ULIRGs are strongly interacting or merging galaxies (see Fig. 3.6).

3.1.3 The bimodal color distribution of galaxies

The classification of galaxies by morphology, given by the Hubble classification scheme (Fig. 3.2), has the disadvantage

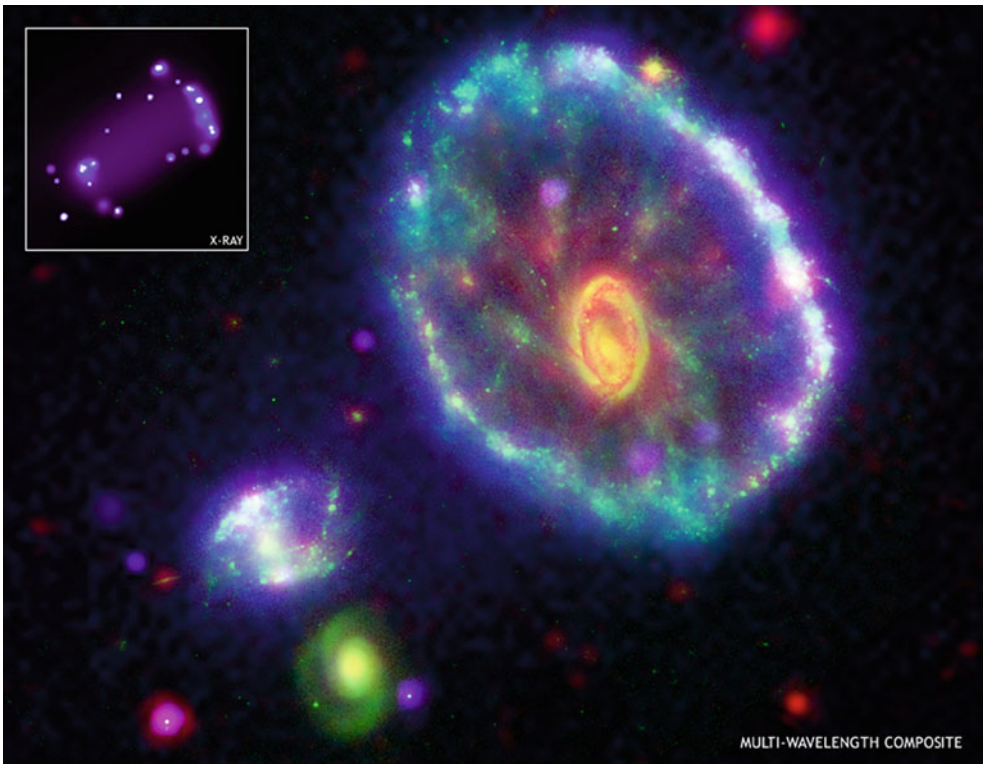


Fig. 3.5 The Cartwheel galaxy is shown as a color composite, based on data with four different telescopes: the *green color* shows the optical light as seen with the HST, the *red color* shows the infrared emission as seen by Spitzer, in *blue* the ultraviolet emission as seen by GALEX is displayed, and *purple* shows the X-ray light observed with Chandra. This galaxy has a very unusual morphology, which is due to a collision with one of the smaller galaxies seen towards the lower left, some 200 million years ago. Before the collision it was probably a normal spiral galaxy, but the collision created a shock wave which swept up gas and formed a large ring, in which very active star formation started to occur. This intense star formation is seen clearly through its UV and X-ray emission, and some of the star-forming regions are also very luminous in the infrared. Many of the massive stars formed in this starburst exploded as supernovae, leaving behind neutrons stars and probably black holes. If these compact objects have a companion star, they can accrete matter and can become powerful X-ray sources, like the X-ray binaries seen in the Milky Way. We will see later that the triggering of starbursts through galaxy collisions is a very common phenomenon. Credit: Composite: NASA/JPL/Caltech/P.Appleton et al. X-ray: NASA/CXC/A. Wolter & G. Trinchieri et al.

that morphologies of galaxies are not easy to quantify. Traditionally, this was done by visual inspection but of course this method bears some subjectivity of the researcher doing it and requires a lot of experience. Furthermore, this visual inspection is time consuming and cannot be performed on large samples of galaxies.[2] Various techniques and related software were developed to perform such a classification automatically, in many cases with significant success, including the reproducibility of galaxy classification between different methods. Nevertheless, quite a number of problems remain, such as the inclination dependence of the morphological appearance of a galaxy.

Even automatic classifications cannot be applied to galaxies for which the angular resolution of the imaging is not much better than the angular size of galaxies, that is, for distant objects. An alternative to morphological classification is provided by the colors of galaxies, which can be obtained from broad-band multi-color imaging. Colors are much easier to measure than morphology, in particular for very small galaxies. In addition, the physical properties of galaxies may be better characterized by their colors than by their morphology—the colors yield information about the stellar population, whereas the morphology is determined by the dynamics of the galaxy.

Using photometric measurements and spectroscopy from the Sloan Digital Sky Survey (see Sect. 8.1.2), the colors and absolute magnitudes of low-redshift galaxies have been studied; their density distribution in a color-magnitude diagram is plotted in the left-hand side of Fig. 3.7. We see immediately that there are two density peaks of the galaxy distribution

[2]The morphological classification recently had a revival, with the Galaxy Zoo project. Its goal was to obtain the morphological classification of millions of galaxies from the Sloan Digital Sky Survey (SDSS), carried out by the general public—i.e., everyone interested could participate. Every galaxy was seen and classified by many different participants, so that misclassifications of individuals get corrected 'democratically'. In its first year of existence, more than 150 000 people participated, yielding the classification of $\sim 50 \times 10^6$ galaxies. A large number of publications are based on the results of the Galaxy Zoo project.

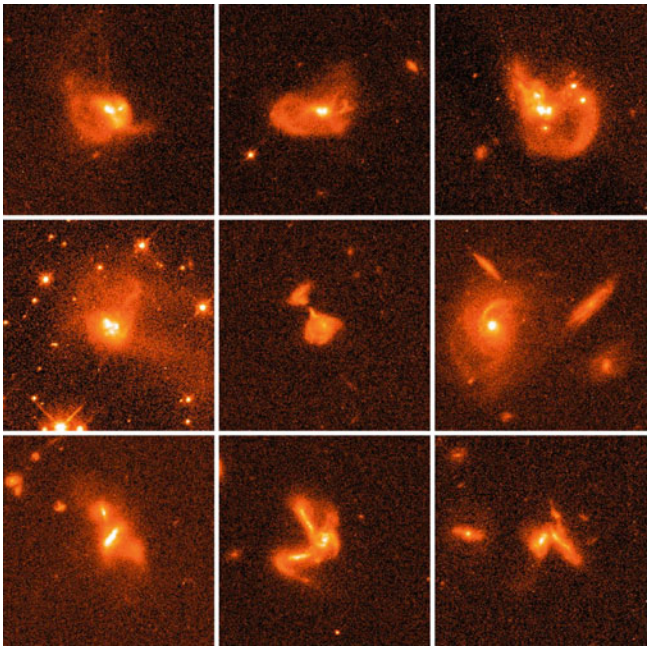


Fig. 3.6 This mosaic of 9 HST images shows different ULIRGs in collisional interaction between two or more galaxies. Credit: NASA, STScI, K. Borne, L. Colina, H. Bushouse & R. Lucas

in this diagram: one at high luminosities and red color, the other at significantly fainter absolute magnitudes and much bluer color. It appears that the galaxies are distributed at and around these two density peaks, hence galaxies tend to be either luminous and red, or less luminous and blue. We can also easily see from this diagram that the distribution of red and blue galaxies with respect to their luminosity is different, the former one being more shifted towards larger luminosity.

We can next consider the color distribution of galaxies at a fixed absolute magnitude M_r. This is obtained by plotting the galaxy number density along vertical cuts through the left-hand side of Fig. 3.7. When this is done for different M_r, it turns out that the color distribution of galaxies is bimodal: over a broad range in absolute magnitude, the color distribution has two peaks, one at red, the other at blue $u - r$. Again, this fact can be seen directly from Fig. 3.7. For each value of M_r, the color distribution of galaxies can be very well fitted by the sum of two Gaussian functions. The central colors of the two Gaussians are shown by the two dashed curves in the left panel of Fig. 3.7. They become redder the more luminous the galaxies are. This luminosity-dependent reddening is considerably more pronounced for the blue population than for the red galaxies.

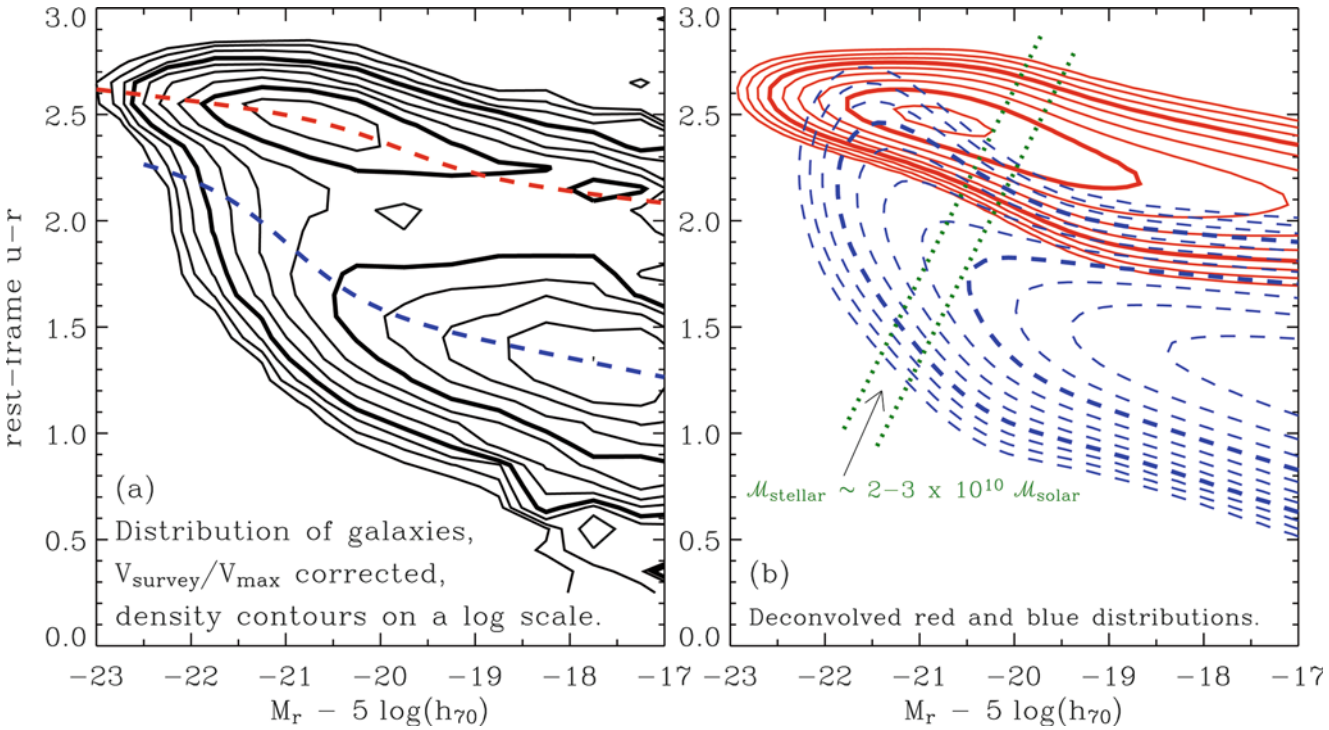


Fig. 3.7 The density of galaxies in color-magnitude space. The color of $\sim$ 70 000 galaxies with redshifts $0.01 \leq z \leq 0.08$ from the Sloan Digital Sky Survey is measured by the rest-frame $u - r$, i.e., after a (small) correction for their redshift was applied. The density contours, which were corrected for selection effects, are logarithmically spaced, with a factor of $\sqrt{2}$ between consecutive contours. (**a**) The measured distribution is shown. Obviously, two peaks of the galaxy density are clearly visible, one at a red color of $u - r \sim 2.5$ and an absolute magnitude of $M_r \sim -21$, the other at a bluer color of $u - r \sim 1.3$ and significantly fainter magnitudes. (**b**) Corresponds to the modeled galaxy density, as is described in the text. Reused with permission from I.K. Baldry, M.L. Balogh, R. Bower, K. Glazebrook & R.C. Nichols 2004, *Color bimodality: Implications for galaxy evolution*, in: THE NEW COSMOLOGY: Conference on Strings and Cosmology, R. Allen (ed.), Conference Proceeding 743, p. 106, Fig. 1 (2004). ©2004, American Institute of Physics

To see how good this fit indeed is, the right-hand side of Fig. 3.7 shows the galaxy density as obtained from the two-Gaussian fits, with solid contours corresponding to the red galaxies and dashed contours to the blue ones. We thus conclude that the local galaxy population can be described as a bimodal distribution in $u-r$ color, where the characteristic color depends slightly on absolute magnitude. The galaxy distribution at bright absolute magnitudes is dominated by red galaxies, whereas for less luminous galaxies the blue population dominates.

The mass-to-light ratio of a red stellar population is larger than that of a blue population, since the former no longer contains massive luminous stars. The difference in the peak absolute magnitude between the red and blue galaxies therefore corresponds to an even larger difference in the stellar mass of these two populations. Red galaxies in the local Universe have on average a much higher stellar mass than blue galaxies. This fact is illustrated by the two dotted lines in the right-hand panel of Fig. 3.7 which correspond to lines of constant stellar mass of ~ 2–$3\times10^{10}M_{\odot}$. This seems to indicate a very characteristic mass scale for the galaxy distribution: most galaxies with a stellar mass larger than this characteristic mass scale are red, whereas most of those with a lower stellar mass are blue.

Obviously, these statistical properties of the galaxy distribution must have an explanation in terms of the evolution of galaxies; we will come back to this issue in Chap. 9. Furthermore, in Sect. 3.6 we will relate the morphological classification to that in color-magnitude space. But first we will describe the properties of elliptical and spiral galaxies in more detail in the next two sections.

3.2 Elliptical Galaxies

3.2.1 Classification

The general term 'elliptical galaxies' (or ellipticals, for short) covers a broad class of galaxies which differ in their luminosities and sizes—some of them are displayed in Fig. 3.8. A rough subdivision is as follows:

- *Normal ellipticals.* This class includes giant ellipticals (gE's), those of intermediate luminosity (E's), and compact ellipticals (cE's), covering a range in absolute magnitudes from $M_B \sim -23$ to $M_B \sim -15$.
- *Dwarf ellipticals* (dE's). These differ from the cE's in that they have a significantly smaller surface brightness and a lower metallicity.
- *cD galaxies*. These are extremely luminous (up to $M_B \sim -25$) and large (up to $R \lesssim 1$ Mpc) galaxies that are only found near the centers of dense clusters of galaxies. Their surface brightness is very high close to the center, they have an extended diffuse envelope, and they have a very high M/L ratio. As we will discuss in Sect. 6.3.4, it is not clear whether the extended envelope actually 'belongs' to the galaxy or is part of the galaxy cluster in which the cD is embedded, since such clusters are now known to have a population of stars located outside of the cluster galaxies.
- *Blue compact dwarf galaxies.* These 'blue compact dwarfs' (BCD's) are clearly bluer (with $\langle B-V\rangle$ between 0.0 and 0.3) than the other ellipticals, and contain an appreciable amount of gas in comparison.
- *Dwarf spheroidals* (dSph's) exhibit a very low luminosity and surface brightness. They have been observed down to $M_B \sim -8$. Due to these properties, they have thus far only been observed in the Local Group.

Thus elliptical galaxies span an enormous range (more than 10^6) in luminosity and mass, as is shown by the compilation in Table 3.1.

3.2.2 Brightness profile

The brightness profiles of normal E's and cD's follow approximately a de Vaucouleurs profile [see (2.40) or (2.42), respectively] over a wide range in radius, as is illustrated in Fig. 3.9. The effective radius $R_{\rm e}$ is strongly correlated with the absolute magnitude M_B, as can be seen in Fig. 3.10, with rather little scatter. In comparison, the dE's and the dSph's clearly follow a different distribution. Owing to the relation (2.43) between luminosity, effective radius and central surface brightness, an analogous relation exists for the average surface brightness $\mu_{\rm ave}$ (unit: B-mag/arcsec^2) within $R_{\rm e}$ as a function of M_B. In particular, the surface brightness in normal E's decreases with increasing luminosity, while it increases for dE's and dSph's.

Yet another way of expressing this correlation is by eliminating the absolute luminosity, obtaining a relation between effective radius $R_{\rm e}$ and surface brightness $\mu_{\rm avg}$. This form is then called the *Kormendy relation.*

The de Vaucouleurs profile provides good fits for normal E's, whereas for E's with exceptionally high (or low) luminosity the profile decreases more slowly (or rapidly) for larger radii. The profile of cD's extends much farther out and is not properly described by a de Vaucouleurs profile (Fig. 3.11), except in it innermost part. It appears that cD's are similar to E's but embedded in a very extended, luminous halo. Since cD's are only found in the centers of massive clusters of galaxies, a connection must exist between this morphology and the environment of these galaxies; we shall return to this topic in Sect. 6.3.4. In contrast to these classes of ellipticals, diffuse dE's are often better described by an exponential profile. In fact, the large recent surveys allowed a much better characterization of the brightness profiles of ellipticals and variations amongst them, as will be discussed in Sect. 3.6.

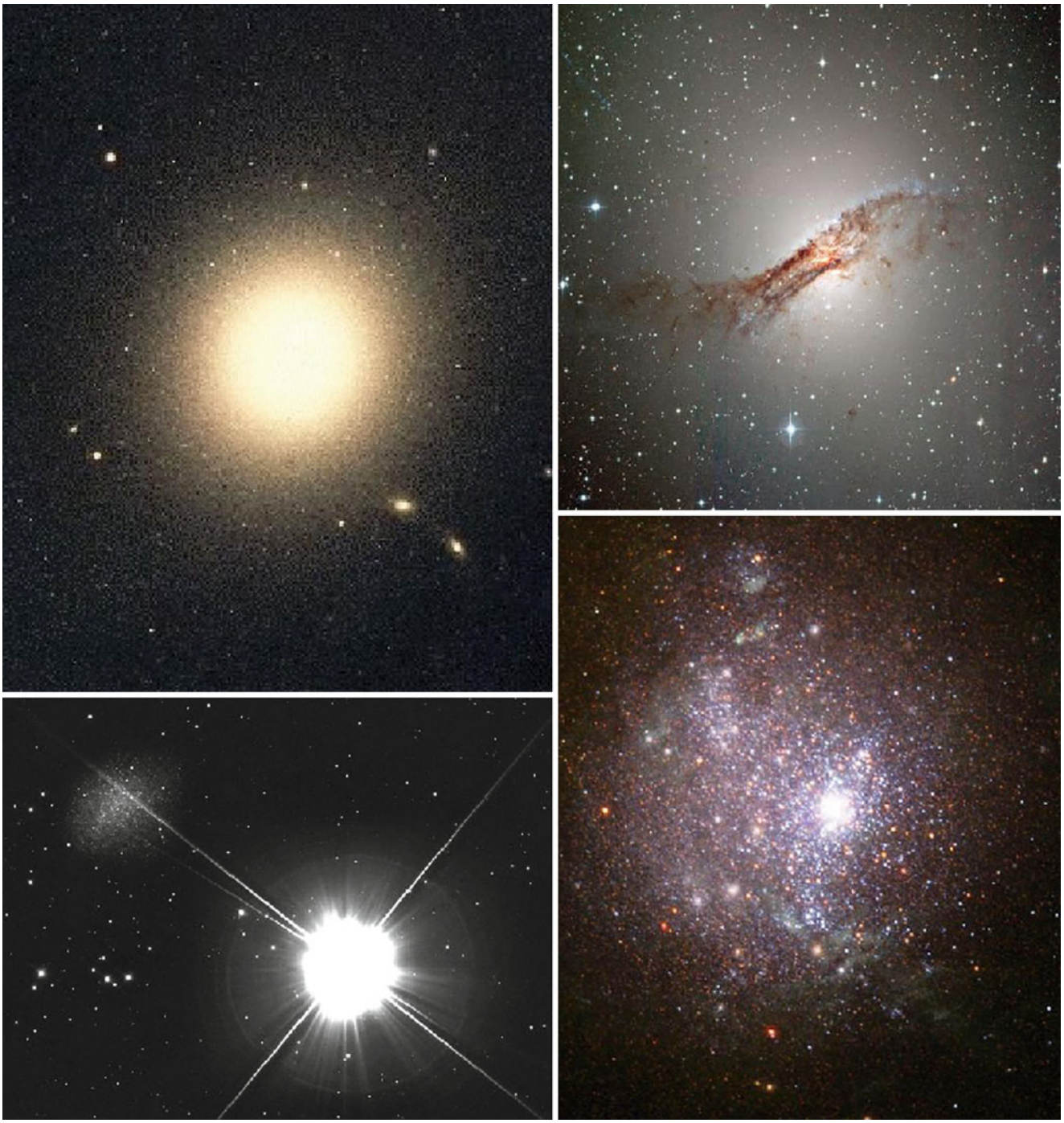


Fig. 3.8 Different types of elliptical galaxies. *Upper left*: the cD galaxy M87 in the center of the Virgo galaxy cluster; *upper right*: Centaurus A, a giant elliptical galaxy with a very distinct dust disk and an active galactic nucleus; *lower left*: the galaxy Leo I (located near the upper left corner of the image) belongs to the nine known *dwarf spheroidals* in the Local Group; *lower right*: NGC 1705, a dwarf irregular, shows indications of massive star formation—a super star cluster and strong galactic winds. Credit: *Top left*: Digital Sky Survey, ESO. *Top right*: ESO. *Bottom left*: Michael Breite, www.skyphoto.de. *Bottom right*: NASA, ESA and The Hubble Heritage Team (STScI/AURA); acknowledgement: M. Tosi (INAF, Osservatorio Astronomico di Bologna)

Cores and extra light. As indicated in Fig. 3.9, the brightness profile can differ significantly from a de Vaucouleurs profile in the very central part; in the example shown, the central brightness profile lies well below the $r^{1/4}$ fit. In this case, the central brightness profile is said to have a core, or a light deficit (relative to the extrapolation of the de Vaucouleurs profile towards the center). Ellipticals with a core are typically very luminous (and correspondingly

Table 3.1 Characteristic values for early-type galaxies

	S0	cD	E	dE	dSph	BCD
M_B	-17 to -22	-22 to -25	-15 to -23	-13 to -19	-8 to -15	-14 to -17
M ($M_\odot$)	10^{10}–10^{12}	10^{13}–10^{14}	10^8–10^{13}	10^7–10^9	10^7–10^8	$\sim 10^9$
D_{25} (kpc)	10–100	300–1000	1–200	1–10	0.1–0.5	<3
$\langle M/L_B \rangle$	~ 10	>100	10–100	1–10	5–100	0.1–10
$\langle S_N \rangle$	~ 5	~ 15	~ 5	4.8 ± 1.0	–	–

D_{25} denotes the diameter at which the surface brightness has decreased to 25 B-mag/arcsec^2, S_N is the 'specific frequency', a measure for the number of globular clusters in relation to the visual luminosity [see (3.18)], and M/L is the mass-to-light ratio in Solar units (the values of this table are taken from the book by Carroll & Ostlie)

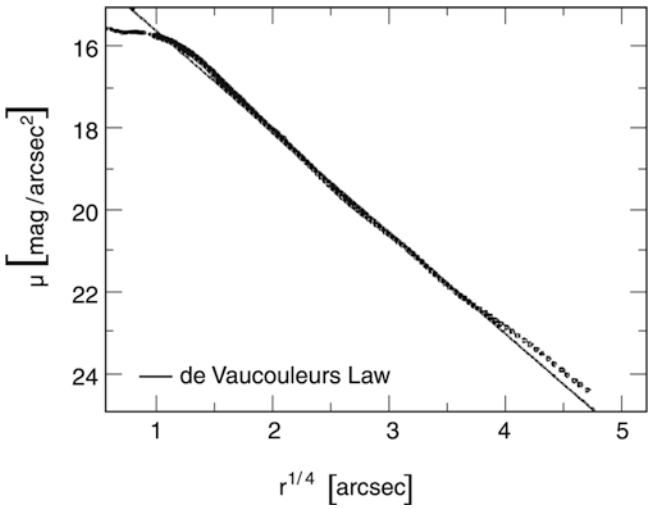


Fig. 3.9 Surface brightness profile of the galaxy NGC 4472, fitted by a de Vaucouleurs profile (*solid curve*). The de Vaucouleurs profile describes a linear relation between the logarithm of the intensity (i.e., linear on a magnitude scale) and $r^{1/4}$; for this reason, it is also called an $r^{1/4}$-law

have a high stellar mass). Also the opposite effect occurs: some ellipticals, typically those of lower luminosity, have an excess of light in their cores, relative to the extrapolation of the brightness profile fit at larger radii.

3.2.3 Composition of elliptical galaxies

Except for the BCD's, elliptical galaxies appear red when observed in the optical, which suggests an old stellar population. It was once believed that ellipticals contain neither gas nor dust, but these components have now been found, though at a much lower mass-fraction than in spirals. For example, in some ellipticals hot gas ($\sim 10^7$K) is detected by its X-ray emission. Furthermore, Hα emission lines of warm gas ($\sim 10^4$ K) are observed, as well as cold gas (~ 100 K) in the HI (21 cm) and CO molecular lines. Many (up to 75 % of the population) of the normal ellipticals contain visible amounts of dust, partially manifested as a dust disk (see the upper right panel of Fig. 3.8 for a prominent example). They frequently show extended HI disks, up to ~ 200 kpc in diameter. However, the estimated mass of the cold atomic and molecular gas is typically less than 1 % of the stellar mass in ellipticals. This amount of gas is actually smaller than expected from the gas release from the stellar population due to its evolution, e.g., in the form of stellar winds, planetary nebulae etc. The fate of the bulk of this gas is currently unclear.

Whereas the largest fraction of the stellar population in ellipticals is old—as we will see soon, most of the stars in present-day ellipticals must have formed some 10 billion years ago—there are spectroscopic indications for a low level of recent star formation. This has been further supported from UV-observations carried out by the GALEX satellite which showed that $\sim$ 15 % of galaxies which are red in optical colors (and thus located in the upper peak of the color-magnitude diagram in Fig. 3.7) show a strong UV-excess. Spatially resolved imaging of early-type galaxies with an UV-excess at wavelengths of $\lambda \sim 1500$ Å showed that about 75 % of them do indeed clearly display star-formation activity (see Fig. 3.12 for four examples). The UV emission is more extended that the optical image of the galaxies, i.e., the star formation occurs in the outer parts of the galaxies. The star-formation rate is sufficiently low ($\sim 0.5 M_\odot$/yr) as to have a vanishing effect on the optical colors of these galaxies. The large radii at which the UV emission is seen suggests that the stars do not form from gas which has been located in the gravitational potential of the galaxy; instead, it is likely that it is gas infalling from the surrounding medium. We have seen direct evidence for such infalling gas in the Milky Way (see Sect. 2.3.7), and in Chap. 10 we will explain that such gas infall is indeed expected in our model of structure formation in the Universe.

The metallicity of ellipticals and S0 galaxies increases towards the galaxy center, as derived from color gradients. Also in S0 galaxies the bulge appears redder than the disk. The Spitzer Space Telescope, launched in 2003, detected a spatially extended distribution of warm dust in S0 galaxies, organized in some sort of spiral structure. Cold dust was also found in ellipticals and S0 galaxies.

This composition of ellipticals clearly differs from that of spiral galaxies and needs to be explained by models of the

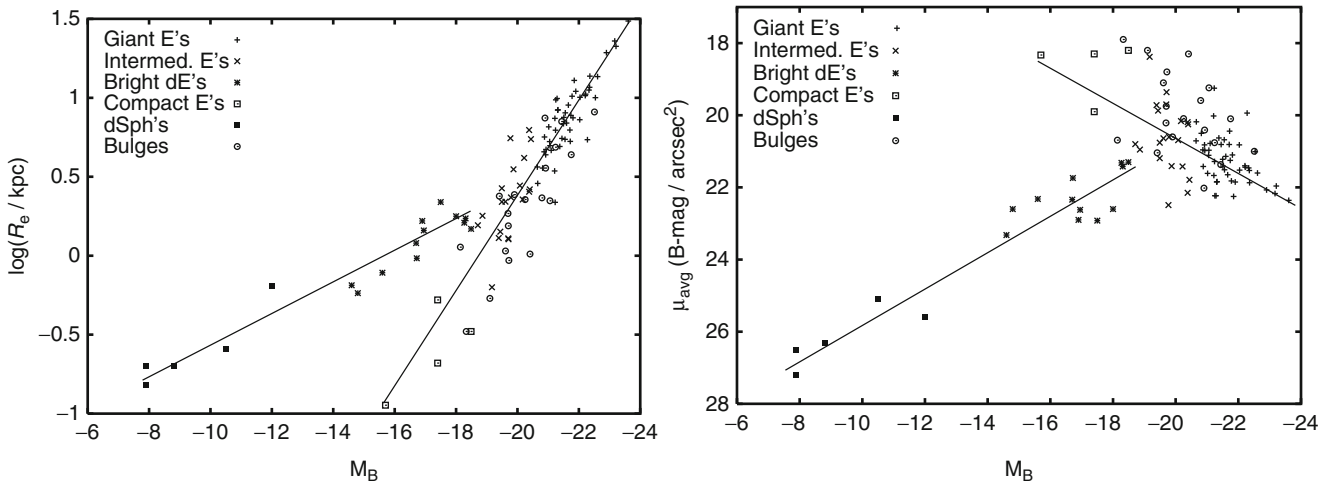


Fig. 3.10 *Left panel*: effective radius R_e versus absolute magnitude M_B; the correlation for normal ellipticals is different from that of dwarfs. *Right panel*: average surface brightness μ_{avg} versus M_B; with increasing luminosity, the surface brightness of normal ellipticals decreases, while for dwarf ellipticals and spheroidals it increases. Source: R. Bender et al. 1992, *Dynamically hot galaxies. I - Structural properties*, ApJ 399, 462. ©AAS. Reproduced with permission

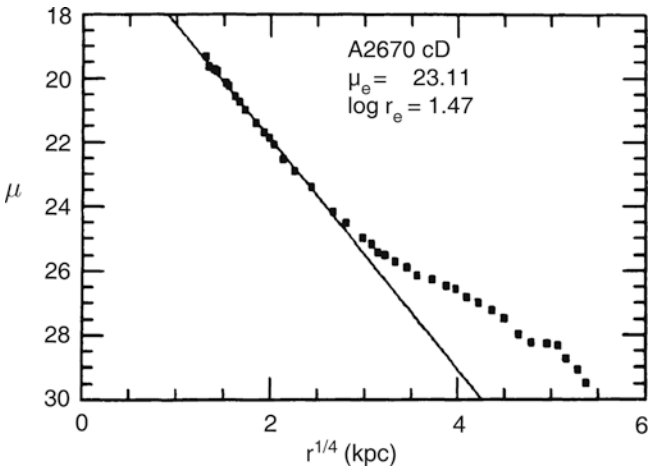


Fig. 3.11 Comparison of the brightness profile of a cD galaxy, the central galaxy of the cluster of galaxies Abell 2670, with a de Vaucouleurs profile. The light excess for large radii is clearly visible. Source: J.M. Schombert 1986, *The structure of brightest cluster members. I - Surface photometry*, ApJS 60, 603, p. 618, Fig. 1. ©AAS. Reproduced with permission

formation and evolution of galaxies. We will see later that the cosmic evolution of elliptical galaxies is also observed to be different from that of spirals.

3.2.4 Dynamics of elliptical galaxies

Analyzing the morphology of elliptical galaxies raises a simple question: *Why are ellipticals not round?* A simple explanation would be rotational flattening, i.e., as in a rotating self-gravitating gas ball, the stellar distribution bulges outwards at the equator due to centrifugal forces, as is also the case for the Earth. If this explanation was correct, the rotational velocity v_{rot}, which is measurable in the relative Doppler shift of absorption lines, would have to be of about the same magnitude as the velocity dispersion of the stars σ_v that is measurable through the Doppler broadening of lines. More precisely, by means of stellar dynamics one can show that for the rotational flattening of an axially symmetric, oblate[3] galaxy, the relation

$$\left(\frac{v_{rot}}{\sigma_v}\right)_{iso} \approx \sqrt{\frac{\epsilon}{1-\epsilon}} \tag{3.1}$$

has to be satisfied, where 'iso' indicates the assumption of an isotropic velocity distribution of the stars. However, for very luminous ellipticals one finds that, in general, $v_{rot} \ll \sigma_v$, so that rotation cannot be the major cause of their ellipticity (see Fig. 3.13). In addition, many ellipticals are presumably triaxial, so that no unambiguous rotation axis is defined. Thus, luminous ellipticals are in general *not* rotationally flattened. For less luminous ellipticals and for the bulges of disk galaxies, however, rotational flattening can play an important role. The question remains of how to explain a stable elliptical distribution of stars without bulk rotation.

The brightness profile of an elliptical galaxy is defined by the distribution of its stellar orbits. Let us assume that the gravitational potential is given. The stars are then placed into this potential, with the initial positions and velocities following a specified distribution. If this distribution is not

[3] If $a \geq b \geq c$ denote the lengths of the major axes of an ellipsoid, then it is called an oblate spheroid (= rotational ellipsoid) if $a = b > c$, whereas a prolate spheroid is specified by $a > b = c$. If all three axes are different, it is called triaxial ellipsoid.

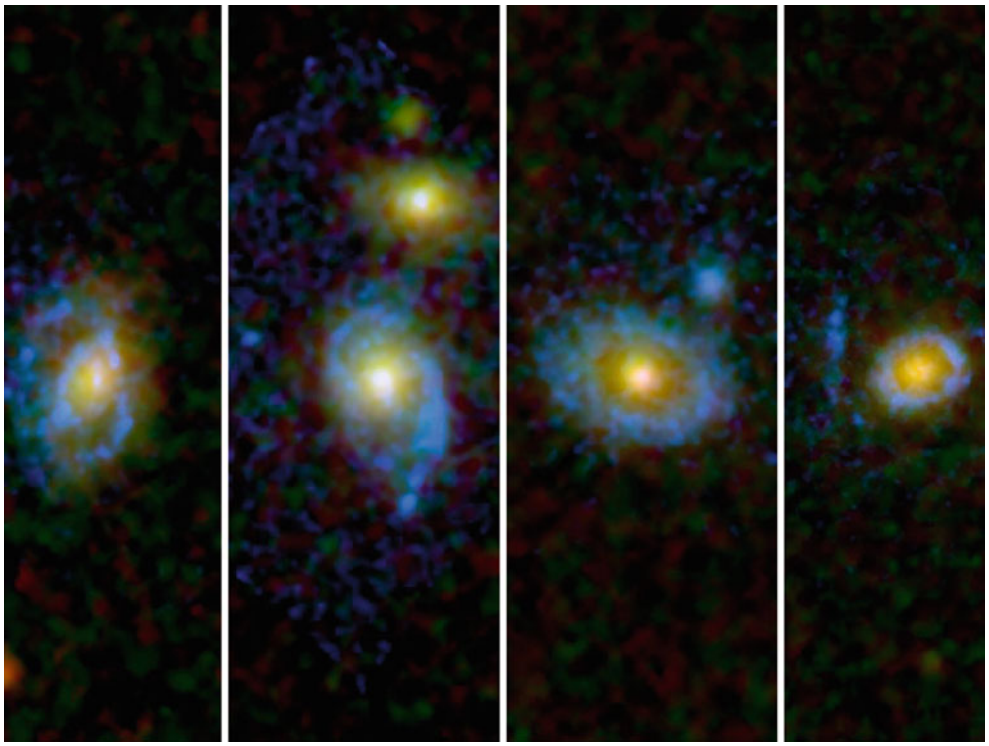


Fig. 3.12 Four early-type galaxies, observed at UV and optical wavelengths with HST. The optical emission is shown in *green*, *yellow* and *red*, whereas the UV emission is shown in *blue*. Credit: NASA/ESA/JPL-Caltech/STScI/UCLA

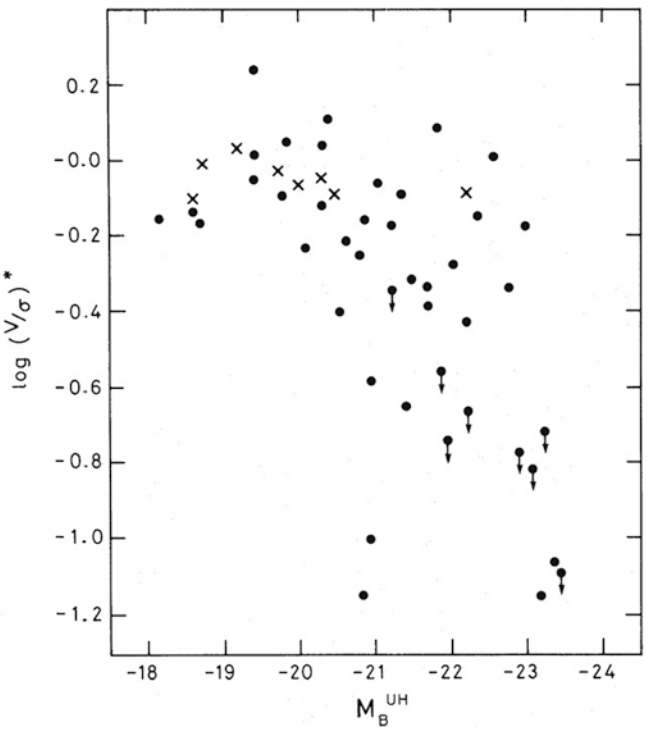


Fig. 3.13 The rotation parameter $\left(\frac{v_{\rm rot}}{\sigma_v}\right) / \left(\frac{v_{\rm rot}}{\sigma_v}\right)_{\rm iso}$ of elliptical galaxies, here denoted by $(V/\sigma)^*$, plotted as a function of absolute magnitude. *Dots* denote elliptical galaxies, *crosses* the bulges of disk galaxies; *arrows* indicate that the corresponding dot is an upper limit on the rotation parameter. One sees that the luminous ellipticals rotate far too slow to explain their ellipticity as being due to rotational flattening, whereas lower-luminosity objects can be rotationally flattened. Source: R.L. Davies et al. 1983, *The kinematic properties of faint elliptical galaxies*, ApJ 266, 41, p. 49, Fig. 4. ©AAS. Reproduced with permission

isotropic in velocity space, the resulting light distribution will in general not be spherical. For instance, one could imagine that the orbital planes of the stars have a preferred direction, but that an equal number of stars exists with positive and negative angular momentum L_z, so that the total stellar distribution has no net angular momentum and therefore does not rotate. Each star moves along its orbit in the gravitational potential, where the orbits are in general not closed. If an initial distribution of stellar orbits is chosen such that the statistical properties of the distribution of the orbits are invariant in time, then one will obtain a stationary system. If, in addition, the distribution is chosen such that the respective mass distribution of the stars will generate exactly the originally chosen gravitational potential, one arrives at a self-gravitating equilibrium system. In general, it is a difficult mathematical problem to construct such self-gravitating equilibrium systems. Furthermore, as we will see, elliptical galaxies also contain a dark matter component, whose gravitational potential adds to that of the stars.

Relaxation time-scale. The question now arises whether such an equilibrium system can also be stable in time. One might expect that close encounters of pairs of stars would cause a noticeable disturbance in the distribution of orbits. These pair-wise collisions could then lead to a 'thermalization' of the stellar orbits.[4] To examine this question we need

[4]Note that in a gas like air, scattering between molecules occurs frequently, which drives the velocity distribution of the molecules towards an isotropic Maxwellian, i.e., the thermal distribution.

to estimate the time-scale for such collisions and the changes in direction they cause.

For this purpose, we consider the relaxation time-scale by pair collisions in a system of N stars of mass m, total mass $M = Nm$, extent R, and a mean stellar density of $n = 3N/(4\pi R^3)$. We define the relaxation time $t_{\rm relax}$ as the characteristic time in which a star changes its velocity direction by $\sim 90°$ due to pair collisions with other stars. By simple calculation (see below), we find that

$$t_{\rm relax} \approx \frac{R}{6v}\frac{N}{\ln(N/2)} , \tag{3.2}$$

or

$$\boxed{t_{\rm relax} = \frac{t_{\rm cross}}{6}\frac{N}{\ln(N/2)}} , \tag{3.3}$$

where $t_{\rm cross} = R/v$ is the crossing time-scale, i.e. the time it takes a star to cross the stellar system. If we now consider a typical galaxy, with $t_{\rm cross} \sim 10^8\,$yr, $N \sim 10^{12}$ [thus $\ln(N/2) \sim 30$], then we find that the relaxation time is much longer than the age of the Universe. This means that *pair collisions do not play any role in the evolution of stellar orbits.* The dynamics of the orbits are determined solely by the large-scale gravitational field of the galaxy. In Sect. 7.5.1, we will describe a process called violent relaxation which most likely plays a central role in the formation of galaxies and which is probably also responsible for the stellar orbits establishing an equilibrium configuration.

We thus conclude that the stars behave like a collisionless gas: elliptical galaxies are stabilized by (dynamical) pressure, and they are elliptical because the stellar distribution is anisotropic in velocity space. This corresponds to an anisotropic pressure—where we recall that the pressure of a gas is nothing but the momentum transport of gas particles due to their thermal motion.

Derivation of the collisional relaxation time-scale. We consider a star passing by another one, with the impact parameter b being the minimum distance between the two. From gravitational deflection, the star attains a velocity component perpendicular to the incoming direction of

$$\boxed{v_\perp^{(1)} \approx a\,\Delta t \approx \left(\frac{Gm}{b^2}\right)\left(\frac{2b}{v}\right) = \frac{2Gm}{bv}} , \tag{3.4}$$

where a is the acceleration at closest separation and Δt the 'duration of the collision', estimated as $\Delta t = 2b/v$ (see Fig. 3.14). Equation (3.4) can be derived more rigorously by integrating the perpendicular acceleration along the orbit. A star undergoes many collisions, through which the perpendicular velocity components will accumulate; these form two-dimensional vectors perpendicular to the original direction. After a time t we have $\boldsymbol{v}_\perp(t) = \sum_i \boldsymbol{v}_\perp^{(i)}$. The expectation value of this vector

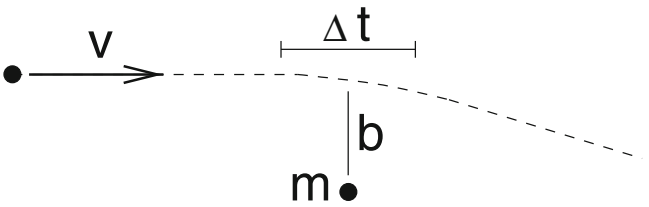


Fig. 3.14 Sketch related to the derivation of the dynamical time-scale

is $\langle \boldsymbol{v}_\perp(t)\rangle = \sum_i \left\langle \boldsymbol{v}_\perp^{(i)}\right\rangle = 0$ since the directions of the individual $\boldsymbol{v}_\perp^{(i)}$ are random. But the mean square velocity perpendicular to the incoming direction does not vanish,

$$\left\langle |\boldsymbol{v}_\perp|^2\,(t)\right\rangle = \sum_{ij}\left\langle \boldsymbol{v}_\perp^{(i)}\cdot\boldsymbol{v}_\perp^{(j)}\right\rangle = \sum_i \left\langle \left|\boldsymbol{v}_\perp^{(i)}\right|^2\right\rangle \neq 0 , \tag{3.5}$$

where we set $\left\langle \boldsymbol{v}_\perp^{(i)}\cdot\boldsymbol{v}_\perp^{(j)}\right\rangle = 0$ for $i \neq j$ because the directions of different collisions are assumed to be uncorrelated. The velocity $\boldsymbol{v}_\perp$ performs a so-called *random walk*. To compute the sum, we convert it into an integral where we have to integrate over all collision parameters b. During time t, all collision partners with impact parameters within db of b are located in a cylindrical shell of volume $(2\pi b\,{\rm d}b)\,(vt)$, so that

$$\begin{aligned}\left\langle |\boldsymbol{v}_\perp|^2\,(t)\right\rangle &= \int 2\pi\, b\,{\rm d}b\, v\, t\, n\, \left|v_\perp^{(1)}\right|^2 \\ &= 2\pi\left(\frac{2Gm}{v}\right)^2 v\,t\,n\int\frac{{\rm d}b}{b} .\end{aligned} \tag{3.6}$$

The integral cannot be performed from 0 to ∞. Thus, it has to be cut off at $b_{\rm min}$ and $b_{\rm max}$ and then yields $\ln(b_{\rm max}/b_{\rm min})$. Due to the finite size of the stellar distribution, $b_{\rm max} = R$ is a natural choice. Furthermore, our approximation which led to (3.4) will certainly break down if $v_\perp^{(1)}$ is of the same order of magnitude as v; hence we choose $b_{\rm min} = 2Gm/v^2$. With this, we obtain $b_{\rm max}/b_{\rm min} = Rv^2/(2Gm)$. The exact choice of the integration limits is not important, since $b_{\rm min}$ and $b_{\rm max}$ appear only logarithmically. Next, using the virial theorem, $|E_{\rm pot}| = 2E_{\rm kin}$, and thus $GM/R = v^2$ for a typical star, we get $b_{\rm max}/b_{\rm min} \approx N/2$. Thus,

$$\boxed{\left\langle |\boldsymbol{v}_\perp|^2\,(t)\right\rangle = 2\pi\left(\frac{2Gm}{v}\right)^2 v\,t\,n\,\ln(N/2)} . \tag{3.7}$$

We define the relaxation time $t_{\rm relax}$ by $\left\langle |\boldsymbol{v}_\perp|^2\,(t_{\rm relax})\right\rangle = v^2$, i.e., the time after which the perpendicular velocity roughly equals the infall velocity:

$$\begin{aligned}t_{\rm relax} &= \frac{1}{2\pi n v}\left(\frac{v^2}{2Gm}\right)^2\frac{1}{\ln(N/2)} \\ &= \frac{1}{2\pi n v}\left(\frac{M}{2Rm}\right)^2\frac{1}{\ln(N/2)} = \frac{R}{6v}\frac{N}{\ln(N/2)} ,\end{aligned} \tag{3.8}$$

from which we finally obtain (3.3).

The Jeans equation. The behavior of stars in an elliptical galaxy is thus that of collisionless particles in a gravitational potential. The equation governing the density of stars as a function of position, velocity, and time, i.e., the phase-space density $f(\boldsymbol{r},\boldsymbol{v},t)$, is the collisionless Boltzmann equation. Without going into any detail, we shall quote one special result from the Boltzmann equation, which applies to the simplest case: Consider a spherically symmetric gravitational potential

$\Phi(r)$, in which stars are orbiting. We assume that the system is stationary, so that the phase-space density f does not depend on time. Furthermore, the system is assumed to have no net rotation. The stellar distribution is assumed to be spherically symmetric as well, and the velocity distribution in the plane perpendicular to the radius vector should be isotropic. In spherical coordinates, this means that the velocity dispersion in the θ-direction is the same as that in the φ-direction, $\langle v_\theta^2 \rangle = \langle v_\varphi^2 \rangle$. However, the velocity dispersion $\langle v_r^2 \rangle$ in the radial direction is allowed to be different from that in the tangential direction. We quantify the anisotropy of the velocity distribution by the parameter

$$\beta = 1 - \frac{\langle v_\theta^2 \rangle}{\langle v_r^2 \rangle} \,. \tag{3.9}$$

For example, if all stars would be on circular orbits, then $\langle v_r^2 \rangle = 0$, corresponding to $\beta = -\infty$. Conversely, in the case that all stars are on radial orbits, one has $\beta = 1$. If the velocity distribution is isotropic, then $\beta = 0$. From the collisionless Boltzmann equation, one obtains the Jeans equation

$$\frac{1}{n} \frac{\mathrm{d}\left(n \langle v_r^2 \rangle\right)}{\mathrm{d}r} + 2\frac{\beta \langle v_r^2 \rangle}{r} = -\frac{\mathrm{d}\Phi}{\mathrm{d}r} \,, \tag{3.10}$$

relating the local volume density of particles

$$n(r) = \int \mathrm{d}^3 v \; f(x, \boldsymbol{v}) \,,$$

and the velocity distribution characterized by $\langle v_r^2 \rangle (r)$ and $\beta(r)$ to the gravitational potential $\Phi(r)$.

Suppose we can measure the density of stars $n(r)$, by mapping the surface brightness of an elliptical galaxy, assuming a mean stellar luminosity (which then yields the column density of stars, i.e., the projected stellar number density), and calculating $n(r)$ from the projected density; for spherically symmetric distributions, these two are uniquely related to each other. Furthermore, suppose we obtain the line-of-sight velocity dispersion as a function of projected radius from spectroscopically determining the width of stellar absorption lines. This measured line-of-sight velocity dispersion depends on $n(r)$, $\langle v_r^2 \rangle (r)$ and the anisotropy parameter $\beta(r)$. With $n(r)$ determined from the projected number density, the observed velocity dispersion then depends on the two functions $\langle v_r^2 \rangle (r)$ and $\beta(r)$. Thus, the latter two cannot be determined separately from measurements of the observed line-of-sight velocity dispersion. This has an immediate consequence for the determination of the mass profile: let $\Phi(r) = -GM(r)/r$, then the mass profile of the galaxy is described by

$$v_\mathrm{c}^2(r) \equiv \frac{GM(r)}{r} = -\langle v_r^2 \rangle \left(\frac{\mathrm{d}\ln n}{\mathrm{d}\ln r} + \frac{\mathrm{d}\ln \langle v_r^2 \rangle}{\mathrm{d}\ln r} + 2\beta \right) \,, \tag{3.11}$$

where $v_\mathrm{c}(r)$ is the velocity that a particle on a circular orbit with radius r has in this potential. Since β and $\langle v_r^2 \rangle$ cannot be determined separately, the mass profile of the galaxy cannot be measured. Or phrased differently, the mass estimate depends on the assumed anisotropy of the stellar orbits, so that mass profile and anisotropy are degenerate.

Therefore, even in the simplest case of maximum symmetry, the determination of the mass profile of elliptical galaxies is problematic. This is the reason why it is much more difficult to make accurate statements about the mass of ellipticals as obtained from stellar kinematics than it is for spiral galaxies, where the rotation curve yields the mass profile directly. Breaking the degeneracy between the radial velocity dispersion and the anisotropy parameter is possible, however, if one studies not only the line width of the stellar absorption lines, but also their shape. This shape depends on higher-order moments of the velocity distribution, and can be used to estimate $\langle v_r^2 \rangle (r)$ and $\beta(r)$ separately.

3.2.5 Indicators of a complex evolution

The isophotes (that is, the curves of constant surface brightness) of many of the normal elliptical galaxies are well approximated by ellipses. These elliptical isophotes with different surface brightnesses are concentric to high accuracy, with the deviation of the isophote's center from the center of the galaxy being typically $\lesssim 1\,\%$ of its extent. However, in many cases the ellipticity varies with radius, so that the value for ϵ is not a constant. In addition, a few percent of ellipticals show a so-called isophote twist: the orientation of the semi-major axis of the isophotes changes with the surface brightness, and thus with radius. This indicates that elliptical galaxies are not spheroidal, but triaxial systems (or that there is some intrinsic twist of their axes).

Although the light distribution of ellipticals appears rather simple at first glance, a more thorough analysis reveals that the kinematics can be quite complicated. For example, dust disks are not necessarily perpendicular to any of the principal axes, and the dust disk may rotate in a direction opposite to the galactic rotation. In addition, ellipticals may also contain (weak) stellar disks.

Boxiness and diskiness. The so-called boxiness parameter describes the deviation of the isophotes' shape from that of an ellipse. Consider the shape of an isophote. If it is described by an ellipse, then after a suitable choice of the coordinate system, $\theta_1 = a\cos t$, $\theta_2 = b\sin t$, where a and b are the two semi-axes of the ellipse and $t \in [0, 2\pi]$ parametrizes the curve. The distance $r(t)$ of a point from the center is

$$r(t) = \sqrt{\theta_1^2 + \theta_2^2} = \sqrt{\frac{a^2 + b^2}{2} + \frac{a^2 - b^2}{2} \cos(2t)} \,.$$

Deviations of the isophote shape from this ellipse are now expanded in a Taylor series, where the term $\propto \cos(4t)$ describes the lowest-order correction that preserves the symmetry of the ellipse with respect to reflection in the two coordinate axes. The modified curve is then described by

$$\boldsymbol{\theta}(t) = \left(1 + \frac{a_4 \cos(4t)}{r(t)}\right) \begin{pmatrix} a\cos t \\ b\sin t \end{pmatrix} \,, \tag{3.12}$$

with $r(t)$ as defined above. The parameter a_4 thus describes a deviation from an ellipse: if $a_4 > 0$, the isophote appears more disk-like, and if $a_4 < 0$, it becomes rather boxy (see Fig. 3.15). In most elliptical galaxies we typically find

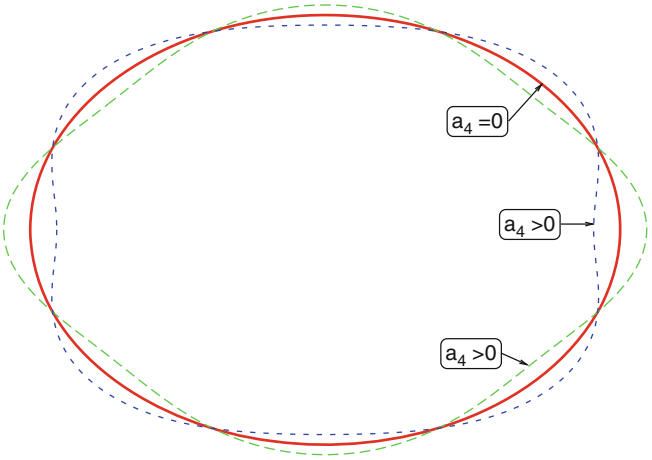


Fig. 3.15 Sketch to illustrate boxiness and diskiness. The *solid red curve* shows an ellipse ($a_4 = 0$), the *green dashed curve* a disky ellipse ($a_4 > 0$), and the *blue dotted curve* a boxy ellipse ($a_4 < 0$). In most elliptical galaxies, the deviations in the shape of the isophotes from an ellipse are considerably smaller than in this sketch

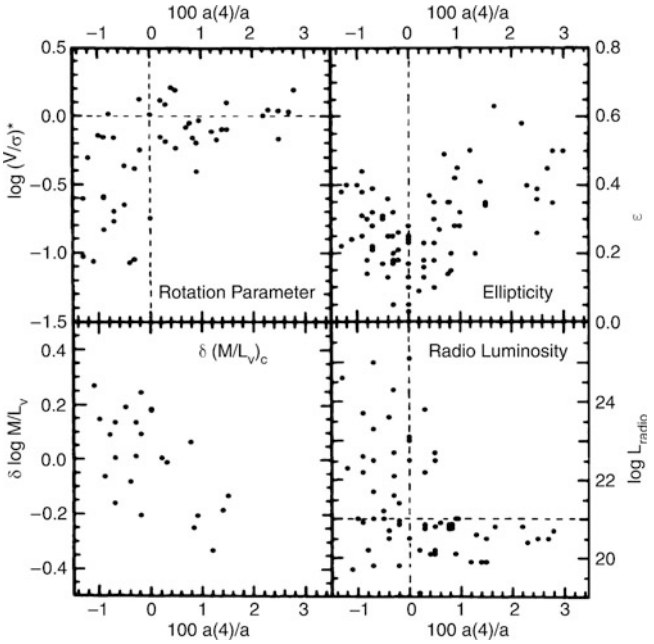


Fig. 3.16 Correlations of a_4/a with some other properties of elliptical galaxies. $100a(4)/a$ (corresponding to a_4/a) describes the deviation of the isophote shape from an ellipse in percent. Negative values denote boxy ellipticals, positive values disky ellipticals. The *upper left panel* shows the rotation parameter discussed in Sect. 3.2.4; at the *lower left*, the deviation from the average mass-to-light ratio is shown. The *upper right panel* shows the ellipticity, and the *lower right panel* displays the radio luminosity at 1.4 GHz. Obviously, there is a correlation of all these parameters with the boxiness parameter. Source: J. Kormendy & S. Djorgovski 1989, *Surface photometry and the structure of elliptical galaxies*, ARA&A 27, 235, Fig. 3, p. 259. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 27 ©1989 by Annual Reviews www.annualreviews.org

$|a_4/a| \sim 0.01$, thus only a small deviation from the elliptical form.

Correlations of a_4 with other properties of ellipticals. At first sight, such apparently small deviations from an exact elliptical shape of isophotes seems to be of little importance. Surprisingly however, we find that the parameter a_4/a is strongly correlated with other properties of ellipticals (see Fig. 3.16). The ratio $\left(\frac{v_{\rm rot}}{\sigma_v}\right) / \left(\frac{v_{\rm rot}}{\sigma_v}\right)_{\rm iso}$ (upper left in Fig. 3.16) is of order unity for disky ellipses ($a_4 > 0$) and, in general, significantly smaller than one for boxy ellipticals. From this we conclude that 'diskies' are in part rotationally supported, whereas the flattening of 'boxies' is mainly caused by the anisotropic distribution of their stellar orbits in velocity space. The mass-to-light ratio is also correlated with a_4: boxies (diskies) have a value of M/L in their core which is larger (smaller) than the mean elliptical of comparable luminosity. A very strong correlation exists between a_4/a and the radio luminosity of ellipticals: while diskies are weak radio emitters, boxies show a broad distribution in $L_{\rm radio}$. These correlations are also seen in the X-ray luminosity, since diskies are weak X-ray emitters and boxies have a broad distribution in $L_{\rm X}$. This bimodality becomes even more obvious if the radiation contributed by compact sources (e.g., X-ray binary stars) is subtracted from the total X-ray luminosity, thus considering only the diffuse X-ray emission. Ellipticals with a different sign of a_4 also differ in the kinematics of their stars: boxies often have cores spinning against the general direction of rotation (counter-rotating cores), which is rarely observed in diskies.

About 70 % of the ellipticals are diskies. The transition between diskies and S0 galaxies may be continuous along a sequence of varying disk-to-bulge ratio.

Shells and ripples. In many of the early-type galaxies that are not member galaxies of a cluster, sharp discontinuities in the surface brightness are found, a kind of shell structure ('shells' or 'ripples'). They are visible as elliptical arcs curving around the center of the galaxy (see Fig. 3.17). Such sharp edges can only be formed if the corresponding distribution of stars is 'cold', i.e., they must have a very small velocity dispersion, since otherwise such coherent structures would smear out on a very short time-scale. As a comparison, we can consider disk galaxies that likewise contain sharp structures, namely the thin stellar disk. Indeed, the stars in the disk have a very small velocity dispersion, $\sim$ 20 km/s, compared to the rotational velocity of typically 200 km/s. Probably a better example are the stellar streams discovered in the halo of the Milky Way (Sect. 2.3.6) which remain coherent through orbiting the Galaxy only

Fig. 3.17 As a result of galaxy collisions and mergers, coherent stellar structures are formed, as seen in the galaxy NGC 474, one of the most spectacular examples found so far. The galaxy has multiple luminous shells and a complex structure of tidal tails, witnesses of its past violent history. Credit: Jean-Charles Cuillandre (CFHT) & Giovanni Anselmi (Coelum)/Canada-France-Hawaii Telescope/Coelum

because the velocity dispersion of stars belonging to these streams is much smaller than the characteristic rotational velocity.

These peculiarities in ellipticals are not uncommon. Indicators for shells and other tidal features can be found in about 70 % of the early-type galaxies, and about a third of them show boxy isophotes.

Boxiness, counter-rotating cores, and shells and ripples are all indicators of a complex evolution that is probably caused by past interactions and mergers with other galaxies.

We will proceed with a discussion of this interpretation in Chap. 10.

3.3 Spiral galaxies

3.3.1 Trends in the sequence of spirals

Looking at the sequence of early-type spirals (i.e., Sa's or SBa's) to late-type spirals, we find a number of differences that can be used for classification (see Fig. 3.18):

- a decreasing luminosity ratio of bulge and disk, with $L_{\rm bulge}/L_{\rm disk} \sim 0.3$ for Sa's and ~ 0.05 for Sc's,

Fig. 3.18 Types of spiral galaxies. *Top left*: M94, an Sab galaxy. *Top middle*: M51, an Sbc galaxy. *Top right*: M101, an Sc galaxy. *Lower left*: M83, an SBa galaxy. *Lower middle*: NGC 1365, an SBb galaxy. *Lower right*: M58, an SBc galaxy. Credit: *Top left*: Jacobus Kapteyn Telescope, ING Archive and Nik Szymanek. *Top middle*: William Herschel Telescope, ING Archive, Javier Méndez and Nik Szymanek. *Top right*: INT, Peter Bunclark and Nik Szymanek. *Lower left*: European Southern Observatory. *Lower middle*: European Southern Observatory. *Lower right*: JKT, Johan Knapen and Nik Szymanek

- an increasing opening angle of the spiral arms, from $\sim 6°$ for Sa's to $\sim 18°$ for Sc's,
- and an increasing brightness structure along the spiral arms: Sa's have a 'smooth' distribution of stars along the spiral arms, whereas the light distribution in the spiral arms of Sc's is resolved into bright knots of stars and HII regions.

Compared to ellipticals, the spirals cover a distinctly smaller range in absolute magnitude (and mass). They are limited to $-16 \gtrsim M_B \gtrsim -23$ and $10^9 M_\odot \lesssim M \lesssim 10^{12} M_\odot$, respectively. Characteristic parameters of the various types of spirals are compiled in Table 3.2.

Bars are common in spiral galaxies, with $\sim 70\,\%$ of all disk galaxies containing a large-scale stellar bar. Such a bar perturbs the axial symmetry of the gravitational potential in a galaxy, which may have a number of consequences. One of them is that this perturbation can lead to a redistribution of angular momentum of the stars, gas, and dark matter. In addition, by perturbing the orbits, gas can be driven towards the center of the galaxy which may have important consequences for triggering nuclear activity and enhanced star formation (see Chap. 5).

3.3.2 Brightness profile

The light profile of the bulge of spirals is described by a de Vaucouleurs profile to a first approximation—see (2.40) and (2.42)—while the disk follows an exponential brightness profile, as is the case for our Milky Way. Expressing these distributions of the surface brightness in $\mu \propto -2.5 \log(I)$, measured in mag/arcsec^2, we obtain

$$\boxed{\mu_{\rm bulge}(R) = \mu_{\rm e} + 8.3268 \left[\left(\frac{R}{R_{\rm e}} \right)^{1/4} - 1 \right]} \qquad (3.13)$$

Table 3.2 Characteristic values for spiral galaxies

	Sa	Sb	Sc	Sd/Sm	Im/Ir
M_B	-17 to -23	-17 to -23	-16 to -22	-15 to -20	-13 to -18
M ($M_\odot$)	10^9–10^{12}	10^9–10^{12}	10^9–10^{12}	10^8–10^{10}	10^8–10^{10}
$\langle L_{bulge}/L_{tot}\rangle_B$	0.3	0.13	0.05	–	–
Diam. (D_{25},kpc)	5–100	5–100	5–100	0.5–50	0.5–50
$\langle M/L_B\rangle$ ($M_\odot/L_\odot$)	6.2 ± 0.6	4.5 ± 0.4	2.6 ± 0.2	~ 1	~ 1
V_{max} range(km s^{-1})	163–367	144–330	99–304	–	50–70
Opening angle	$\sim 6°$	$\sim 12°$	$\sim 18°$	–	–
$\mu_{0,B}$ (mag arcsec^{-2})	21.52 ± 0.39	21.52 ± 0.39	21.52 ± 0.39	22.61 ± 0.47	22.61 ± 0.47
$\langle B-V\rangle$	0.75	0.64	0.52	0.47	0.37
$\langle M_{gas}/M_{tot}\rangle$	0.04	0.08	0.16	0.25 (Scd)	–
$\langle M_{H_2}/M_{HI}\rangle$	2.2 ± 0.6 (Sab)	1.8 ± 0.3	0.73 ± 0.13	0.19 ± 0.10	–
$\langle S_N\rangle$	1.2 ± 0.2	1.2 ± 0.2	0.5 ± 0.2	0.5 ± 0.2	–

V_{max} is the maximum rotation velocity, thus characterizing the flat part of the rotation curve. The opening angle is the angle under which the spiral arms branch off, i.e., the angle between the tangent to the spiral arms and the circle around the center of the galaxy running through this tangential point. S_N is the specific abundance of globular clusters as defined in (3.18). The values in this table are taken from the book by B.W. Carroll & D.A. Ostlie 1996, *Introduction to Modern Astrophysics*, Addison Wesley

and

$$\boxed{\mu_{disk}(R) = \mu_0 + 1.09\left(\frac{R}{h_R}\right)}\,. \tag{3.14}$$

Here, μ_e is the surface brightness at the effective radius R_e which is defined such that half of the bulge luminosity is emitted within R_e [see (2.41)]. The central surface brightness and the scale-length of the disk are denoted by μ_0 and h_R, respectively. It has to be noted that μ_0 is not directly measurable since μ_0 is *not* the central surface brightness of the galaxy, only that of its disk component. To determine μ_0, the exponential law (3.14) is extrapolated from large R inwards to $R = 0$, or more precisely, by fitting the sum of an exponential and a bulge component to the total light profile of the galaxy.

The brightness profile of spiral disks perpendicular to the disk can be studied exclusively in edge-on spirals. From them, one finds that it is in general well described by an exponential law of the form (2.32) or, equivalently, of the form (2.36). The scale-height h_z of the disk is almost independent of the galacto-centric radius R, and between galaxies scales roughly linearly with the rotational velocity of the disk. The typical value for the ratio of scale-height to scale-length is $h_z/h_R \sim 0.07$—indeed, the disks of spiral galaxies are thin. The flattest galaxies are those of late Hubble type.

Bulges and pseudobulges. As mentioned before, the brightness profile of bulges follows approximately that of a de Vaucouleurs profile. However, in some spiral galaxies bulges were found which behave differently than these 'classical' bulges; one calls them *pseudobulges*. In contrast to classical bulges, they follow more an exponential profile, are typically flatter, and have significant rotational support. Furthermore, whereas classical bulges lie on the same sequence in the effective radius vs. absolute magnitude diagram as the ellipticals (see Fig. 3.10), pseudobulges do not. They have lower luminosity for a given size.

In many cases it is very difficult to distinguish between both types of bulges photometrically. However, spectroscopy aids a lot in this distinction. In fact, some bulges of spirals have two components, i.e., both a classical bulge and a pseudobulge.

The differences in the two types of bulges suggest that they should have a different origin. Classical bulges behave like a small elliptical galaxy. As we will discuss in detail later, it is believed that ellipticals form through merging events of galaxies which 'heats up' the stellar velocity distribution, i.e., turns ordered velocity fields of disk galaxies into random orbits which are characteristic for ellipticals. Therefore, in the current model of galaxy evolution, classical bulges are also formed as a result of merger events. In contrast, the ordered rotation of pseudobulges suggests that they have evolved from the disk population. For example, symmetry perturbations of the gravitational field caused by a bar can generate random velocity components of stars perpendicular to the plane of the disk, and thus thicken the disk population in the inner part of a galaxy.

Whereas pseudobulges may provide important insights into the evolution of galaxies, they are a sub-dominant component in the population of galaxies. It is estimated that classical bulges together contain about ten times more stars than pseudobulges. Therefore, whenever we use the term 'bulge' in the following, we implicitly mean the classical bulges.

Some late-type spiral galaxies seem to have no bulge component. Some of them show instead a nuclear stellar cluster at their center. These nuclear star clusters appear

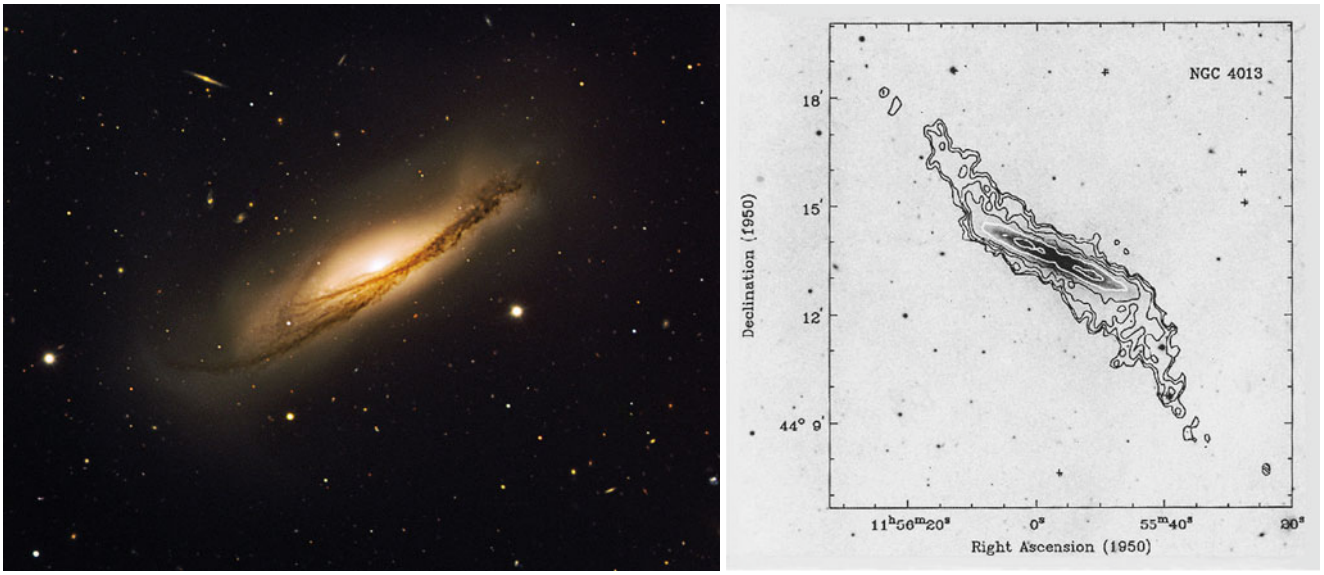


Fig. 3.19 *Left panel*: This 6′.2 × 5′ VLT-image of the edge-on spiral NGC 3190 shows an example of a warped disk galaxy. NGC 3190 is the dominant member of a compact group of galaxies (HCG 44), where the other members (one elliptical and two additional spirals) are outside the field-of-view of this image. The galaxy also contains an AGN in its center. *Right panel*: The edge-on spiral galaxy NGC 4013, with the optical emission shown in *grey*, superimposed by the intensity of the 21 cm-emitting gas. The warping of the gas disk starts at about the radius where the disk becomes invisible in optical light. Credit: *Left*: European Southern Observatory/H. Boffin. *Right*: R. Bottema 1995, *The prodigious warp of NGC 4013. II. Detailed observations of the neutral hydrogen gas*, A&A 295, 605, p. 609, Fig. 4. ©ESO. Reproduced with permission

at first sight to be similar to globular clusters. However, their stellar population is quite different from that of the old globular clusters in our Galaxy, as their light is dominated by a relatively young stellar population, although their stellar mass is totally dominated by an old population. In some respect, these nuclear star clusters share properties with the peculiar Galactic globular ω Centauri, which also shows a broad range of stellar ages and an inhomogeneous chemical abundance. Therefore, it has been hypothesized that ω Centaurus is the remnant of a merger of a lower mass galaxy with the Milky Way.

Freeman's law. When Ken Freeman analyzed a sample of spiral galaxies, he found the remarkable result that the central surface brightness μ_0 of disks has a very low spread, i.e., it is very similar for different galaxies (*Freeman's law, 1970*). For Sa's to Sc's, a value of $\mu_0 = 21.52 \pm 0.39$ B-mag/arcsec^2 is observed, and for Sd spirals and later types, $\mu_0 = 22.61 \pm 0.47$ B-mag/arcsec^2. This result was critically discussed, for example with regard to its possible dependence on selection effects. Their importance is not implausible since the determination of precise photometry of galaxies is definitely a lot easier for objects with a high surface brightness. After accounting for such selection effects in the statistical analysis of galaxy samples, Freeman's law was confirmed for 'normal' spiral galaxies.

However, galaxies exist which have a significantly lower surface brightness, the *low surface brightness galaxies (LSBs)*. They seem to form a separate class of (disk) galaxies whose central surface brightness is often two or more magnitudes fainter than the canonical value given by Freeman's law, and thus much lower than the brightness of the night sky, so that searching for these LSBs is problematic and requires very accurate data reduction and subtraction of the sky background. These LSB galaxies seem to transform their gas much more slowly into stars than normal spirals; indeed, combining UV- and IR-data from GALEX and Spitzer reveal that LSBs show little extinction, i.e., a very low dust fraction and little molecular gas.

Warps in disks. The disks of galaxies are not always lying in a plane—disks can be warped. In this case, the plane in which the orbit of stars and gas rotate around the galactic center at a given radius R changes its inclination with R. The warping can sometimes be observed from the distribution of stars (Fig. 3.19), but more frequently from the (more extended) distribution of neutral hydrogen gas and the velocity field as measured from its 21 cm-emission. The latter is of course also altered by the change of the orientation of the orbital planes. The origin of warps in galaxies is not well understood. One possibility would be that they are generated by interactions with other galaxies which seriously perturb the orbits of stars and gas. Indeed, the galaxy shown in the left panel of Fig. 3.19 is the dominant member of a compact galaxy group, and thus subject to tidal forces from the other group members and the group as a whole. However,

this is a rather extreme case. In most cases, warps start at radii beyond the optical radius of a galaxy and thus are visible only in the distribution and motion of gas; the right-hand panel of Fig. 3.19 shows an example of this kind. Indeed, the majority of galaxies with warps in their outer gas disks seem to have no significant companion.

Stellar halo. Whereas the bulge and the disk can be studied in spirals even at fairly large distances, the stellar halo has too low a surface brightness to be seen in distant galaxies. However, our neighboring galaxy M31, the Andromeda galaxy, can be studied in quite some detail. In particular, the brightness profile of its stellar halo can be studied more easily than that of the Milky Way, taking advantage of our 'outside' view. This galaxy should be quite similar to our Galaxy in many respects. A stellar halo of red giant branch (RGB) stars was detected in M31, which extends out to more than 150 kpc from its center. The brightness profile of this stellar distribution indicates that for radii $r \lesssim 20\,\mathrm{kpc}$ it follows the extrapolation from the brightness profile of the bulge. However, for larger radii it exceeds this extrapolation, showing a power-law profile which corresponds to a radial density profile of approximately $\rho \propto r^{-3}$, similar to that observed in our Milky Way. Furthermore, stellar streams from disrupted galaxies were also clearly detected in M31, as in the Galaxy. It thus seems that stellar halos form a generic property of spirals. Unfortunately, the corresponding surface brightness is so small that there is little hope of detecting such a halo in other spirals for which individual stars can no longer be resolved and classified.

The thick disk in other spirals can only be studied if they are oriented edge-on. In these cases, a thick disk can indeed be observed as a stellar population outside the plane of the disk and well beyond the scale-height of the thin disk. As is the case for the Milky Way, the scale-height of a stellar population increases with its age, increasing from young main-sequence stars to old asymptotic giant branch (AGB) stars. For luminous disk galaxies, the thick disk does not contribute substantially to the total luminosity; however, in lower-mass disk galaxies with rotational velocities $\lesssim 120\,\mathrm{km/s}$, the thick disk stars can contribute nearly half the luminosity and may actually dominate the stellar mass. In this case, the dominant stellar population of these galaxies is old, despite the fact that they appear blue.

Sizes of disks. Typically, the optical radius of a spiral galaxy extends out to about four exponential scale lengths, after which the surface brightness, and thus the stellar surface density, displays a break. The characteristic surface brightness at which this occurs is $\mu_\mathrm{B} \approx 25.5\,\mathrm{mag\,arcsec^{-2}}$. Although there are many exceptions to this behavior, it still prevails in the majority of spirals. In contrast to the stellar distribution, neutral gas is observed (due to its 21 cm emission of neutral hydrogen) to considerably larger radius, typically a factor of two beyond the break radius.

3.3.3 The Schmidt–Kennicutt law of star formation

If we now take into account that stars form out of gas, and the gas distribution is much more extended than the stellar distribution, then it appears that stars can only form at places in the disk where the gas mass density exceeds a certain value. Indeed, Marteen Schmidt discovered in 1959 a relation between the surface mass density of gas, Σ_gas (measured in units of $M_\odot\,\mathrm{pc^{-2}}$) and the star-formation rate per unit area, Σ_SFR (measured in units of $M_\odot\,\mathrm{yr^{-1}\,kpc^{-2}}$), of the form

$$\Sigma_\mathrm{SFR} \propto \Sigma_\mathrm{gas}^N \,, \tag{3.15}$$

with a power-law index of $N \approx 1.4$. The connection between the two quantities was later examined in detail by Rob Kennicutt, and the relation (3.15) is known as Schmidt–Kennicutt law; including the normalization, one finds

$$\frac{\Sigma_\mathrm{SFR}}{M_\odot\,\mathrm{yr^{-1}\,kpc^{-2}}} = (2.5 \pm 0.7) \times 10^{-4} \left(\frac{\Sigma_\mathrm{gas}}{M_\odot\,\mathrm{pc^{-2}}} \right)^{1.4\pm0.15} . \tag{3.16}$$

Due to the apparent absence of star formation in the outer part of spiral galaxies, one often complements this relation with a cut-off at a specific value for Σ_gas.

In Fig. 3.20, recent results of the star-formation rate in galaxies are compiled, ranging from low-surface brightness galaxies (diamonds) to starburst galaxies (open triangles). Here, the gas surface density is taken to be the sum of atomic and molecular gas, where the density of molecular hydrogen is estimated from the abundance of the CO molecule, assuming a constant conversion factor between these two species. The three diagonal lines indicate the star-formation rate at which the available gas reservoir would be consumed in star formation on a time-scale of 10^8, 10^9, and $10^{10}\,\mathrm{yr}$ (from top to bottom). A global power-law fit to these data would yield a result very similar to (3.15), with an index $N \sim 1.4$. However, the figure suggests that there are different regimes of star formation activity, indicated by the two vertical lines. The regime shown on the right part of the figure is occupied by starburst galaxies. Using only those, a linear relation between Σ_SFR and Σ_gas seems to describe the data quite well. The regime between the two vertical lines is occupied by normal spiral galaxies, and again, restricting a power-law fit solely to them, a linear relation with $N \sim 1$ provides a good approximation. For low values of the gas density (the left

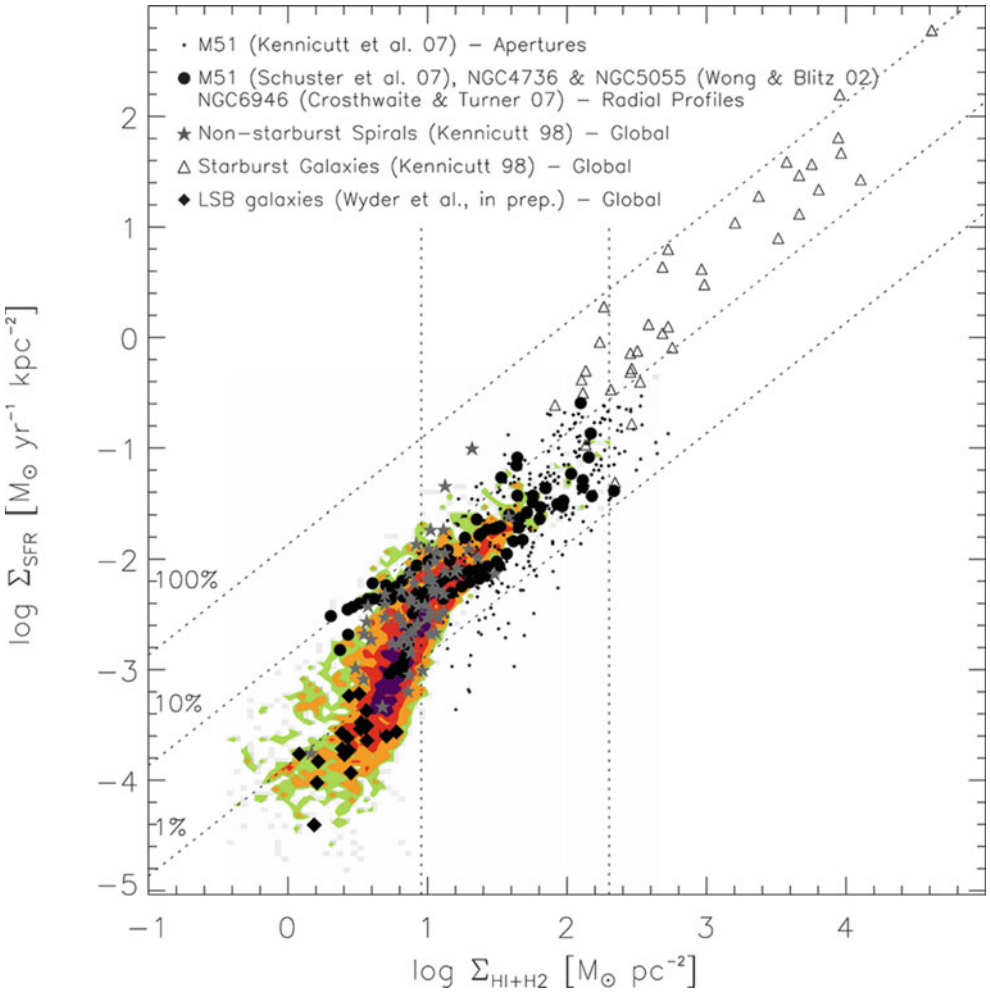


Fig. 3.20 The star formation surface density $\Sigma_{\rm SFR}$ as a function of the surface mass density $\Sigma_{\rm HI+H_2}$ of the sum of atomic and molecular gas. The colored-shading shows results from subregions of nearby spiral and late-type dwarf galaxies. Symbols show measurements from either regions or radial bins (*dots* and *black circles*), or disk-averaged estimates of normal spiral galaxies (*asterisks*). *Open triangles* correspond to starburst galaxies, *diamonds* to low-surface brightness galaxies. The *diagonal lines* indicate a star-formation rate in which 1, 10 or 100% of the gas is consumed in star formation within 10^8 yr. The *two vertical lines* indicate characteristic values of the projected gas density. Source: F. Bigiel et al. 2008, *The Star Formation Law in Nearby Galaxies on Sub-Kpc Scales*, AJ 136, 2846, p. 2869, Fig. 15. ©AAS. Reproduced with permission

part of the figure), the star-formation rate seems to decrease rapidly. Thus, an index $N \approx 1.4$ in the Schmidt–Kennicutt law is obtained as a global fit which makes no distinction between the three different regimes just outlined.

Furthermore, the figure indicates a density threshold in the Schmidt–Kennicutt law. Whereas UV-observations with the GALEX satellite, as well as measurements of Hα emission (which stems from the HII regions around hot stars), showed that star formation can occur well beyond the optical break radius (see Fig. 3.21), it appears that the corresponding level of star formation is rather low, as indicated also in Fig. 3.20.

A better understanding of the origin of the Schmidt–Kennicutt relation is obtained if one considers the dependence of the star-formation rate on the density of atomic and molecular gas separately. This yields the result that $\Sigma_{\rm SFR}$ is essentially proportional to $\Sigma_{\rm H_2}$. This relation is not very surprising, since we know that star formation occurs in molecular clouds and thus one expects that the molecular gas density controls the star-formation rate.

On the other hand, the star-formation rate exhibits a rather steep dependence on the density of atomic gas, which seems to saturate at a value of about $10\,M_\odot\,{\rm pc}^{-2}$, indicated by the left of the vertical lines in Fig. 3.20. But the densities of atomic and molecular gas are not unrelated; molecules form from atoms, and so the dependence of $\Sigma_{\rm SFR}$ on the density of atomic gas could be a secondary effect—$\Sigma_{\rm SFR}$ depends mostly (or even solely) on the molecular density. The Schmidt–Kennicutt relation can then be understood as a combination of the proportionality between $\Sigma_{\rm SFR}$ and $\Sigma_{\rm H_2}$ and the molecular fraction of the gas, which increases with the gas density.

The fact that the starburst galaxies seem to have a higher star-formation rate at a given molecular density may indicate that different physical processes are operating for them; indeed, we have seen before that at least a large fraction of ULIRGs are recent mergers, which are thought to be triggering the starburst.

The efficiency of star formation as a function of the surface mass density of gas provides a possible explanation of the observed break in the optical surface brightness of spiral galaxies. However, the fact that the warps observed in neutral hydrogen typically start to occur at this break radius may also imply that the origin of the break may be due to other effects. For example, the disk inside the break radius may have been assembled rather quickly in the formation history of spirals, whereas the matter lying further out could have been added later on in the evolution, e.g., due to the accretion of mass from the surrounding medium.

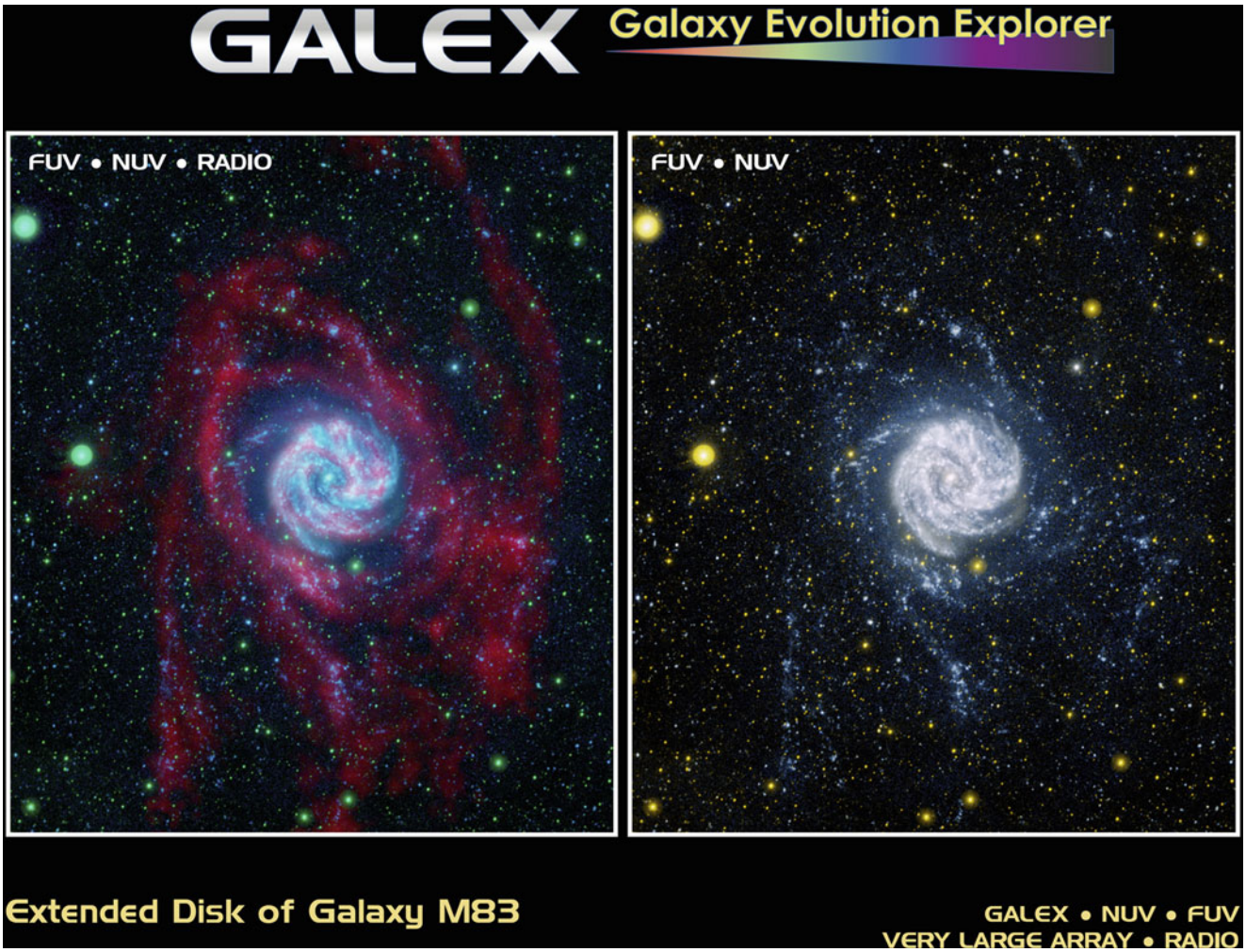


Fig. 3.21 The galaxy M83 (the Pinwheel galaxy) seen by the GALEX satellite in UV-light (*right*), and combined with the 21 cm emission from atomic hydrogen shown in *red* (*left*). The gas disk is far more extended than the stellar disk of this face-on spiral galaxy. However, the UV-image clearly shows that new stars are formed even at $\sim$ 20 kpc from the center of this galaxy, i.e., outside the optical break radius. Credit: NASA/JPL-Caltech/VLA/MPIA

3.3.4 Rotation curves and dark matter

The rotation curves of other spiral galaxies are easier to measure than that of the Milky Way because we are able to observe them 'from outside'. These measurements are achieved by utilizing the Doppler effect, where the inclination of the disk, i.e., its orientation with respect to the line-of-sight, has to be accounted for. The inclination angle is determined from the observed axis ratio of the disk, assuming that disks are intrinsically axially symmetric (except for the spiral arms). Mainly the stars and HI gas in the galaxies are used as luminous tracers, where the observable HI disk is in general significantly more extended than the stellar disk. Therefore, the rotation curves measured from the 21 cm line typically extend to much larger radii than those from optical stellar spectroscopy.

Like our Milky Way, other spirals also rotate considerably faster in their outer regions than one would expect from Kepler's law and the distribution of visible matter (see Fig. 3.22).

> The rotation curves of spirals do not decrease for $R \geq h_R$, as one would expect from the light distribution, but are basically flat. We therefore conclude that spirals are surrounded by a halo of dark matter. The density distribution of this dark halo can be derived from the rotation curves.

To see how the density distribution of the dark matter can be derived from the rotation curves, we employ the force balance between gravitation and centrifugal acceleration, as described by the Kepler rotation law,

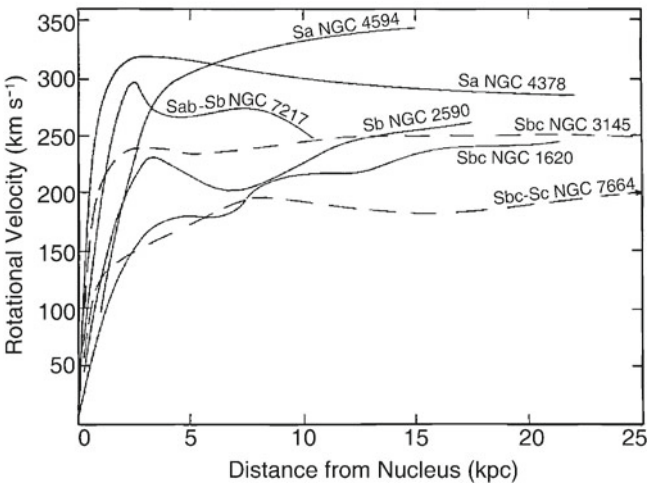


Fig. 3.22 Examples of rotation curves of spiral galaxies. They are all flat in the outer region and do not behave as expected from Kepler's law if the galaxy consisted only of luminous matter. Also striking is the fact that the amplitude of the rotation curve is higher for early-type than for late-type spirals. Source: V. Rubin et al. 1978, *Extended rotation curves of high-luminosity spiral galaxies. IV—Systematic dynamical properties, SA through SC*, ApJ 225, L107, p. L109, Fig. 3. ©AAS. Reproduced with permission

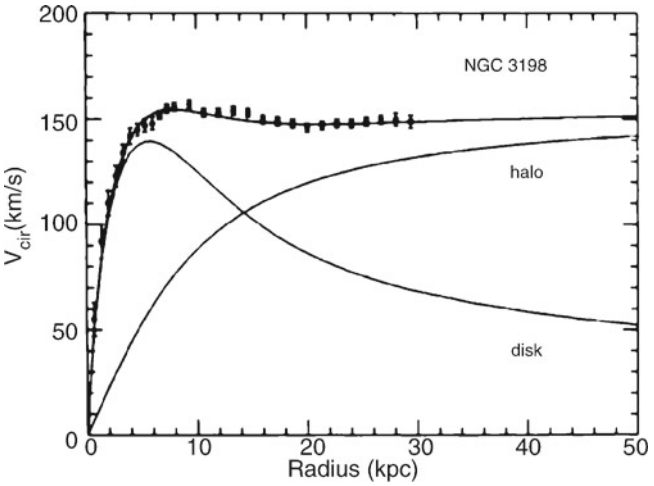


Fig. 3.23 The flat rotation curves of spiral galaxies cannot be explained by visible matter alone. The example of NGC 3198 demonstrates the rotation curve which would be expected from the visible matter alone (curve labeled 'disk'). To explain the observed rotation curve, a dark matter component has to be present (curve labeled 'halo'). However, the decomposition into disk and halo mass is not unambiguous because for it to be so it would be necessary to know the mass-to-light ratio of the disk. In the case considered here, a 'maximum disk' was assumed, i.e., it was assumed that the innermost part of the rotation curve is produced solely by the visible matter in the disk. Source: T.S. van Albada et al. 1985, *Distribution of dark matter in the spiral galaxy NGC 3198*, ApJ 295, 305, p. 309, Fig. 4. ©AAS. Reproduced with permission

$$v^2(R) = GM(R)/R \ ,$$

from which one directly obtains the mass $M(R)$ within a radius R. The rotation curve expected from the visible matter distribution is[5]

$$v^2_{\rm lum}(R) = GM_{\rm lum}(R)/R \ .$$

$M_{\rm lum}(R)$ can be determined by assuming a plausible value for the mass-to-light ratio M/L of the luminous matter. This value is obtained either from the spectral light distribution of the stars, together with knowledge of the properties of stellar populations, or by fitting the innermost part of the rotation curve (where the mass contribution of dark matter can presumably be neglected), assuming that M/L is independent of radius for the stellar population. From this estimate of the mass-to-light ratio, the discrepancy between $v^2_{\rm lum}$ and v^2 yields the distribution of the dark matter, $v^2_{\rm dark} = v^2 - v^2_{\rm lum} = GM_{\rm dark}/R$, or

$$\boxed{M_{\rm dark}(R) = \frac{R}{G}\left[v^2(R) - v^2_{\rm lum}(R)\right]} \ . \qquad (3.17)$$

[5]This consideration is strongly simplified insofar as the given relations are only valid in this form for spherical mass distributions. The rotational velocity produced by an oblate (disk-shaped) mass distribution is more complicated to calculate; for instance, for an exponential mass distribution in a disk, the maximum of $v_{\rm lum}$ occurs at $\sim 2.2h_R$, with a Kepler decrease, $v_{\rm lum} \propto R^{-1/2}$, at larger radii.

An example of this decomposition of the mass contributions is shown in Fig. 3.23.

The decomposition of the rotation curve into a contribution due to the baryonic matter—i.e., mainly stars—and one due to dark matter is not unique, since it depends on the mass-to-light ratio of the stellar population. Assuming M/L to be constant, one can obtain an upper bound on M/L by fitting the innermost part of the rotation curve by the stellar contribution only. This then yields the 'maximum disk' model for rotation curves. The corresponding values obtained for M/L are often larger than those obtained from stellar population models, and thus the maximum disk model is not necessary well justified.

The assumption of an almost constant M/L can be tested in a different way. From stellar dynamics, one can derive a relation between the surface density of the disk, $\Sigma(R)$, the scale-height of the disk, h_z, and the velocity dispersion of stars perpendicular to the disk, $\sigma_z(R)$, which in the case of flat rotation curves reads

$$\sigma_z(R) = \sqrt{c'\pi G\Sigma(R)h_z(R)} \ ,$$

where c' is a constant depending on the vertical profile of the disk; for an exponential profile, $c' = 3/2$. We noted before that the scale-height h_z is observed in edge-on spirals to be independent of R. If M/L is a constant, independent of R, then the observed brightness profile of disks implies that $\Sigma(R) \propto \exp(-R/h_R)$. The foregoing relation then shows that also the velocity dispersion σ_z should follow an exponential in R, but $\propto \exp(-R/2h_R)$, due to the square root. Thus, the scale-length on which the velocity dispersion decreases should be twice the scale-length of the surface brightness. Indeed, spectroscopic studies of face-on spirals have shown this to be the case—a result which is consistent with an (almost) constant M/L of the disk.

The corresponding density profiles of the dark matter halos seem to be flat in the inner region, and decreasing as R^{-2} at large radii. It is remarkable that $\rho \propto R^{-2}$ implies a mass profile $M \propto R$, i.e., the mass of the halo increases linearly with the radius for large R. As long as the extent of the halo is undetermined the total mass of a galaxy will be unknown. Since the observed rotation curves are flat out to the largest radius for which 21 cm emission can still be observed, a lower limit for the radius of the dark halo can be obtained, $R_{\rm halo} \gtrsim 30h^{-1}$kpc. Inside the optical radius of a disk, the dark matter comprises about 2/3 of the total mass.

To derive the density profile out to even larger radii, other observable objects in an orbit around the galaxies are needed. Potential candidates for such luminous tracers are satellite galaxies—companions of other spirals, like the Magellanic Clouds are for the Milky Way. Because we cannot presume that these satellite galaxies move on circular orbits around their parent galaxy, conclusions can be drawn based only on a statistical sample of satellites. These analyses of the relative velocities of satellite galaxies around spirals still give no indication of an 'edge' to the halo, leading to a lower limit for the radius of $R_{\rm halo} \gtrsim 100h^{-1}$kpc.

Correlations of rotation curves with galaxy properties. The form and amplitude of the rotation curves of spirals are correlated with their luminosity and their Hubble type. The larger the luminosity of a spiral, the steeper the rise of $v(R)$ in the central region, and the larger the maximum rotation velocity $v_{\rm max}$. This latter fact indicates that the mass of a galaxy increases with luminosity, as expected. For the characteristic values of the various Hubble types, one finds $v_{\rm max} \sim 300\,{\rm km/s}$ for Sa's, $v_{\rm max} \sim 175\,{\rm km/s}$ for Sc's, whereas Irr's have a much lower $v_{\rm max} \sim 70\,{\rm km/s}$. For equal luminosity, $v_{\rm max}$ is higher for earlier types of spirals. However, the shape (not the amplitude) of the rotation curves of different Hubble types is similar, despite the fact that they have a different brightness profile as seen, for instance, from the varying bulge-to-disk ratio. This point is another indicator that the rotation curves cannot be explained by visible matter alone.

Dark matter in ellipticals. For elliptical galaxies the mass estimate, and thus the detection of a possible dark matter component, is significantly more complicated, since the orbits of stars are substantially more complex than in spirals. In particular, the mass estimate from measuring the stellar velocity dispersion via line widths depends on the anisotropy of the stellar orbits, which is a priori unknown. Nevertheless, in recent years it has been unambiguously proven that dark matter also exists in ellipticals. First, the degeneracy between the anisotropy of the orbits and the mass determination was broken by detailed kinematic analysis. Second, in some ellipticals hot gas was detected from its X-ray emission. As we will see in Sect. 6.4 in the context of clusters of galaxies, the temperature of the gas allows an estimate of the depth of the potential well, and therefore the mass. Both methods reveal that ellipticals are also surrounded by a dark halo.

The gravitational lens effect offers another way to determine the masses of galaxies up to very large radii. With this method we cannot study individual galaxies but only the mean mass properties of a galaxy population. The results of these measurements confirm the large size of dark halos in spirals and in ellipticals (see Sect. 7.7).

The quest for dark matter. These results leave us with a number of obvious questions. What is the nature of the dark matter? What are the density profiles of dark halos, how are they determined, and where is the 'boundary' of a halo? Does the fact that galaxies with $v_{\rm rot} \lesssim 100\,{\rm km/s}$ have no prominent spiral structure mean that a minimum dark matter mass (or mass concentration) needs to be exceeded in order for spiral arms to form?

Some of these questions will be examined later, but here we point out that the major fraction of the mass of (spiral) galaxies consists of non-luminous matter. The fact that we do not know what this matter consists of leaves us with the question of whether this invisible matter is a new, yet unknown, form of matter. Or is the dark matter less exotic, normal (baryonic) matter that is just not luminous for some reason (for example, because it did not form any stars)? We will see in Chap. 4 that the issue of dark matter is not limited to galaxies, but is also clearly present on a cosmological scale; furthermore, the dark matter cannot be baryonic. A currently unknown form of matter is, therefore, revealing itself in the rotation curves of spirals. We will pick up this issue in Sect. 4.4.6 after we have excluded the possibility that the dark matter is composed of unseen baryons.

3.3.5 Stellar populations and gas fraction

The color of spiral galaxies depends on their Hubble type, with later types being bluer; e.g., one finds $B-V \sim 0.75$ for Sa's, 0.64 for Sb's, 0.52 for Sc's, and 0.4 for Irr's. This means that the fraction of massive young stars increases along the Hubble sequence towards later spiral types. This conclusion is also in agreement with the findings for the light distribution along spiral arms where we clearly observe active star formation regions in the bright knots in the spiral arms of Sc's. Furthermore, this color sequence is also in agreement with the decreasing bulge fraction towards later types.

The formation of stars requires gas, and the mass fraction of gas is larger for later types, as can be measured, for instance, from the 21 cm emission of HI, from Hα and from CO emission. Characteristic values for the ratio

$\langle M_{\rm gas}/M_{\rm baryons}\rangle$ are about 0.04 for Sa's, 0.08 for Sb's, 0.16 for Sc's, and 0.25 for Irr's. It thus appears that early-type spirals have been more efficient in the past in turning their gas into stars. In addition, the fraction of molecular gas relative to the total gas mass is smaller for later Hubble types. The dust mass is less than 1 % of the gas mass, or about 0.1 % of the total baryonic mass.

Dust, in combination with hot stars, is the main source of far-infrared (FIR) emission from galaxies. Sc galaxies emit a larger fraction of FIR radiation than Sa's, and barred spirals have stronger FIR emission than normal spirals. The FIR emission arises from dust heated by the UV radiation of hot stars and then re-radiating this energy in the form of thermal emission.

Dust extinction affects the total optical luminosity that is emitted from a spiral galaxy. Depending on the spatial distribution of the dust relative to that of the stars, the extinction can be direction-dependent. Or in other words: the optical luminosity of a spiral galaxy is not necessarily emitted isotropically, but the mean extinction can be higher if a galaxy is seen edge-on. The occurrence of this effect and its strength can be studied, due to the fact that extinction is related to reddening. By studying the mean color of spirals of fixed near-IR luminosity as a function of observed axis ratio, i.e., as a function of inclination angle, one finds that edge-on spirals are redder than face-on galaxies. From the dependence of the reddening on the inclination angle of a large sample of SDSS galaxies, one finds that the typical extinction of an edge-on spiral is 0.7, 0.6, 0.5 and 0.4 magnitudes in the u-, g-, r-, and i-bands, respectively. Hence, spiral galaxies are not really transparent. This effect seems to be weaker for lower-mass spirals, indicating that their relative dust content is smaller than that of high-mass spirals. This is in accord with the fact that the metallicity of lower-mass galaxies is smaller than that of more massive ones (see Fig. 3.40 below).

A prominent color gradient is observed in spirals: they are red in the center and bluer in the outer regions. We can identify at least two reasons for this trend. The first is a metallicity effect, as the metallicity is increasing inwards and metal-rich stars are redder than metal-poor ones, due to their higher opacity. Second, the color gradient can be explained by star formation. Since the gas fraction in the bulge is lower than in the disk, less star formation takes place in the bulge, resulting in a stellar population that is older and redder in general. Furthermore, it is found that the metallicity of spirals increases with luminosity.

Abundance of globular clusters. The number of globular clusters is higher in early types and in more luminous galaxies. The *specific abundance* of globular clusters in a galaxy is defined as their number, normalized to a galaxy of absolute magnitude $M_V = -15$. This can be done by scaling the observed number $N_{\rm t}$ of globular clusters in a galaxy of visual luminosity L_V or absolute magnitude M_V, respectively, to that of a fiducial galaxy with $M_V = -15$, corresponding to a luminosity of $L_V = L_{15}$:

$$S_N = N_{\rm t}\,\frac{L_{15}}{L_V} = N_{\rm t}\,10^{0.4(M_V+15)}\,. \qquad (3.18)$$

If the number of globular clusters were proportional to the luminosity (and thus roughly to the stellar mass) of a galaxy, then this would imply a constant S_N. However, this is not the case: For Sa's and Sb's we find $S_N \sim 1.2$, whereas $S_N \sim 0.5$ for Sc's. S_N is larger for ellipticals and largest for cD galaxies.

3.3.6 Spiral structure

The spiral arms are the bluest regions in spirals and they contain young stars and HII regions. For this reason, the brightness contrast of spiral arms increases as the wavelength of the (optical) observation decreases. In particular, the spiral structure is very prominent in a blue filter, as is shown impressively in Fig. 3.24.

Naturally, the question arises as to the nature of the spiral arms. Probably the most obvious answer would be that they are material structures of stars and gas, rotating around the galaxy's center together with the rest of the disk. However, this scenario cannot explain spiral arm structure since, owing to the differential rotation, they would wind up much more tightly than observed within only a few rotation periods.

Rather, it is suspected that spiral arms are a wave structure, the velocity of which does not coincide with the physical velocity of the stars. Spiral arms are quasi-stationary density waves, regions of higher density (possibly 10–20 % higher than the local disk environment). If the gas, on its orbit around the center of the galaxy, enters a region of higher density, it is compressed, and this compression of molecular clouds results in an enhanced star formation rate. This accounts for the blue color of spiral arms. Since low-mass (thus red) stars live longer, the brightness contrast of spiral arms is lower in red light, whereas massive blue stars are born in the spiral arms and soon after explode there as SNe. Indeed, only few blue stars are found outside spiral arms.

The generation of spiral arms may be induced by a non-axially symmetric perturbation of the gravitational potential of a disk galaxy. Such perturbation can be due to a massive bar in its center, or by companion galaxies. Figure 3.25 shows a particularly impressive case for the latter possibility, together with a multi-color view of this galaxy. The fact that about 65 % of luminous spirals in the local Universe have a

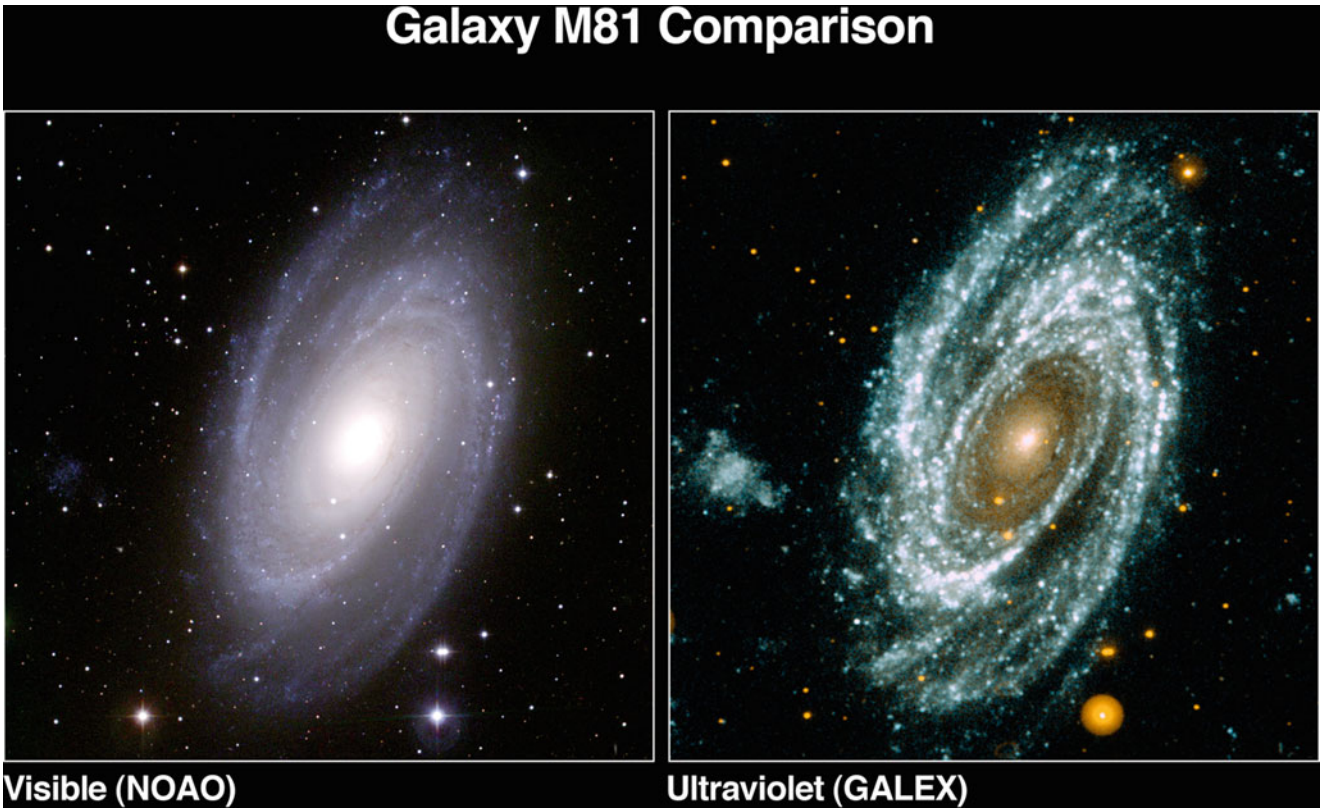


Fig. 3.24 The galaxy M81 in optical light (*left*) and the UV (*right*). The spiral arms are much more prominent in the UV than in optical light, showing that star formation occurs almost exclusively in spiral arms. Note the absence of any visible UV-emission in the center of the galaxy, indicating the lack of hot stars there. Credit: NASA/JPL-Caltech/NOAO

central bar may indicate that bars play an important role in the formation of spiral structure.

In order to better understand density waves we may consider, for example, the waves on the surface of a lake. Peaks at different times consist of different water particles, and the velocity of the waves is by no means the bulk velocity of the water.

3.3.7 Halo gas in spirals

In Sect. 2.3.7 we showed that the disk of the Milky Way is surrounded by gas in its halo, seen as high-velocity clouds in form of neutral hydrogen, absorption by highly ionized species, and X-ray emission from a hot gas. These properties are not unique to the Galaxy.

Corona in spirals. Hot gas resulting from supernovae and their subsequent evolution may expand out of the disk and thereby be ejected to form a hot gaseous halo of a spiral galaxy. We might therefore suspect that such a 'coronal' gas exists outside the galactic disk. While the existence of this coronal gas has long been suspected, the detection of its X-ray emission was first made possible with the ROSAT satellite in the early 1990s. However, the limited angular resolution of ROSAT rendered the distinction between diffuse emission and clusters of discrete sources difficult. Finally, the Chandra observatory unambiguously detected the coronal gas in a number of spiral galaxies. As an example, Fig. 3.26 shows the spiral galaxy NGC 4631.

Neutral hydrogen, in the form of high-velocity clouds, has been detected outside the disk of many spiral galaxies, generally within $\sim$ 10 kpc of the disk. It is estimated that about 25 % of luminous spirals have more than $10^8 M_\odot$ of neutral hydrogen gas outside the disk. In those cases where the HI mass is large it can be related to tidal features or streams indicating the accretion of smaller galaxies—like the Magellanic stream in our Milky Way, which contains most of the neutral hydrogen in our halo.

Ionized gas. Detecting ionized gas and hot gas outside the plane of spiral galaxies is in fact easier in external galaxies than it is in the Milky Way. Whereas emission of Hα is detected outside of the disk is other spirals, the clearest signals are due to absorption lines in the spectra of background sources, usually quasars, which can trace gas with low column density. When the redshifts of the absorption lines

Fig. 3.25 A multi-color view of the spiral galaxy M51, also called the Whirlpool Galaxy, together with its smaller companion NGC 5195 at the *top*. The image has a size of 2′.2 × 3′.2 and different colors indicate: *purple*: X-ray emission as observed by Chandra; *green*: HST optical imaging; *red*: infrared emission as seen by Spitzer; *blue*: UV-radiation observed by GALEX. The spiral arms are the location of young stars, and the dusted heated by them is clearly seen from the IR emission. The X-ray emission mainly comes from compact sources like accreting neutron stars and black holes (called X-ray binaries), but also from diffuse emission by hot gas between the young stars. Simulations indicate that the small companion may have passed through M51 in the recent past; in any case, its physical closeness certainly perturbs the gravitational field of M51 which may be the origin for the very pronounced spiral structure and, at the same time, the increased level of star-formation activity, through added compression of the gas. Credit: X-ray: NASA/CXC/Wesleyan Univ./R.Kilgard et al; UV: NASA/JPL-Caltech; Optical: NASA/ESA/S. Beckwith & Hubble Heritage Team (STScI/AURA); IR: NASA/JPL-Caltech/ Univ. of AZ/R. Kennicutt

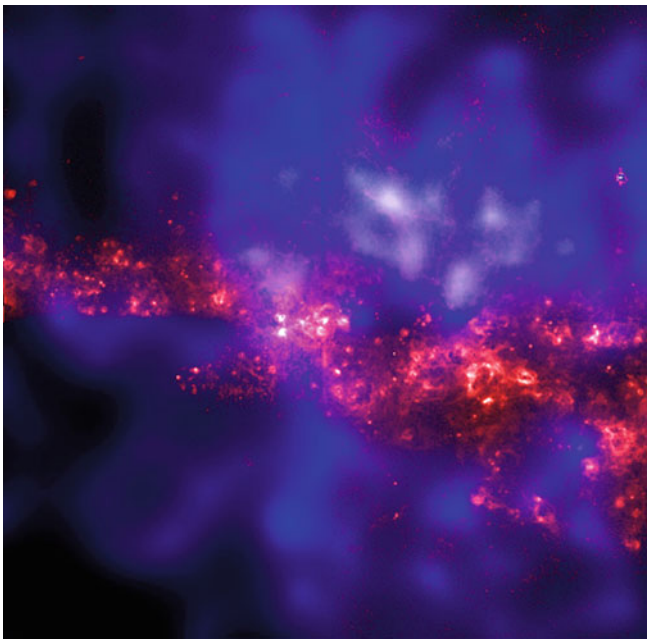


Fig. 3.26 The spiral galaxy NGC 4631. The optical (HST) image of the galaxy is shown in *red*; the many luminous areas are regions of very active star formation. The SN explosions of massive stars eject hot gas into the halo of the galaxy. This gas (at a temperature of $T \sim 10^6$ K) emits X-ray radiation, shown as the *blue* diffuse emission as observed by the Chandra satellite. The image has a size of 2′.5. Credit: X-ray: NASA/UMass/D. Wang et al., Optical: NASA/HST/D. Wang et al.; Research article: D. Wang et al. 2001, *Chandra Detection of a Hot Gaseous Corona around the Edge-on Galaxy NGC 4631*, ApJ 555, L99

are correlated with the position of foreground galaxies, one can study the extend and covering factor of the gas giving rise to this absorption. One finds that the Lyα absorption has a covering factor of almost unity within $\sim$ 300 kpc of luminous galaxies, and the covering factor of MgII absorption within 100 kpc is about 50 %. Hence, warm ionized gas extends to large separation from galaxies. In addition, in star-forming galaxies, one also finds highly ionized species (like OVI) at large distance from the disk.

Taken together, we find that the halo gas in other spirals shares the properties of that in the Milky Way. The gaseous halo is a busy place, where one meets metal-enriched gas driven out by energetic processes in the galactic disk, and low-metallicity gas falling in from larger distances, providing new raw material for continued star formation. The various phases of the gas in halos are essentially in pressure equilibrium, i.e., their density scales inversely to their temperature.

3.4 Scaling relations

The properties of a galaxy are characterized by a number of quantities, such as luminosity, size, mass, rotational velocity or velocity dispersion, color, star-formation rate etc. At first sight one might think that galaxies can exist where these different quantities take on a large range of values. However, this is not the case: The properties of isolated galaxies seem to be determined by just a few parameters, from which the others follow.

As a first example of that fact, we will show in this section that the kinematic properties of spirals and ellipticals are closely related to their luminosity. As we shall discuss below, spirals follow the *Tully–Fisher relation* (Sect. 3.4.1), whereas elliptical galaxies obey the *Faber–Jackson relation* (Sect. 3.4.2) and are located in the *fundamental plane* (Sect. 3.4.3). These scaling relations are a very important tool for distance estimations, as will be discussed in Sect. 3.9. Furthermore, these scaling relations express relations between galaxy properties which any successful model of galaxy evolution must be able to explain. Here we will describe these scaling relations and discuss their physical origin.

3.4.1 The Tully–Fisher relation

Using 21 cm observations of spiral galaxies, in 1977 R. Brent Tully and J. Richard Fisher found that the maximum rotation velocity of spirals is closely related to their luminosity, following the relation

$$\boxed{L \propto v_{\rm max}^{\alpha}}\,, \tag{3.19}$$

where the power-law index (i.e., the slope) of the Tully–Fisher relation is about $\alpha \sim 4$. The larger the wavelength of the filter in which the luminosity is measured, the smaller the dispersion of the Tully–Fisher relation (see Fig. 3.27). This is to be expected because radiation at larger wavelengths is less affected by dust absorption and by the current star formation rate, which may vary to some extent between individual spirals. Furthermore, it is found that the value of α increases with the wavelength of the filter: The Tully–Fisher relation is steeper in the red, which follows from the fact that more massive, or more luminous galaxies—i.e., those with larger $v_{\rm max}$—are redder, as can be seen from Fig. 3.7. The dispersion of galaxies around the relation (3.19) in the near-infrared (e.g., in the H-band) is about 10 %.

Because of this close correlation, the luminosity of spirals can be estimated quite precisely by measuring the rotational velocity. The determination of the (maximum) rotational velocity is independent of the galaxy's distance. By comparing the luminosity, as determined from the Tully–Fisher relation, with the measured flux, one can then estimate the distance of the galaxy—without utilizing the Hubble relation!

The measurement of $v_{\rm max}$ is obtained either from a spatially resolved rotation curve, by measuring $v_{\rm rot}(\theta)$, which can be done with optical spectroscopy or, for relatively nearby galaxies, also with spatially resolved 21 cm spectroscopy. Alternatively, one can observe an integrated spectrum of the 21 cm line of HI that has a Doppler width corresponding to about $2v_{\rm max}$ (see Fig. 3.28). The Tully–Fisher relation shown in Fig. 3.27 was determined by measuring the width of the 21 cm line.

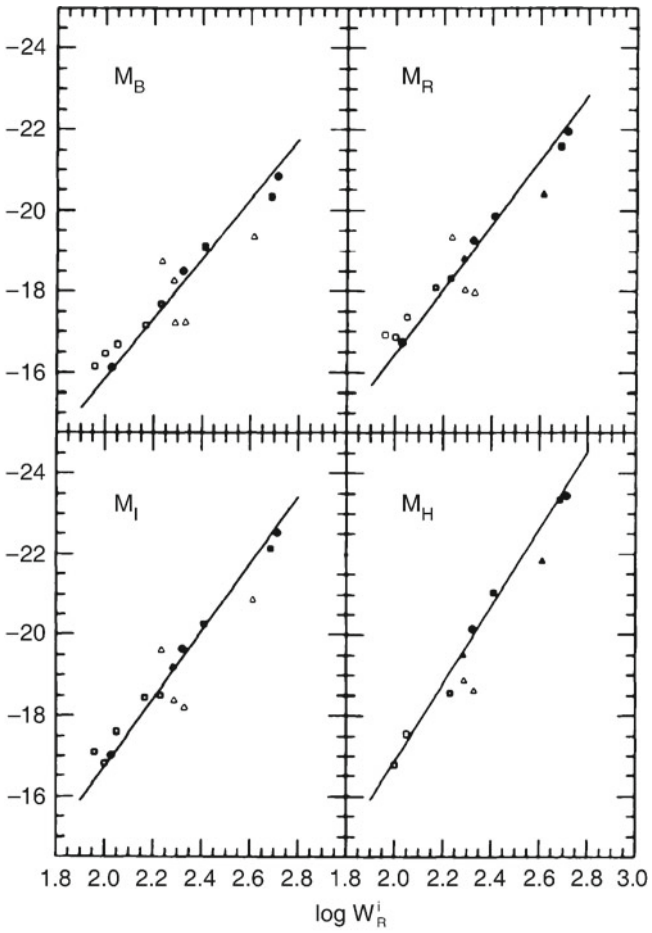


Fig. 3.27 The Tully–Fisher relation for galaxies in the Local Group (*dots*), in the Sculptor group (*triangles*), and in the M81 group (*squares*). The absolute magnitude is plotted as a function of the width of the 21 cm profile which indicates the maximum rotation velocity (see Fig. 3.28). *Filled symbols* represent galaxies for which independent distance estimates were obtained, either from RR Lyrae stars, Cepheids, or planetary nebulae. For galaxies represented by *open symbols*, the average distance of the respective group is used. The *solid line* is a fit to similar data for the Ursa-Major cluster, together with data of those galaxies for which individual distance estimates are available (*filled symbols*). The larger dispersion around the mean relation for the Sculptor group galaxies is due to the group's extent along the line-of-sight. Source: M.J. Pierce & R.B. Tully 1992, *Luminosity-line width relations and the extragalactic distance scale. I—Absolute calibration*, ApJ 387, 47, p. 51, Fig. 1. ©AAS. Reproduced with permission

Explaining the Tully–Fisher-relation. The shapes of the rotation curves of spirals are very similar to each other, in particular with regard to their flat behavior in the outer part. The flat rotation curve implies

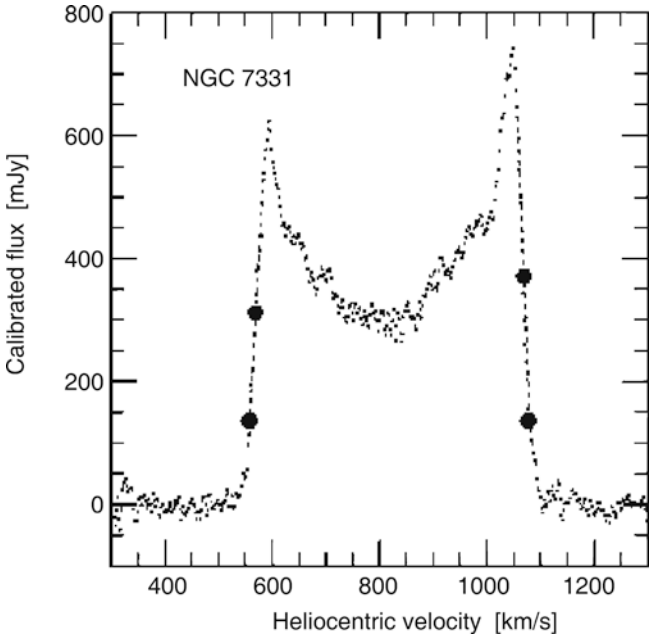


Fig. 3.28 21 cm profile of the galaxy NGC 7331. The *bold dots* indicate 20 and 50 % of the maximum flux; these are of relevance for the determination of the line width from which the rotational velocity is derived. Source: L.M. Macri et al. 2000, *A Database of Tully–Fisher Calibrator Galaxies*, ApJS 128, 461, p. 467, Fig. 5. ©AAS. Reproduced with permission

$$\boxed{M = \frac{v_{\max}^2 R}{G}}\,, \tag{3.20}$$

where the value of the distance R from the center of the galaxy is chosen to be in the range of the flat part of the rotation curve, i.e., where $v_{\rm rot}(R) \approx V_{\max}$. We note that the exact value of R is not important; of course, $M = M(R)$ in (3.20). By re-writing (3.20),

$$L = \left(\frac{M}{L}\right)^{-1} \frac{v_{\max}^2 R}{G}\,, \tag{3.21}$$

and replacing R by the mean surface brightness $\langle I\rangle = L/R^2$, we obtain

$$L = \left(\frac{M}{L}\right)^{-2} \left(\frac{1}{G^2\,\langle I\rangle}\right) v_{\max}^4\,. \tag{3.22}$$

This is the Tully–Fisher relation *if* M/L and $\langle I\rangle$ are the same for all spirals. As discussed previously, the latter is in fact suggested by Freeman's law (Sect. 3.3.2). Since the shapes of rotation curves for spirals seem to be very similar, the radial dependence of the ratio of luminous to dark matter may also be quite similar among spirals. Furthermore, since the mass-to-light ratios of a stellar population as measured from the red or infrared emission do not depend strongly on its age, the constancy of M/L could also be valid if dark matter is included.

Although the line of argument presented above is far from a rigorous derivation of the Tully–Fisher-relation, it nevertheless makes the existence of such a scaling relation plausible.

Mass-to-light ratio of spirals. We are unable to determine the total mass of a spiral because the extent of the dark halo is unknown. Thus we can measure M/L only within a fixed radius. We shall define this radius as R_{25}, the radius at which the surface brightness attains the value of 25 mag/arcsec^2 in the B-band[6]; then spirals follow the relation

$$\boxed{\log\left(\frac{R_{25}}{\rm kpc}\right) = -0.249 M_B - 4.00}\,, \tag{3.23}$$

independently of their Hubble type. Within R_{25} one finds $M/L_B = 6.2$ for Sa's, 4.5 for Sb's, and 2.6 for Sc's. This trend does not come as a surprise because late types of spirals contain more young, blue and luminous stars.

The baryonic Tully–Fisher relation. The above 'derivation' of the Tully–Fisher relation is based on the assumption of a constant M/L value, where M is the total mass (i.e., including dark matter). Let us assume that (1) the ratio of baryons to dark matter is constant, and furthermore that (2) the stellar populations in spirals are similar, so that the ratio of stellar mass to luminosity is a constant. Even under these assumptions we would expect the Tully–Fisher relation to be valid only if the gas does not, or only marginally, contribute to the baryonic mass. However, low-mass spirals contain a significant fraction of gas, so we should expect that the Tully–Fisher relation does not apply to these galaxies. Indeed, it is found that spirals with a small $v_{\max} \lesssim 100\,{\rm km/s}$ deviate significantly from the Tully–Fisher relation—see Fig. 3.29a.

Since the luminosity is approximately proportional to the stellar mass, $L \propto M_*$, the Tully–Fisher relation is a relation between $v_{\max}$ and M_*. Adding the mass of the gas, which can be determined from the strength of the 21 cm line and molecular emission, to the stellar mass, a much tighter correlation is obtained, see Fig. 3.29b. It reads

$$\boxed{M_{\rm disk} = 2\times 10^9\, h^{-2}\, M_\odot \left(\frac{v_{\max}}{100\,{\rm km/s}}\right)^4}\,, \tag{3.24}$$

and is valid over five orders of magnitude in disk mass $M_{\rm disk} = M_* + M_{\rm gas}$. If no further baryons exist in spirals

[6] We point out explicitly once more that the surface brightness does not depend on the distance of a source.

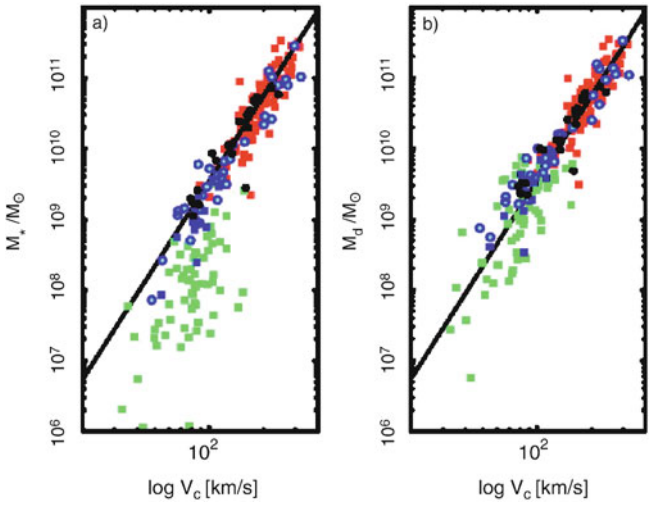


Fig. 3.29 (**a**) The mass contained in stars as a function of the rotational velocity V_c for spirals. This stellar mass is computed from the luminosity by multiplying it with a suitable stellar mass-to-light ratio which depends on the chosen filter and which can be calculated from stellar population models. This is the 'classical' Tully–Fisher relation. *Squares* and *circles* denote galaxies for which V_c was determined from the 21 cm line width or from a spatially resolved rotation curve, respectively. The *colors* of the *symbols* indicate the filter band in which the luminosity was measured: H (*red*), K' (*black*), I (*green*), B (*blue*). (**b**) Instead of the stellar mass, here the sum of the stellar and gaseous mass is plotted. The gas mass was derived from the flux in the 21 cm line, $M_{\rm gas} = 1.4 M_{\rm HI}$, corrected for helium and metals. The *line* in both plots is the Tully–Fisher relation with a slope of $\alpha = 4$. Source: S. McGaugh et al. 2000, *The Baryonic Tully-Fisher Relation*, ApJ 533, L99, p. L100, Fig. 1. ©AAS. Reproduced with permission

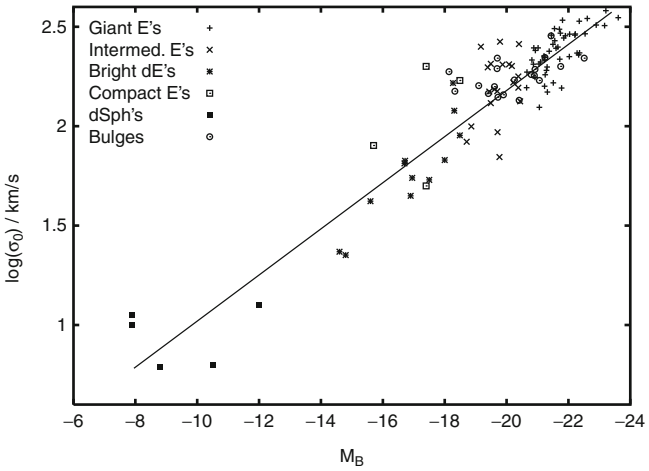


Fig. 3.30 The Faber–Jackson relation expresses a relation between the velocity dispersion and the luminosity of elliptical galaxies. It can be derived from the virial theorem. Data from R. Bender et al. 1992, ApJ 399, 462

(such as, e.g., MACHOs), this close relation means that the ratio of baryons and dark matter in spirals is constant over a very wide mass range.

3.4.2 The Faber–Jackson relation

A relation for elliptical galaxies, analogous to the Tully–Fisher relation, was found by Sandra Faber and Roger Jackson. They discovered that the velocity dispersion in the center of ellipticals, σ_0, scales with luminosity (see Fig. 3.30),

$$\boxed{L \propto \sigma_0^4}\,, \quad \text{or} \quad \boxed{\log(\sigma_0) = -0.1 M_B + \text{const.}}\,. \tag{3.25}$$

'Deriving' the Faber–Jackson scaling relation is possible under the same assumptions as for the Tully–Fisher relation. However, the dispersion of ellipticals about this relation is larger than that of spirals about the Tully–Fisher relation.

3.4.3 The fundamental plane

The Tully–Fisher and Faber–Jackson relations specify a connection between the luminosity and a kinematic property of galaxies. As we discussed previously, various relations exist between the parameters of elliptical galaxies. Thus one might wonder whether a relation exists between observables of elliptical galaxies for which the dispersion is smaller than that of the Faber–Jackson relation. Such a relation was indeed found and is known as the *fundamental plane*.

To explain this relation, we will consider the various relations between the parameters of ellipticals. In Sect. 3.2.2 we saw that the effective radius of normal ellipticals and cD's, i.e., excluding dwarfs, is related to the luminosity (see Fig. 3.10). This implies a relation between the surface brightness and the effective radius,

$$\boxed{R_{\rm e} \propto \langle I \rangle_{\rm e}^{-0.83}}\,, \tag{3.26}$$

where $\langle I \rangle_{\rm e}$ is the average surface brightness within the effective radius, so that

$$\boxed{L = 2\pi R_{\rm e}^2 \langle I \rangle_{\rm e}}\,. \tag{3.27}$$

From this, a relation between the luminosity and $\langle I \rangle_{\rm e}$ results,

$$L \propto R_{\rm e}^2 \langle I \rangle_{\rm e} \propto \langle I \rangle_{\rm e}^{-0.66} \text{ or approximately } \boxed{\langle I \rangle_{\rm e} \propto L^{-1.5}}\,. \tag{3.28}$$

Hence, more luminous ellipticals have smaller surface brightnesses, as is also shown in Fig. 3.10. By means of the Faber–Jackson relation, L is related to σ_0, the central velocity dispersion, and therefore, σ_0, $\langle I \rangle_{\rm e}$, and $R_{\rm e}$ are related to each other. The distribution of elliptical galaxies in the three-dimensional parameter space $(R_{\rm e}, \langle I \rangle_{\rm e}, \sigma_0)$ is located close to a plane defined by

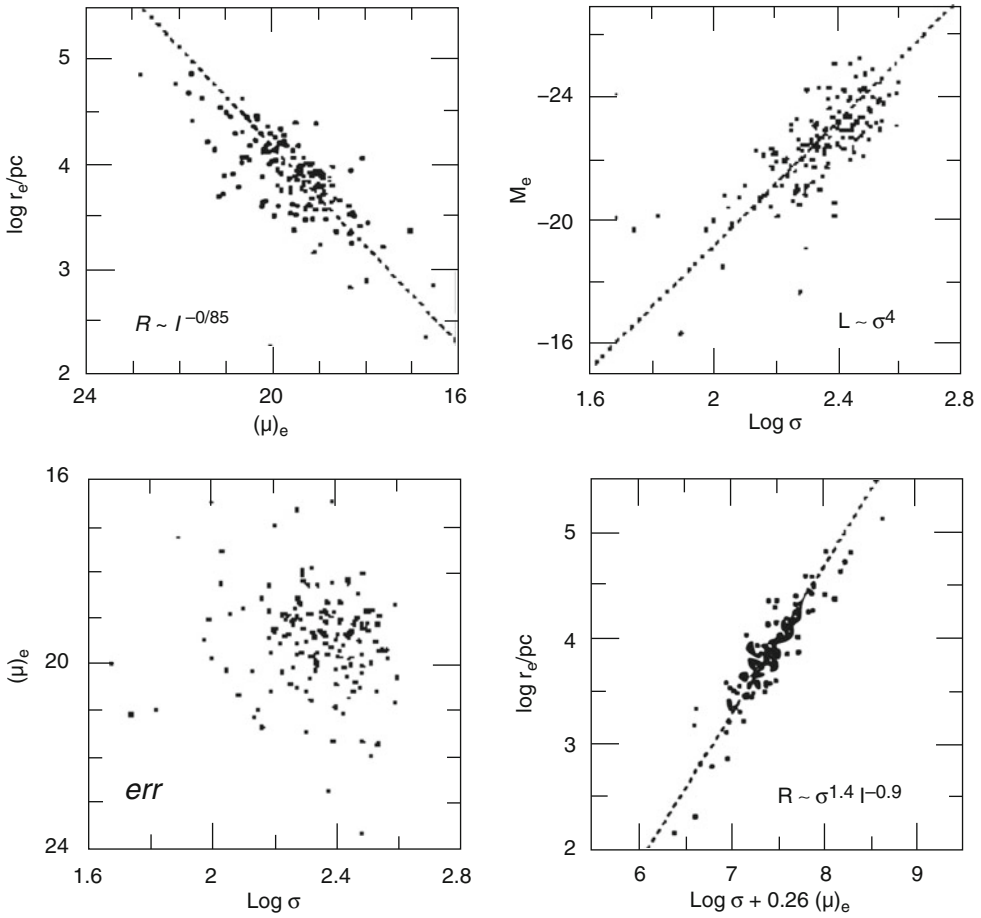


Fig. 3.31 Projections of the fundamental plane onto different two-parameter planes. *Upper left*: the relation between radius and mean surface brightness within the effective radius. *Upper right*: Faber–Jackson relation. *Lower left*: the relation between mean surface brightness and velocity dispersion shows the fundamental plane viewed from above. *Lower right*: the fundamental plane viewed from the side—the linear relation between radius and a combination of surface brightness and velocity dispersion. Source: J. Kormendy & S. Djorgovski 1989, *Surface photometry and the structure of elliptical galaxies*, ARA&A 27, 235, Fig. 2, p. 255. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 27 ©1989 by Annual Reviews www.annualreviews.org

$$\boxed{R_\mathrm{e} \propto \sigma_0^{1.4}\,\langle I\rangle_\mathrm{e}^{-0.85}}\,. \tag{3.29}$$

Writing this relation in logarithmic form, we obtain

$$\boxed{\log R_\mathrm{e} = 0.34\,\langle\mu\rangle_\mathrm{e} + 1.4\log\sigma_0 + \text{const.}}\,, \tag{3.30}$$

where $\langle\mu\rangle_\mathrm{e}$ is the average surface brightness within R_e, measured in mag/arcsec^2. Equation (3.30) defines a plane in this three-dimensional parameter space that is known as the *fundamental plane (FP)*. Different projections of the fundamental plane are displayed in Fig. 3.31.

How can this be explained? The mass within R_e can be derived from the virial theorem, $M \propto \sigma_0^2 R_\mathrm{e}$. Combining this with (3.27) yields

$$R_\mathrm{e} \propto \frac{L}{M}\,\frac{\sigma_0^2}{\langle I\rangle_\mathrm{e}}\,. \tag{3.31}$$

If the mass-to-light ratio M/L was the same for all ellipticals, then (3.31) deviates significantly from the observed fundamental plane (3.29). This deviation is often called the *tilt* of the fundamental plane. This tilt can be accounted for if the mass-to-light ratio varies systematically with the galaxy mass. To wit, (3.31) agrees with the FP in the form of (3.29) if

$$\frac{L}{M}\,\frac{\sigma_0^2}{\langle I\rangle_\mathrm{e}} \propto \frac{\sigma_0^{1.4}}{\langle I\rangle_\mathrm{e}^{0.85}}\,,$$

or

$$\frac{M}{L} \propto \frac{\sigma_0^{0.6}}{\langle I\rangle_\mathrm{e}^{0.15}} \propto \frac{M^{0.3}}{R_\mathrm{e}^{0.3}}\,\frac{R_\mathrm{e}^{0.3}}{L^{0.15}}\,.$$

Hence, the FP follows from the virial theorem provided

$$\left(\frac{M}{L}\right) \propto M^{0.2} \quad \text{or} \quad \left(\frac{M}{L}\right) \propto L^{0.25}\,, \quad \text{respectively,} \tag{3.32}$$

i.e., if the mass-to-light ratio of galaxies increases slightly with mass. Since the luminosity is approximately proportional to the stellar mass, in particular for an old stellar population as found in ellipticals, then (3.32) implies that the ratio of dynamical mass M to stellar mass increases with mass. This increase of M/L with mass could in principle have its origin in a changing M_*/L with increasing mass or luminosity, since we have seen in Fig. 3.7 that more luminous red galaxies are somewhat redder, indicating an older stellar population, which in turn would imply an increasing M_*/L. However, this effect is far too small to explain the tilt of the fundamental plane. In addition, the tilt is also seen if the fundamental plane is studied at near-IR wavelengths, for which M_*/L is much less age dependent. Hence we conclude that the tilt in the fundamental plane is not related to properties of the stellar population.

Like the Tully–Fisher relation, the fundamental plane is an important tool for distance estimations, as will be discussed more thoroughly later.

3.4.4 D_n-σ relation

Another scaling relation for ellipticals which is of substantial importance in practical applications is the D_n-σ relation. D_n is defined as the mean diameter of an ellipse within which the average surface brightness I_n corresponds to a value of 20.75 mag/arcsec^2 in the B-band. If we now assume that all ellipticals have a self-similar brightness profile, $I(R) = I_\mathrm{e}\, f(R/R_\mathrm{e})$, with $f(1) = 1$, then the luminosity within D_n can be written as

$$\begin{aligned} I_n \left(\frac{D_n}{2}\right)^2 \pi &= 2\pi I_\mathrm{e} \int_0^{D_n/2} \mathrm{d}R\; R\, f(R/R_\mathrm{e}) \\ &= 2\pi I_\mathrm{e} R_\mathrm{e}^2 \int_0^{D_n/(2R_\mathrm{e})} \mathrm{d}x\; x\, f(x)\,, \end{aligned}$$

where in the last step we changed the integration variable to $x = R/R_\mathrm{e}$. For a de Vaucouleurs profile we have approximately $f(x) \propto x^{-1.2}$ in the relevant range of radius. Computing the integral with this expression, we obtain

$$D_n \propto R_\mathrm{e}\, I_\mathrm{e}^{0.8}\,. \tag{3.33}$$

Replacing R_e by the fundamental plane (3.29) then results in

$$D_n \propto \sigma_0^{1.4}\, \langle I \rangle_\mathrm{e}^{-0.85}\, I_\mathrm{e}^{0.8}\,.$$

Since we assumed a self-similar brightness profile, we have $\langle I \rangle_\mathrm{e} \propto I_\mathrm{e}$, and thus we finally find

$$\boxed{D_n \propto \sigma_0^{1.4}\, I_\mathrm{e}^{0.05}}\,. \tag{3.34}$$

This implies that D_n is nearly independent of I_e and only depends on σ_0. The D_n-σ relation (3.34) describes the properties of ellipticals considerably better than the Faber–Jackson relation and, in contrast to the fundamental plane, it is a relation between only two observables. Empirically, we find that ellipticals follow the normalized D_n-σ relation

$$\frac{D_n}{\mathrm{kpc}} = 2.05 \left(\frac{\sigma_0}{100\,\mathrm{km/s}}\right)^{1.33}\,, \tag{3.35}$$

and they scatter around this relation with a relative width of about 15 %.

3.4.5 Summary: Properties of galaxies on the Hubble sequence

After having discussed the basic properties of the two main types of galaxies and some of the scaling relations they obey, this is a good place to pause and summarize the main points:

- Most luminous galaxies in the local Universe fit onto the Hubble sequence; they are either ellipticals, spirals, or belong to the class of S0 galaxies, which shares some properties with the two other classes.
- Ellipticals and spirals differ not only in their morphology, but in several other respects, for example: (1) Spirals contain a sizable fraction of gas, whereas the gas-to-stellar mass ratio in ellipticals is much smaller. As a consequence, (2) spirals have ongoing star formation, ellipticals not, or only very little. As a further consequence, (3) the light of elliptical galaxies is substantially redder than that of spirals. Obviously, the morphology of galaxies and the properties of their stellar populations are strongly correlated.
- The stars in spirals have a very ordered motion, moving around the galactic center on nearly circular orbits in a common orbital plane, having a velocity dispersion that is much smaller than the orbital velocity; the stars in the disk are called ‘dynamically cold’. In contrast, the motion of stars in ellipticals is largely random, with fairly little coherent velocity; they are dynamically hot.
- Some elliptical galaxies show clear signs of complex structure, which are interpreted as indications of past interaction with other galaxies. In contrast, the disks of spirals are very thin, which means that they have been largely unperturbed for a long while in the past.

- The rotation curves of spiral galaxies are almost flat for large radii, in contrast to what would be expected from the visible mass distribution that declines exponentially outwards. This implies that there is more matter than seen in stars and gas—the galaxies are embedded in a halo of dark matter. Whereas for elliptical galaxies the radial density distribution is more difficult to probe, the presence of dark matter has been verified also for ellipticals.
- Both, spirals and ellipticals, follow scaling relations which connect their luminous properties (luminosity or surface brightness) with their dynamical properties (rotational velocity or velocity dispersion). Hence, the formation and evolution of galaxies and their stellar populations must proceed in a way as to place them onto these scaling relations.

Next, we will consider the properties of stellar populations in somewhat more detail, since they are a key in relating the observed luminous properties of galaxies to their underlying baryonic component.

3.5 Population synthesis

The light of normal galaxies originates from stars. Stellar evolution is largely understood, and the spectral radiation of stars can be calculated from the theory of stellar atmospheres. If the distribution of the number density of stars is known as a function of their mass, chemical composition, and evolutionary stage, we can compute the light emitted by them. The *theory of population synthesis* aims at interpreting the spectrum of galaxies as a superposition of stellar spectra. We have to take into account the fact that the distribution of stars changes over time; e.g., massive stars leave the main sequence after several 10^6 yr, the number of luminous blue stars thus decreases, which means that the spectral distribution of the population also changes in time. The spectral energy distribution of a galaxy thus reflects its history of star formation and stellar evolution. For this reason, simulating different star formation histories and comparing them with observed galaxy spectra provides important clues for understanding the evolution of galaxies. In this section, we will discuss some aspects of the theory of population synthesis; this subject is of tremendous importance for our understanding of galaxy spectra.

3.5.1 Model assumptions

The processes of star formation are not understood in detail; for instance, it is currently impossible to compute the mass spectrum of a group of stars that jointly formed in a molecular cloud. Obviously, high-mass and low-mass stars are born together and form young (open) star clusters. The mass spectra of these stars are determined empirically from observations.

The *initial mass function* (IMF) is defined as the initial mass distribution at the time of birth of the stars, such that $\phi(m)\,\mathrm{d}m$ specifies the fraction of stars in the mass interval of width $\mathrm{d}m$ around m, where the distribution is normalized,

$$\boxed{\int_{m_\mathrm{L}}^{m_\mathrm{U}} \mathrm{d}m\; m\,\phi(m) = 1M_\odot}\;.$$

The integration limits are not well defined. Typically, one uses $m_\mathrm{L} \sim 0.1M_\odot$ because stars less massive than $\approx 0.08M_\odot$ do not ignite their hydrogen (and are thus brown dwarfs), and $m_\mathrm{U} \sim 100M_\odot$, because considerably more massive stars are not observed. Whereas such very massive stars would in any case be difficult to observe because of their very short lifetime, the theory of stellar structure tells us that more massive stars can probably not form a stable configuration due to excessive radiation pressure. The shape of the IMF is also subject to uncertainties; in most cases, the *Salpeter-IMF* is used,

$$\boxed{\phi(m) \propto m^{-2.35}}\;, \tag{3.36}$$

as obtained from investigating the stellar mass spectrum in young star clusters. It is by no means clear whether a universal IMF exists, or whether it depends on specific conditions like metallicity, the mass of the galaxy, cosmic epoch, or other parameters. Given the difficulties of determining the shape of the IMF, apparent variations of the IMF with epoch or environment may be attributed to other effect, such as the specifics of the star-formation history in galaxies. Therefore, there seems to be no clear direct indication that the IMF varies with environment. However, as will be discussed in Chap. 10, some properties of high-redshift galaxies are very difficult to understand if their IMF would be the same as in our neighborhood. It has therefore been suggested that the IMF in starbursts is different from that of quiescent star formation such as we are experiencing in the Milky Way.

The Salpeter-IMF seems to be a good description for stars with $M \gtrsim 1M_\odot$, whereas the IMF for less massive stars is flatter. Note that, due to the steep slope of the IMF, most of the stellar mass is contained in low-mass stars. However, since the luminosity of main-sequence stars depends strongly on mass, approximately as $L \propto M^3$, most of the luminosity comes from high-mass stars (see Problem 3.2).

The *star-formation rate* is the gas mass that is converted into stars per unit time,

$$\psi(t) = -\frac{\mathrm{d}M_\mathrm{gas}}{\mathrm{d}t}\;.$$

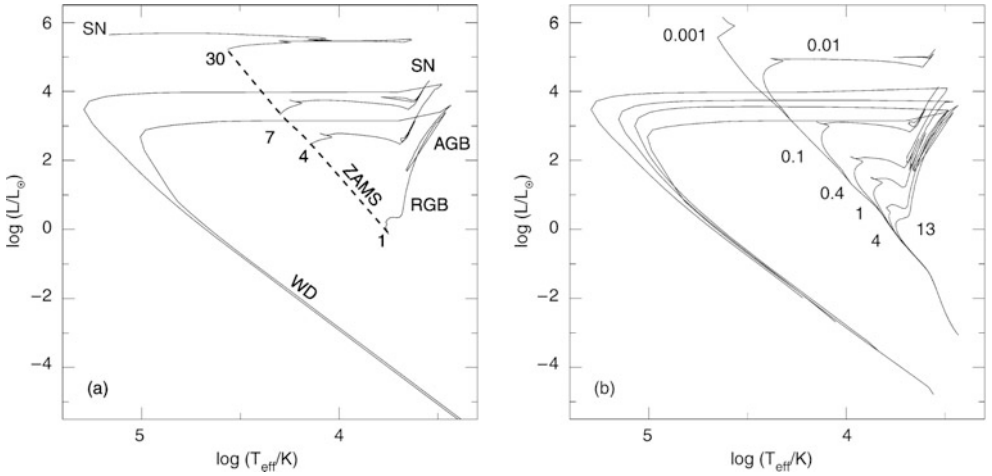


Fig. 3.32 (**a**) Evolutionary tracks in the HRD for stars of different masses, as indicated by the numbers near the tracks (in units of $M_\odot$). The ZAMS (zero age main sequence) is the place of birth in the HRD; evolution moves stars away from the main sequence. Depending on the mass, they explode as a core-collapse SN (for $M \geq 8M_\odot$) or end as a white dwarf (WD). Prior to this, they move along the red giant branch (RGB) and the asymptotic giant branch (AGB). (**b**) Isochrones at different times, indicated in units of 10^9 yr. An isochrone (for a given time t) is a curve connecting the location of stars in the HRD which all have the same age t. Thus, an isochrone shows the distribution of stars from a single star-formation event after a time t. The upper main sequence is quickly depopulated by the rapid evolution of massive stars, whereas the red giant branch is populated over time. Source: S. Charlot 1996, *Spectral Evolution of Galaxies*, Lecture Notes in Physics 470, Springer-Verlag, p. 53

The metallicity Z of the ISM defines the metallicity of the newborn stars, and the stellar properties in turn depend on Z. During stellar evolution, metal-enriched matter is ejected into the ISM by stellar winds, planetary nebulae, and SNe, so that $Z(t)$ is an increasing function of time. This chemical enrichment must be taken into account in population synthesis studies in a self-consistent form.

Let $S_{\lambda,Z}(t')$ be the emitted energy per wavelength and time interval, normalized to an initial total mass of $1M_\odot$, emitted by a group of stars of initial metallicity Z and age t'. The function $S_{\lambda,Z(t-t')}(t')$, which describes this emission at any point t in time, accounts for the different evolutionary tracks of the stars in the Hertzsprung–Russell diagram (HRD)—see Appendix B.2. It also accounts for their initial metallicity (i.e., at time $t - t'$), where the latter follows from the chemical evolution of the ISM of the corresponding galaxy. Then the total spectral luminosity of this galaxy at a time t is given by

$$F_\lambda(t) = \int_0^t \mathrm{d}t' \, \psi(t-t') \, S_{\lambda,Z(t-t')}(t') \, , \qquad (3.37)$$

thus by the convolution of the star formation rate with the spectral energy distribution of the stellar population. In particular, $F_\lambda(t)$ depends on the star formation history.

3.5.2 Evolutionary tracks in the HRD; integrated spectrum

In order to compute $S_{\lambda,Z(t-t')}(t')$, models for stellar evolution and stellar atmospheres are needed. As a reminder, Fig. 3.32a displays the evolutionary tracks in the HRD. Each track shows the position of a star with specified mass in the HRD and is parametrized by the time since its formation. Positions of equal time in the HRD are called *isochrones* and are shown in Fig. 3.32b. As time proceeds, fewer and fewer massive stars exist because they quickly leave the main sequence and end up as supernovae or white dwarfs. The number density of stars along the isochrones depends on the IMF. The spectrum $S_{\lambda,Z(t-t')}(t')$ is then the sum over all spectra of the stars on an isochrone—see Fig. 3.33b.

In the beginning, the spectrum and luminosity of a stellar population are dominated by the most massive stars, which emit intense UV radiation. But after $\sim 10^7$ yr, the flux below 1000 Å is diminished significantly, and after $\sim 10^8$ yr, it hardly exists any more. At the same time, the flux in the NIR increases because the massive stars evolve into red supergiants.

For 10^8 yr $\lesssim t \lesssim 10^9$ yr, the emission in the NIR remains high, whereas short-wavelength radiation is more and more diminished. After $\sim 10^9$ yr, red giant stars (RGB stars) account for most of the NIR production. After $\sim 3 \times 10^9$ yr, the UV radiation increases again slightly, due to blue stars on the horizontal branch into which stars evolve after the AGB phase, and due to white dwarfs which are hot when they are born. Between an age of 4 and 13 billion years, the spectrum of a stellar population evolves fairly little.

Of particular importance is the spectral break located at about 4000 Å which becomes visible in the spectrum after a few 10^7 yr. This break is caused by a strongly changing opacity of stellar atmospheres at this wavelength, mainly due to strong transitions of singly ionized calcium and the

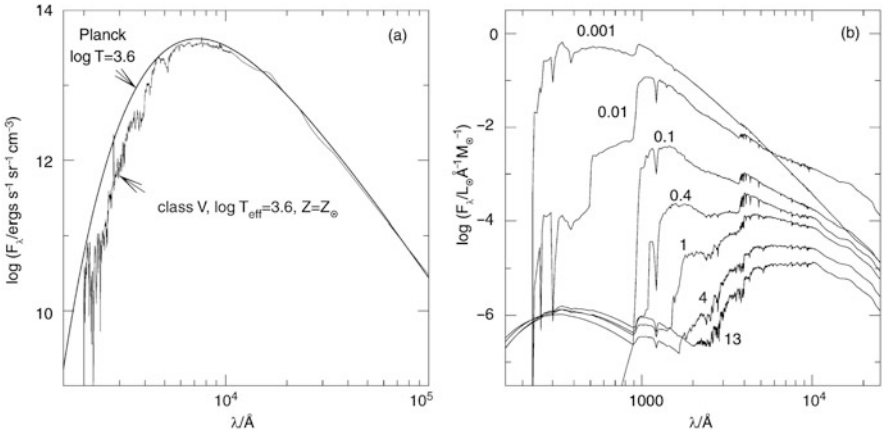


Fig. 3.33 (**a**) Comparison of the spectrum of a main sequence star with a black body spectrum of equal effective temperature. The opacity of the stellar atmosphere causes clear deviations from the Planck spectrum in the UV/optical. (**b**) Spectrum of a stellar population with Solar metallicity that was instantaneously born a time t ago; t is given in units of 10^9 yr. Source: S. Charlot 1996, *Spectral Evolution of Galaxies*, Lecture Notes in Physics 470, Springer-Verlag, p. 53

Balmer lines of hydrogen. This *4000 Å-break* is one of the most important spectral properties of the continuum stellar emission in galaxies; as we will discuss in Sect. 9.1.2, it allows us to estimate the redshifts of early-type galaxies from their photometric properties—so-called photometric redshift estimates.

3.5.3 Color evolution

Detailed spectra of galaxies are often not available. Instead we have photometric images in different broadband filters, since the observing time required for spectroscopy is substantially larger than for photometry. In addition, modern wide-field cameras can obtain photometric data of numerous galaxies simultaneously. From the theory of population synthesis we can derive photometric magnitudes by multiplying model spectra with the filter functions, i.e., the transmission curves of the color filters used in observations, and then integrating over wavelength (A.25). Hence the spectral evolution implies a color evolution, as is illustrated in Fig. 3.34a.

For a young stellar population the color evolution is rapid and the population becomes redder, again because the hot blue stars have a higher mass and thus evolve quickly in the HRD. For the same reason, the evolution is faster in $B-V$ than in $V-K$. It should be mentioned that this color evolution is also observed in star clusters of different ages. The mass-to-light ratio M/L also increases with time because M remains constant while L decreases.

As shown in Fig. 3.34b, the blue light of a stellar population is always dominated by main sequence stars, although at later stages a noticeable contribution also comes from horizontal branch stars. The NIR radiation is first dominated by stars burning helium in their center (this class includes the supergiant phase of massive stars), later by AGB stars, and after $\sim 10^9$ yr by red giants. Main sequence stars never contribute more than 20 % of the light in the K-band. The fact that M/L_K varies only little with time implies that the NIR luminosity is a good indicator for the total stellar mass: the NIR mass-to-light ratio is much less dependent on the age of the stellar population than that for bluer filters.

3.5.4 Star formation history and galaxy colors

Up to now, we have considered the evolution of a stellar population of a common age (called an *instantaneous burst of star formation*). However, star formation in a galaxy takes place over a finite period of time. We expect that the star formation rate decreases over time because more and more matter is bound in stars and thus no longer available to form new stars. Since the star formation history of a galaxy is a priori unknown, it needs to be parametrized in a suitable manner. A 'standard model' of an exponentially decreasing star formation rate was established for this,

$$\boxed{\psi(t) = \tau^{-1} \exp\left[-(t-t_\mathrm{f})/\tau\right] \mathrm{H}(t-t_\mathrm{f})}\,, \qquad (3.38)$$

where τ is the characteristic duration and t_f the onset of star formation. The last factor in (3.38) is the Heaviside step function, $\mathrm{H}(x) = 1$ for $x \geq 0$, $\mathrm{H}(x) = 0$ for $x < 0$. This Heaviside step function accounts for the fact that $\psi(t) = 0$ for $t < t_\mathrm{f}$. We may hope that this simple model describes the

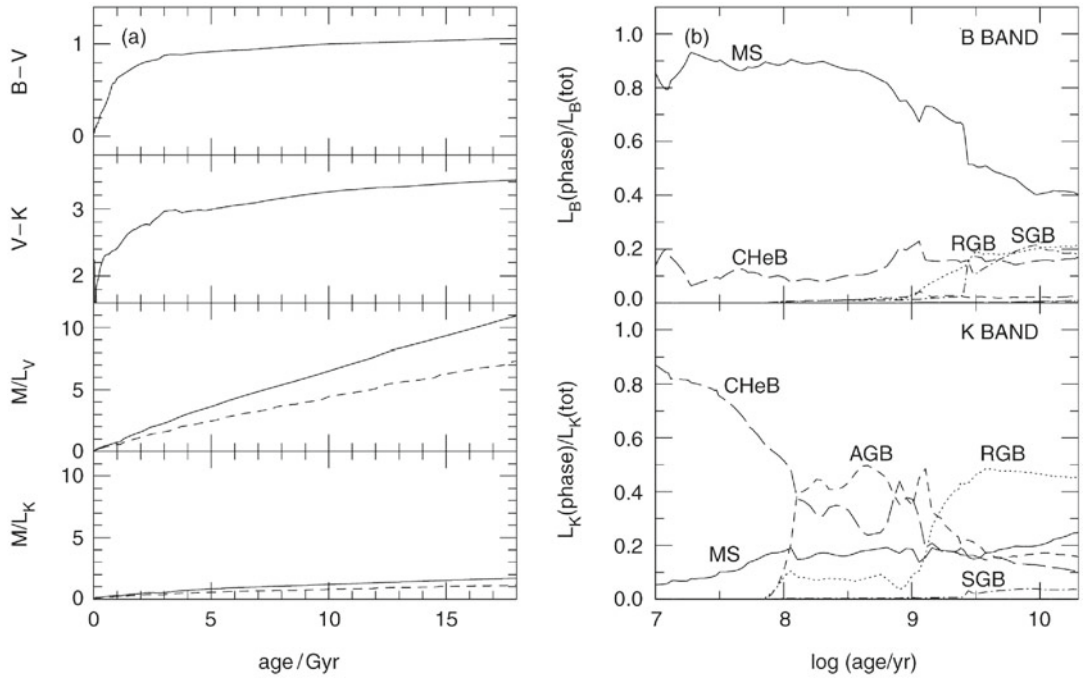


Fig. 3.34 (**a**) For the same stellar population as in Fig. 3.33b, the upper two graphs show the colors $B-V$ and $V-K$ as a function of age. The *lower* two graphs show the mass-to-light ratio M/L in two color bands in Solar units. The *solid curves* show the total M/L (i.e., including the mass that is later returned into the ISM), whereas the *dashed curves* show the M/L of the stars itself. (**b**) The fraction of B- (*top*) and K-luminosity (*bottom*) contributed by stars in their different phases of stellar evolution (MS: main sequence; CHeB: core helium burning stars; SGB: sub-giant branch). Source: S. Charlot 1996, *Spectral Evolution of Galaxies*, Lecture Notes in Physics 470, Springer-Verlag, p. 53

basic aspects of a stellar population. Results of this model are plotted in Fig. 3.35a in a color-color diagram.

From the diagram we find that the colors of the population depend strongly on τ. Specifically, galaxies do not become very red if τ is large because their star formation rate, and thus the fraction of massive blue stars, does not decrease sufficiently. The colors of Sb spirals, for example, are not compatible with a constant star formation rate—except if the total light of spirals is strongly reddened by dust absorption (but there are good reasons why this is not the case). To explain the colors of early-type galaxies we need $\tau \lesssim 4 \times 10^9$ yr. In general, one deduces from these models that a substantial evolution to redder colors occurs for $t \gtrsim \tau$. Since the luminosity of a stellar population in the blue spectral range decreases quickly with the age of the population, whereas increasing age affects the red luminosity much less, we conclude:

> The spectral distribution of galaxies is mainly determined by the ratio of the star formation rate today to the mean star formation rate in the past, $\psi(\text{today})/\langle\psi\rangle$.

One of the achievements of this standard model is that it explains the colors of present day galaxies, which have an age $\gtrsim$ 10 billion years. However, this model is not unambiguous because other star formation histories $\psi(t)$ can be constructed with which the colors of galaxies can be modeled as well.

3.5.5 Metallicity, dust, and HII regions

Predictions of the model depend on the metallicity Z—see Fig. 3.35b. A small value of Z results in a bluer color and a smaller M/L ratio in the V band. The age and metallicity of a stellar population are degenerate in the sense that an increase in the age by a factor X is nearly equivalent to an increase of the metallicity by a factor $0.65X$ with respect to the color of a population. The age estimate of a population from observed colors therefore strongly depends on the assumed value for Z. However, this degeneracy may be broken by taking several colors, or information from absorption-line spectroscopy, into account.

Intrinsic dust absorption will also change the colors of a population. This effect cannot be easily accounted for in the models because it depends not only on the properties of the dust but also on the geometric distribution of dust and stars. For example, it makes a difference whether the dust in a galaxy is homogeneously distributed or concentrated in a thin disk. Empirically, it is found that galaxies show

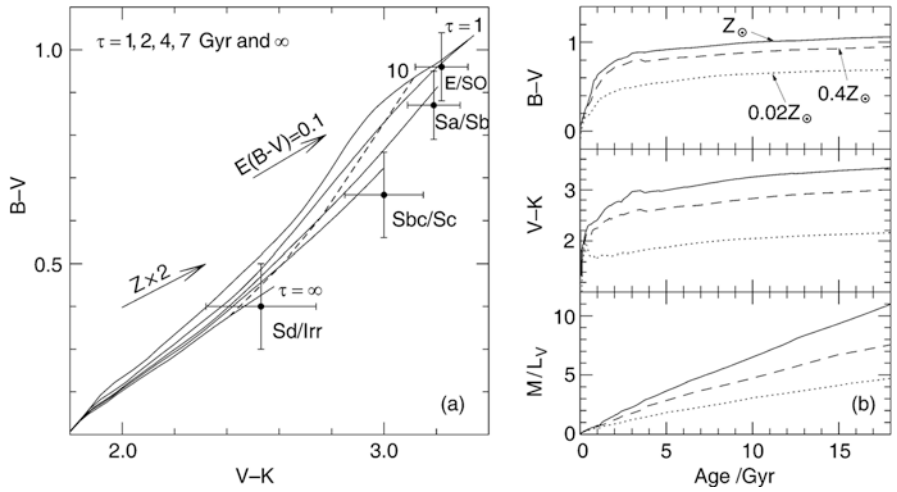


Fig. 3.35 (**a**) Evolution of colors between $0 \leq t \leq 17 \times 10^9$ yr for a stellar population with star-formation rate given by (3.38), for five different values of the characteristic time scale τ ($\tau = \infty$ is the limiting case for a constant star formation rate)—see solid curves. The typical colors for four different morphological types of galaxies are plotted. For each τ, the evolution begins at the *lower left*, i.e., as a blue population in both color indices. In the case of constant star formation, the population never becomes redder than Irr's; to achieve redder colors, τ has to be smaller. The *dashed line* connects points of $t = 10^{10}$ yr on the different curves. Here, a Salpeter IMF and Solar metallicity was assumed. The shift in color obtained by doubling the metallicity is indicated by an *arrow*, as well as that due to an extinction coefficient of $E(B-V) = 0.1$; both effects will make galaxies appear redder. (**b**) The dependence of colors and M/L on the metallicity of the population. Source: S. Charlot 1996, *Spectral Evolution of Galaxies*, Lecture Notes in Physics 470, Springer-Verlag, p. 53

strong extinction during their active phase of star formation, whereas normal galaxies are less affected by extinction, with early-type galaxies (E/S0) affected the least.

From the optical luminosity and colors of a stellar population, one can estimate its stellar mass. Obviously, the stellar mass is the product of the mass-to-light ratio and the luminosity, and M/L can be estimated from the broad-band color—see Fig. 3.34. The presence of dust of course affects these estimates. However, its two effects conspire in a particular way: dust reduces the luminosity that escapes from a stellar population, and reddens the optical light. The latter effect thus leads to the larger estimate of the mass-to-light ratio. As a reasonable approximation, the reduction of the luminosity and the increase of the M/L-estimate compensate such that the estimated stellar mass is fairly insensitive to the presence of dust.

Besides stellar light, the emission by HII regions also contributes to the light of galaxies. It is found, though, that after $\sim 10^7$ yr the emission from gas nebulae only marginally contributes to the broad-band colors of galaxies. However, this nebular radiation is the origin of emission lines in the spectra of galaxies. Therefore, emission lines are used as diagnostics for the star formation rate and the metallicity in a stellar population.

Whereas one might expect that stellar evolution is well understood, as well as the theory of stellar atmospheres where the radiation emitted from stars is formed, the models of population synthesis are still in a state of development. Fairly recently, it was found that a specific type of star—the thermally pulsating AGB stars (TP-AGB)—can significantly affect the emission of a stellar population with ages between 0.5 and 2 Gyr. Depending on how the contribution from these stars are treated in population synthesis models, the resulting predicted spectral flux, and integrated galaxy luminosity, of single-age stellar populations in the above age range can differ substantially. The effect is much smaller in stellar populations with a slowly varying formation rate; therefore, the uncertainties mainly concern post-starburst stellar populations. Hence, if a galaxy is observed about 1 Gyr after a starburst, the estimated amount of star formation during this burst depends on how the population synthesis model treats the effects of TP-AGB stars.

3.5.6 The spectra of galaxies

At the end of this section we shall consider the typical spectra of different galaxy types. They are displayed for six galaxies of different Hubble types in Fig. 3.36. To make it easier to compare them, they are all plotted in a single diagram where the logarithmic flux scale is arbitrarily normalized (since this normalization does not affect the shape of the spectra).

It is easy to recognize the general trends in these spectra: the later the Hubble type, (1) the bluer the overall spectral distribution, (2) the stronger the emission lines, (3) the weaker the absorption lines, and (4) the smaller the 4000 Å-break in the spectra. From the above discussion, we would also expect these trends if the Hubble sequence is considered an ordering of galaxy types according to the characteristic age of their stellar population or according to their star-

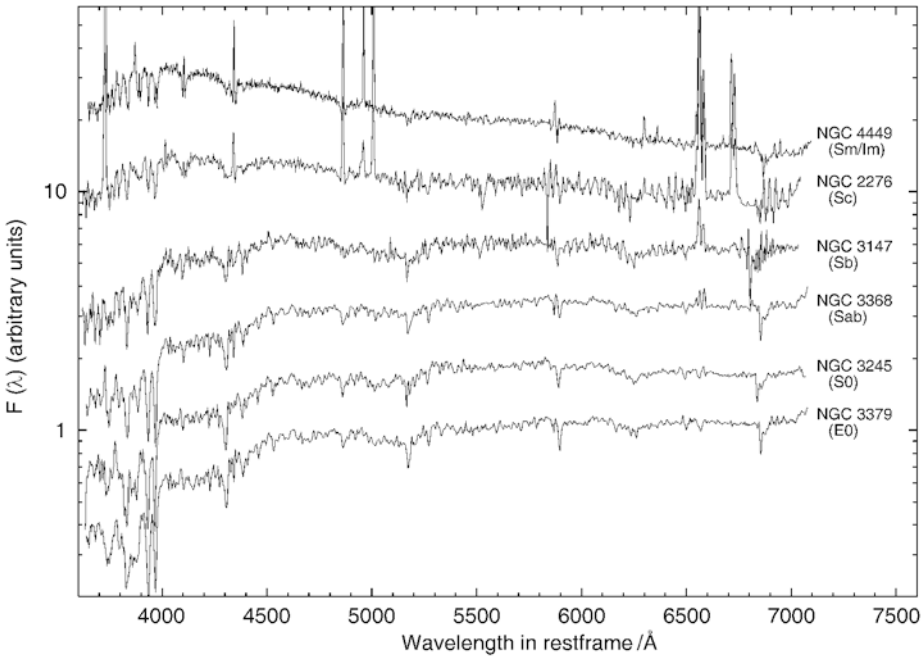


Fig. 3.36 Spectra of galaxies of different types, where the spectral flux is plotted logarithmically in arbitrary units. The spectra are ordered according to the Hubble sequence, with early types at the *bottom* and late-type spectra at the *top*. Data from R. Kennicutt 1992, ApJS 79, 255

formation rate. Elliptical and S0 galaxies essentially have no star formation activity, which renders their spectral energy distribution dominated by red stars. Furthermore, in these galaxies there are no HII regions where emission lines could be generated. The old stellar population produces a pronounced 4000 Å break, which corresponds to a jump by a factor of ~ 2 in the spectra of early-type galaxies. It should be noted that the spectra of ellipticals and S0 galaxies are quite similar.

By contrast, Sc spirals and irregular galaxies have a spectrum which is dominated by emission lines, where the Balmer lines of hydrogen as well as nitrogen and oxygen lines are most pronounced. The relative strength of these emission lines are characteristic for HII-regions, implying that most of this line emission is produced in the ionized regions surrounding young stars. For irregular galaxies, the spectrum is nearly totally dominated by the stellar continuum light of hot stars and the emission lines from HII-regions, whereas clear contributions by cooler stars can be identified in the spectra of Sc spiral galaxies.

The spectra of Sa and Sb galaxies form a kind of transition between those of early-type galaxies and Sc galaxies. Their spectra can be described as a superposition of an old stellar population generating a red continuum with absorption features and a young population with its blue continuum and its emission lines. This can be seen in connection with the decreasing contribution of the bulge to the galaxy luminosity towards later spiral types.

The properties of the spectral light distribution of different galaxy types, as briefly discussed here, is described and interpreted in the framework of population synthesis. This gives us a detailed understanding of stellar populations as a function of the galaxy type. Extending these studies to spectra of high-redshift galaxies allows us to draw conclusions about the evolutionary history of their stellar populations.

3.5.7 Summary

After this somewhat lengthy section, we shall summarize the most important results of population synthesis here:

- A simple model of star formation history reproduces the colors of today's galaxies fairly well.
- (Most of) the stars in elliptical and S0 galaxies are old—the earlier the Hubble type, the older the stellar population.
- Detailed models of population synthesis provide information about the star formation history, and predictions by the models can be compared with observations of galaxies at high redshift (and thus smaller age).

We will frequently refer to results from population synthesis in the following chapters. For example, we will use them to interpret the colors of galaxies at high redshifts and the different spatial distributions of early-type and late-type galaxies (see Chap. 6). Also, we will present a method of estimating the redshift of galaxies from their broad-band colors (photometric redshifts). As a special case of this method, we will discuss the efficient selection of galaxies at very high redshift (Lyman-break galaxies, LBGs, see Chap. 9). Because the color and luminosity of a galaxy are changing even when no star formation is taking place, tracing back such a *passive evolution* allows us to distinguish this passive aging process from episodes of star formation and other processes.

3.6 The population of luminous galaxies

We started this chapter with the classification of galaxies, according to morphology and according to their colors. After discussing the properties of elliptical and spiral galaxies in some detail, we are now ready to ask the obvious question: what is the relation between ellipticals and spirals on the one hand, and red and blue galaxies on the other? How are these two classification schemes related? Furthermore, we may look for other global properties of galaxies that either correlate strongly with color, or with morphology.

The Sérsic brightness profile. As we have seen, the brightness distribution of disks follows in general an exponential profile, whereas bulges of disk galaxies and the light profile of ellipticals are better described with a de Vaucouleurs profile. Thus, the brightness profiles of galaxies are expected to correlate well with their morphological type. J. Sérsic introduced the brightness profile

$$\log\left(\frac{I(R)}{I_\mathrm{e}}\right) = -b_n\left[\left(\frac{R}{R_\mathrm{e}}\right)^{1/n} - 1\right], \qquad (3.39)$$

hence called *Sérsic brightness profile*, where n is called the *Sérsic index*. As was the case for the de Vaucouleurs profile, the effective radius R_e is chosen such that half of the luminosity comes from within the circle of radius R_e. The coefficient b_n must be chosen such that this property is fulfilled; to good approximation, one finds $b_n \approx 1.999n - 0.327$. I_e is the surface brightness at R_e. If $n = 4$, (3.39) reduces to the de Vaucouleurs law, whereas for $n = 1$, an exponential surface brightness distribution is obtained; in this way, the Sérsic law provides a generalization of, and includes these two brightness profiles. The larger n, the more concentrated the light profile is in the central part, and at the same time, the higher is the surface brightness for large R; see Fig. 3.37.

The Sérsic profile provides a convenient parametrization of the brightness profiles of galaxies, and can be used to classify them, by getting the best fit of their light profile with (3.39). In this way, n, R_e and I_e is obtained. The fit is not expected to be a good one in all cases; for example, if one considers an Sa spiral galaxy, for which the bulge contributes substantially to the total light, a single Sérsic profile for both the (de Vaucouleurs) bulge and the (exponential) disk will not necessarily provide an accurate fit. In this case, one would expect that n lies between 1 and 4, depending on the relative strength of the bulge. Indeed, n correlates well with the bulge-to-disk ratio of galaxies. For nearby galaxies, where the brightness profile is well resolved, multi-component Sérsic models are usually fitted, with an inner one for describing the bulge component and the outer one fitting the disk.

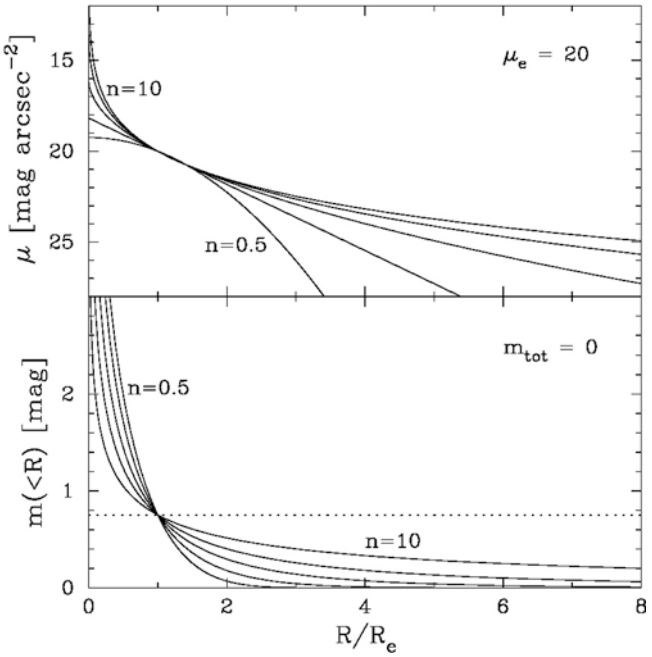


Fig. 3.37 The Sérsic profile, plotted for various values of n. In the *upper panel*, the surface brightness is plotted as a function of R/R_e, where all profiles are chosen to have the same brightness at R_e. The *straight line* (second from bottom) is the exponential profile, $n = 1$. In the *lower panel*, the enclosed flux within R is displayed, again normalized such that all profiles agree at R_e, which is equivalent to say that all profiles have the same total magnitude. Source: A.W. Graham & S.P. Driver 2005, *A concise reference to (projected) Sersic $R^{1/n}$ quantities, including Concentration, Profile Slopes, Petrosian indices, and Kron Magnitudes*, astro-ph/0503176; Credit: NASA/JPL-Caltech

Photometric properties of local galaxies. The SDSS provided the first very large survey of galaxies with homogeneous photometry and spectroscopy, in particular redshift information. Therefore, this survey allowed us to study the statistical properties of galaxy properties in great detail. In Fig. 3.38, the distribution of $\sim 77\,000$ galaxies with $z \leq 0.05$ is shown in terms of photometric parameters, characterizing the luminosity, color, size, and brightness profile of these galaxies. The distribution in absolute magnitude, shown in the lower left panel, indicates that the galaxy sample becomes incomplete for objects less luminous than $M_r \sim -19$, owing to the flux limit of the spectroscopic sample in the SDSS.[7] Lower luminosity galaxies are in the sample only if they are very close to us.

[7]The SDSS spectroscopic sample is flux limited, i.e., it contains (almost) all galaxies in its sky region with a flux $S > S_\mathrm{lim}$. If we restrict the sample to a maximum distance D_max, then the sample is also complete for luminosities $L > 4\pi S_\mathrm{lim} D_\mathrm{max}^2$.

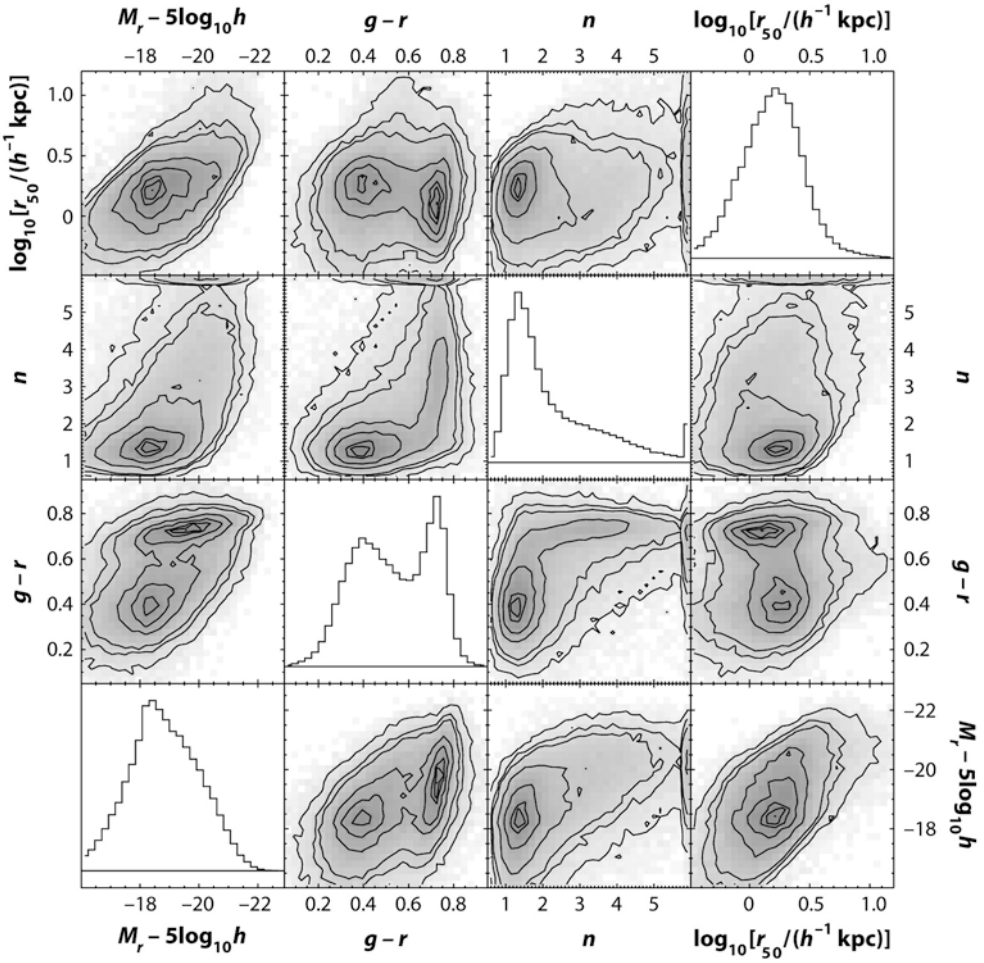


Fig. 3.38 The distribution of photometric properties of galaxies, as obtained from the Sloan Digital Sky Survey. The greyscales and contours in the off-diagonal panels show the number of galaxies in each two-dimensional bin; the darker the bins, the higher is the galaxy number. These distributions are shown for six combinations of the four photometric parameters: absolute magnitude in the r-band M_r (note that for a dimensionless Hubble constant of $h = 0.71$, $5\log h \approx -0.74$), the color $g - r$, the Sérsic index n, and the effective radius, here called r_{50}. Note that panels in the *upper left* part are just mirror images of those in the *lower right* part. The panels on the diagonal show the number distribution of galaxies with respect to the four photometric parameters. Source: M.R. Blanton & J. Moustakas, 2009, *Physical Properties and Environments of Nearby Galaxies*, ARA&A 47, 159, p. 162, Fig. 1. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 47 ©2009 by Annual Reviews www.annualreviews.org

Red sequence, blue cloud, and green valley. The color-magnitude plot (left column, third panel from top) shows essentially the same distribution as in Fig. 3.7, except that here the $g-r$ color is shown (and the absolute magnitude axis is reversed). Galaxies show a clearly bimodal distribution in this space, with two peaks corresponding to luminous red galaxies, and less luminous blue galaxies; these are often called *red-sequence galaxies* and *blue-cloud galaxies*, respectively. This bimodality is also seen in the distribution of galaxy colors, shown by the histogram in the second diagonal panel from the bottom left. The spread of red-sequence galaxies in color is much less than that of the blue-cloud galaxies. Thus, the color of red galaxies is very well defined, also seen by the small slope of their distribution in the color-magnitude diagram. The galaxies located between the red sequence and the blue cloud are called *green-valley galaxies*.

This result is not so surprising after the discussion in Sect. 3.5: the (red) color of an old stellar population depends only little on its exact age. The interpretation of the narrow distribution of the red sequence in color is thus that these galaxies have an old stellar population, with essentially no or very little recent star formation. The slight trend of somewhat redder colors for more luminous galaxies within the red sequence, seen in the color-magnitude diagram, can be due to somewhat higher ages of more luminous galaxies, or higher metallicities. The spread in color of the blue-cloud galaxies presumably reflect different levels of star-formation activities, leading to different mean stellar ages of the population. Here the trend in the color-magnitude diagram is much stronger than for the red sequence: the characteristic color of blue-cloud galaxies correlates significantly with luminosity, in that more luminous galaxies tend to be redder than less luminous ones.

Interestingly, the Sérsic index n correlates clearly with galaxy color, as shown by the second panel in the second row of Fig. 3.38. For blue galaxies, it is strongly concentrated around $n \approx 1$, corresponding to an exponential brightness profile. For red galaxies, there is a much broader distribution; in particular, the de Vaucouleurs value $n = 4$ is not singled out. There is a clear trend that more luminous galaxies are more concentrated than less luminous ones. This behavior holds for the galaxy population as a whole, as well as separately for the blue-cloud and red-sequence galaxies. Consequently, n increases towards redder galaxies on the red sequence. The effective radius correlates strongly with

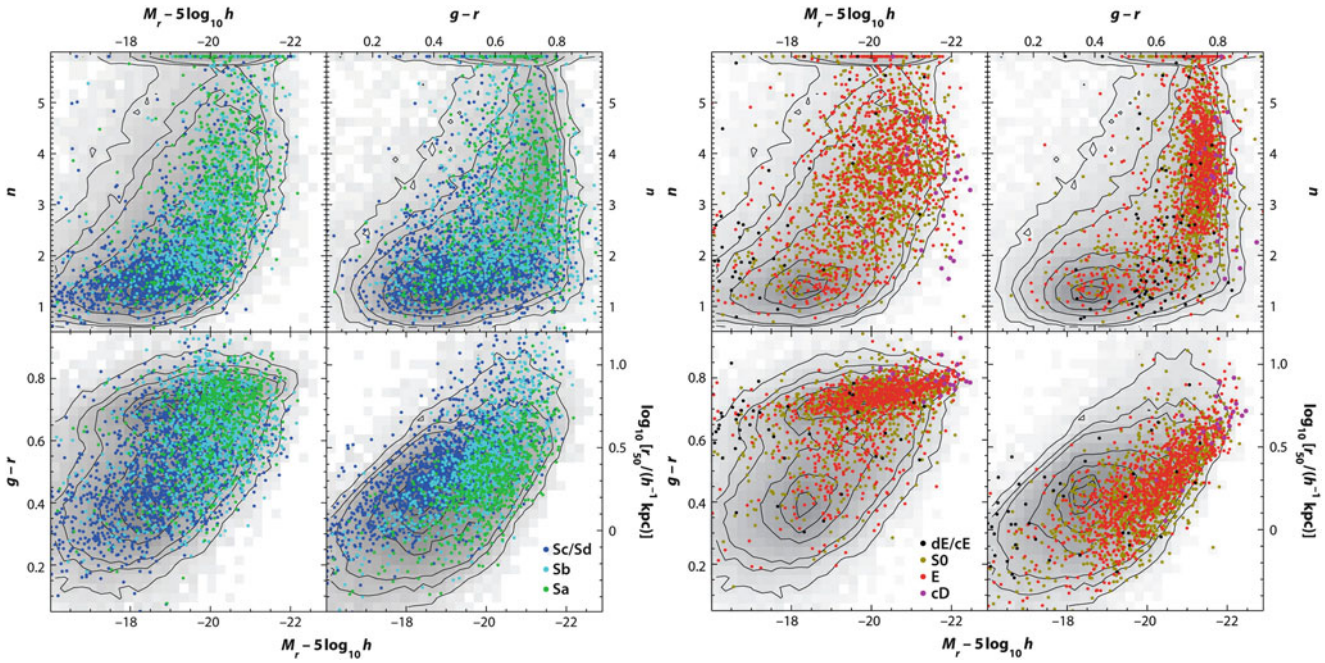


Fig. 3.39 Distribution of galaxies that are morphologically classified as spirals (*left*) and early-type galaxies (*right*), in the same parameter space as in Fig. 3.38. The greyscale and contours are same as in Fig. 3.38. Different types of galaxies are distinguished by differently colored points. Source: M.R. Blanton & J. Moustakas, 2009, *Physical Properties and Environments of Nearby Galaxies*, ARA&A 47, 159, p. 174, 186, Fig. 8, 12. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 47 ©2009 by Annual Reviews www.annualreviews.org

galaxy luminosity, in that more luminous galaxies are more extended.

How does this relate to the Hubble sequence? For a subset of the nearby galaxies in SDSS, a classification by optical morphology is available. These galaxies are plotted in the same photometric parameter space as considered before in Fig. 3.39, where on the left-hand side, spiral galaxies are shown, and early-type galaxies on the right-hand side. The underlying gray-scale and contours are identical with those in Fig. 3.38, showing the distribution of the whole galaxy population.

Considering the color-magnitude diagram (lower left panels) first, we see that the overwhelming majority of early-type galaxies lie on the red sequence, and those that do not are mostly dwarf galaxies. Very few early-type galaxies are located in the blue cloud, which implies that the blue-cloud galaxies are essentially all (star-forming) spirals. This behavior was expected, given our earlier discussion of elliptical galaxies in which ellipticals were described as objects with essentially no ongoing star formation, and consequently red colors.

Surprisingly, the converse is not true: not all red-sequence galaxies are early types. In fact, spiral galaxies are not confined to the blue cloud, but occupy a rather extended region in the color-magnitude diagram, with a clear dependence on type: late-type spirals (Sc/Sd) are on average bluer and less luminous than earlier types. This is expected, based on the relative importance of the (red) bulge compared to the (blue) disk. The influence of relative bulge strength can also be seen in the distribution with respect to Sérsic index n, which clusters around $n \sim 1$ for late-type spirals, but has a broad distribution for earlier types. However, the bulge-to-disk ratio is not the only quantity that determines the color and Sérsic index of spirals. As is seen from the distribution of n for early-type galaxies, it is *not* true that the spheroidal component of galaxies is generally described by a de Vaucouleurs profile with $n = 4$, but n varies substantially among ellipticals. Furthermore, the color of spirals can be substantially affected by dust in their ISM.

The distribution of galaxies with respect to their effective radius (lower right panels) shows a clear tendency—the earlier the galaxy type, the smaller is the effective radius at fixed absolute magnitude. This behavior is not independent of the relation between galaxy types and Sérsic index, with later types having smaller n. Since n describes the concentration of the light distribution, the light of early-type galaxies (with larger n) is more concentrated than that of later types, implying a smaller effective radius at fixed luminosity.

The color-magnitude distribution of S0 galaxies is almost indistinguishable from that of elliptical galaxies, whereas their concentration index n is clearly smaller than that of E's. The latter is expected due to their disk component.

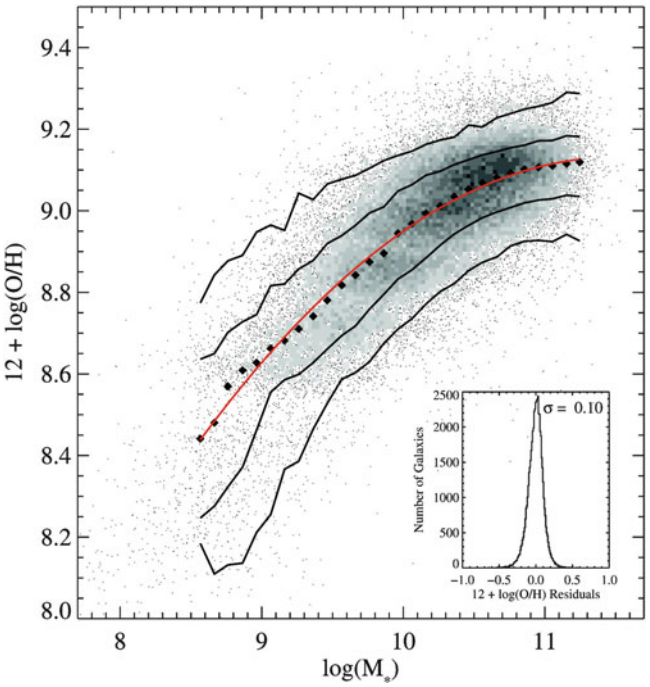


Fig. 3.40 For some 53 000 star-forming galaxies in the SDSS, the oxygen abundance of the interstellar medium is plotted against the stellar mass M_*. The density of points is indicated by the *greyscales* in high-density regions, and individual galaxies are plotted in the less populated parts of the diagram. The *black diamonds* represent the median of the distribution of [O/H] in small bins of M_*, whereas the *red curve* shows a fit through these data points. The *solid black curves* enclose 68 and 95 % of the distribution. There is an unambiguous trend of increased metallicity with stellar mass. The *small inset* shows the distribution of galaxy metallicities around the fit curve; the dispersion of galaxy metallicity around this mean curve is fairly small. Above $M_* \sim 3 \times 10^{10} M_\odot$, the relation between metallicity and stellar mass seems to saturate. Source: C.A. Tremonti et al. 2004, *The Origin of the Mass-Metallicity Relation: Insights from 53,000 Star-forming Galaxies in the Sloan Digital Sky Survey*, ApJ 613, 898, p. 907, Fig. 6. ©AAS. Reproduced with permission

Metallicity. Owing to the availability of spectral information for the SDSS galaxies, the distribution of spectral properties of galaxies can be studied. The metallicity of galaxies, as determined from their relative oxygen abundance [O/H], correlates very strongly with stellar mass, in that more massive galaxies contain more metals (see Fig. 3.40). In particular, the dispersion of galaxy metallicity around the mean trend is fairly small. We will see in Sect. 3.7 that this correlation is expected from models of the chemical evolution of galaxies.

Most interesting is the bimodal distribution of galaxies with respect to their 4000 Å-break, which is a reliable indicator of the luminosity-weighted mean age of a stellar population (see Sect. 3.5.2). Galaxies with a stellar mass $\gtrsim 1.5 \times 10^{10} h^{-2} M_\odot$ typically have a strong 4000 Å-break, whereas lower-mass galaxies have a weak break. The bimodality in this distribution is very similar to that in the color-magnitude diagram, and presumably has the same origin: more massive galaxies have an older stellar population, which renders them redder and having a stronger 4000 Å-break.

Summary. The statistical investigation of a large sample of luminous nearby galaxies yielded remarkable insights into properties of the galaxy population.

> The distribution of galaxies is well structured. Instead of filling the possible space of photometric and spectroscopic parameters more or less uniformly, they are confined largely to well-defined sequences. Indeed, it seems that their properties are largely determined by their luminosity or stellar mass: the more massive a galaxy is, the more likely it is that it is red, has a strong 4000 Å-break, a large Sérsic index, high metallicity, and a large effective radius. Conversely, less luminous galaxies are mostly blue, show a weak 4000 Å-break, have $n \sim 1$, smaller metallicity and smaller effective radius.

There is a more or less strong dispersion of galaxy properties around these general trends, which is probably caused by differences in the recent star-formation activity and different dust contents; in addition, the color of spiral galaxies depends somewhat on their inclination, as discussed in Sect. 3.3.5. The variation of galaxy properties is considerably larger for spirals, whereas early-type galaxies form a fairly uniform population. The dependence of the galaxy population on environmental effects will be discussed in Sect. 6.7.

3.7 Chemical evolution of galaxies

During its evolution, the chemical composition of a galaxy changes. Thus the observed metallicity yields information about the galaxy's star formation history. We expect the metallicity Z to increase with star-formation rate, integrated over the lifetime of the galaxy. We will now discuss a simple model of the chemical evolution of a galaxy, which will provide insight into some of the principal aspects.

We assume that at the formation epoch of the stellar population of a galaxy, at time $t = 0$, no metals were present; hence $Z(0) = 0$. Furthermore, the galaxy did not contain any stars at the time of its birth, so that all baryonic matter was in the form of gas. In addition, we consider the galaxy as a closed system out of which no matter can escape or be added later on by processes of accretion or merger. Finally, we assume that the time-scales of the stellar evolution processes that lead to the metal enrichment of the galaxy are small compared to the evolutionary time-scale of the galaxy. Under

these assumptions, we can now derive a relation between the metallicity and the gas content of a galaxy.

Of the total mass of a newly formed stellar population, part of it is returned to the ISM by supernova explosions and stellar winds. We define this fraction as R, so that the fraction $\alpha = (1 - R)$ of a newly-formed stellar population remains enclosed in stars, i.e., it no longer takes part in the further chemical evolution of the ISM. The value of α depends on the IMF of the stellar population and can be computed from models of population synthesis. Furthermore, let q be the ratio of the mass in metals, which is produced by a stellar population and then returned into the ISM, and the initial total mass of the population. The *yield* $y = q/\alpha$ is defined as the ratio of the mass in metals that is produced by a stellar population and returned into the ISM, and the mass that stays enclosed in the stellar population. The yield can also be calculated from population synthesis models. If $\psi(t)$ is the star formation rate as a function of time, then the mass of all stars formed in the history of the galaxy is given by

$$S(t) = \int_0^t \mathrm{d}t' \; \psi(t') \; ,$$

and the total mass that remains enclosed in stars is $s(t) = \alpha S(t)$. Since we have assumed a closed system for the baryons, the sum of gas mass $g(t)$ and stellar mass $s(t)$ is a constant, namely the baryon mass of the galaxy,

$$g(t) + s(t) = M_\mathrm{b} \quad \Rightarrow \quad \frac{\mathrm{d}g}{\mathrm{d}t} + \frac{\mathrm{d}s}{\mathrm{d}t} = 0 \; . \tag{3.40}$$

The mass of the metals in the ISM is gZ; it changes when stars are formed. Through this formation, the mass of the ISM and thus also that of its metals decreases. On the other hand, metals are also returned into the ISM by processes of stellar evolution. Under the above assumption that the time scales of stellar evolution are small, this return occurs virtually instantaneously. The metals returned to the ISM are composed of metals that were already present at the formation of the stellar population—a fraction R of these will be returned—and newly formed metals. Together, the total mass of the metals in the ISM obeys the evolution equation

$$\frac{\mathrm{d}(gZ)}{\mathrm{d}t} = \psi \, (RZ + q) - Z\psi \; ,$$

where the last term specifies the rate of the metals extracted from the ISM in the process of star formation and the first term describes the return of metals to the ISM by stellar evolution processes. Since $\mathrm{d}S/\mathrm{d}t = \psi$, this can also be written as

$$\frac{\mathrm{d}(gZ)}{\mathrm{d}S} = (R-1)Z + q = q - \alpha Z \; .$$

Dividing this equation by α and using $s = \alpha S$ and the definition of the yield, $y = q/\alpha$, we obtain

$$\frac{\mathrm{d}(gZ)}{\mathrm{d}s} = \frac{\mathrm{d}g}{\mathrm{d}s} Z + g \frac{\mathrm{d}Z}{\mathrm{d}s} = y - Z \; . \tag{3.41}$$

From (3.40) it follows that $\mathrm{d}g/\mathrm{d}s = -1$ and $\mathrm{d}Z/\mathrm{d}s = -\mathrm{d}Z/\mathrm{d}g$, and so we obtain a simple equation for the metallicity,

$$\begin{aligned} g \frac{\mathrm{d}Z}{\mathrm{d}g} &= \frac{\mathrm{d}Z}{\mathrm{d}\ln g} = -y \\ \Rightarrow \quad Z(t) &= -y \, \ln\left(\frac{g(t)}{M_\mathrm{b}}\right) = -y \, \ln(\mu_\mathrm{g}) \; , \end{aligned} \tag{3.42}$$

where $\mu_\mathrm{g} = g/M_\mathrm{b}$ is the fraction of baryons in the ISM, and where we chose the integration constant such that at the beginning, when $\mu_\mathrm{g} = 1$, the metallicity was $Z = 0$. From this relation, we can now see that with decreasing gas content in a galaxy, the metallicity will increase; in our simple model this increase depends only on the yield y. Since y can be calculated from population synthesis models with a typical value of $y \sim 10^{-2}$, (3.42) is a well-defined relation.

If (3.42) is compared with observations of galaxies, one finds that they follow the general trend predicted by (3.42). More gas-rich galaxies tend to have smaller metallicities. For example, the metallicities of Sa-galaxies is in general higher than those of later-type spirals which contain a higher gas mass fraction. However, in detail there are strong deviations from (3.42), which are particularly prominent for low-mass galaxies. While the assumption of an instantaneous evolution of the ISM is fairly well justified, we know from structure formation in the Universe (Chap. 7) that galaxies are by no means isolated systems: their mass continuously changes through accretion and merging processes. In addition, the kinetic energy transferred to the ISM by supernova explosions causes an outflow of the ISM, in particular in low-mass galaxies where the gas is not strongly gravitationally bound. These outflows are directly observed in terms of galactic winds from star-forming galaxies, and may explain the deviations from (3.42) by up to a factor of 10 for the low-mass galaxies. An analysis of star-forming galaxies in the SDSS indicates that galaxies with stellar masses below $\sim 4 \times 10^9 M_\odot$ can lose more than half their metals by outflows. Of course, the observed deviations from relation (3.42) allow us to draw conclusions about these accretion and wind processes.

Also, from observations in our Milky Way we find indications that the model of the chemical evolution sketched above is too simplified. This is known as the *G-dwarf problem*. The model described above predicts that about half of the F- and

G-main sequence stars should have a metallicity of less than a quarter of the Solar value. These stars have a long lifetime on the main sequence, so that many of those observed today should have been formed in the early stages of the Galaxy. Thus, in accordance with our model they should have very low metallicity. However, a low metallicity is in fact observed in only very few of these stars. The discrepancy is far too large to be explained by selection effects. Rather, observations show that the chemical evolution of our Galaxy must have been substantially more complicated than described by our simple model. Indeed, we saw in Sect. 2.3.7 that there is clear evidence for infalling gas towards the Galactic plane, providing new and rather low-metallicity material for star formation.

3.8 Black holes in the centers of galaxies

As we have seen in Sect. 2.6.3, the Milky Way harbors a black hole in its center. Furthermore, it is generally accepted that the energy for the activity of AGNs is generated by accretion onto a black hole (see Sect. 5.3). Thus, the question arises as to whether all (or most) galaxies contain a supermassive black hole (SMBH) in their nuclei. We will pursue this question in this section and show that SMBHs are very abundant indeed. This result then instigates further questions: what distinguishes a 'normal' galaxy from an AGN if both have a SMBH in the nucleus? Is it the mass of the black hole, the rate at which matter is accreted onto it, or the efficiency of the mechanism which is generating the energy?

We will start with a concise discussion of how to search for SMBHs in galaxies, then present some examples for the discovery of such SMBHs. Finally, we will discuss the very tight relationship between the mass of the SMBH and the properties of the stellar component of a galaxy.

3.8.1 The search for supermassive black holes

What is a black hole? A technical answer is that a black hole is the simplest solution of Einstein's theory of general relativity which describes the gravitational field of a point mass. Less technically—though sufficient for our needs—we may say that a black hole is a point mass, or a compact mass concentration, with an extent smaller than its Schwarzschild radius r_S (see below).

The Schwarzschild radius. The first discussion of black holes can be traced back to Laplace in 1795, who considered the following: if one reduces the radius r of a celestial body of mass M, the escape velocity v_{esc} at its surface,

$$v_{esc} = \sqrt{\frac{2GM}{r}} ,$$

will increase. As a thought experiment, one can now see that for a sufficiently small radius, v_{esc} will be equal to the speed of light, c. This happens when the radius decreases to

$$\boxed{r_S := \frac{2GM}{c^2} = 2.95 \times 10^5 \,\mathrm{cm} \left(\frac{M}{M_\odot} \right)} . \qquad (3.43)$$

The radius r_S is named the *Schwarzschild radius*, after Karl Schwarzschild who, in 1916, discovered the point-mass solution of Einstein's field equations. For our purpose we will define a black hole as a mass concentration with a radius smaller than r_S. As we can see, r_S is very small: about 3 km for the Sun, and $r_S \sim 10^{12}$ cm for the SMBH in the Galactic center. At a distance of $D = R_0 \approx 8$ kpc, this corresponds to an angular radius of $\sim 8 \times 10^{-6}$ arcsec. Current observing capabilities are still far from resolving scales of order r_S, except for the VLBI technique which currently comes close to it: The highest angular resolution currently achieved with millimeter-VLBI is a mere factor of ~ 10 away from resolving the Schwarzschild radius for the Galactic black hole that is supposed to coincide with the compact radio source Sgr A*. By performing VLBI studies at sub-millimeter wavelengths in the near future, we may actually be able to 'see' the Schwarzschild radius of a black hole for the first time. The largest observed velocities of stars in the Galactic center, ~ 5000 km/s $\ll c$, indicate that they are still well away from the Schwarzschild radius. We will show in Sect. 5.3.3 that relativistic effects are directly observed in AGNs and that velocities close to c do in fact occur there—which again is a very direct indication of the existence of a SMBH.

If even for the closest SMBH, the one in the GC, the Schwarzschild radius is significantly smaller than the achievable angular resolution, how can we hope to prove that SMBHs exist in other galaxies? Like in the GC, this proof has to be found indirectly by detecting a compact mass concentration incompatible with the mass concentration of the stars observed.

The radius of influence. We consider a mass concentration of mass $M_\bullet$ in the center of a galaxy where the characteristic velocity dispersion of stars (or gas) is σ_v. We compare this velocity dispersion with the characteristic velocity (e.g., the Kepler rotational velocity) around a SMBH at a distance r, given by $\sqrt{GM_\bullet/r}$. From this it follows that, for distances smaller than

$$r_{BH} = \frac{GM_\bullet}{\sigma_v^2} \sim 0.4 \left(\frac{M_\bullet}{10^6 M_\odot} \right) \left(\frac{\sigma_v}{100\,\mathrm{km/s}} \right)^{-2} \mathrm{pc} , \qquad (3.44)$$

the SMBH will significantly affect the kinematics of stars and gas in the galaxy. The corresponding angular scale is

$$\theta_{\rm BH} = \frac{r_{\rm BH}}{D} \sim 0''.1 \left(\frac{M_\bullet}{10^6 M_\odot}\right) \left(\frac{\sigma_v}{100\,{\rm km/s}}\right)^{-2} \left(\frac{D}{1\,{\rm Mpc}}\right)^{-1}, \tag{3.45}$$

where D is the distance of the galaxy. From this we immediately conclude that our success in finding SMBHs will depend heavily on the achievable angular resolution. The HST enabled scientists to make huge progress in this field. The search for SMBHs promises to be successful only in relatively nearby galaxies. In addition, from (3.45) we can see that for increasing distance D the mass $M_\bullet$ has to increase for a SMBH to be detectable at a given angular resolution.

Kinematic evidence. The presence of a SMBH inside $r_{\rm BH}$ is revealed by an increase in the velocity dispersion for $r \lesssim r_{\rm BH}$, which should then behave as $\sigma_v \propto r^{-1/2}$ for $r \lesssim r_{\rm BH}$. If the inner region of the galaxy rotates, one expects, in addition, that the rotational velocity $v_{\rm rot}$ should also increase inwards $\propto r^{-1/2}$.

Problems in detecting these signatures. The practical problems in observing a SMBH have already been mentioned above. One problem is the angular resolution. To measure an increase in the velocities for small radii, the angular resolution needs to be better than $\theta_{\rm BH}$. Furthermore, projection effects play a role because only the velocity dispersion of the projected stellar distribution, weighted by the luminosity of the stars, is measured. Added to this, the kinematics of stars can be rather complicated, so that the observed values for σ and $v_{\rm rot}$ depend on the distribution of orbits and on the geometry of the distribution.

Despite these difficulties, the detection of SMBHs has been achieved in recent years, largely due to the much improved angular resolution of optical telescopes (like the HST) and to improved kinematic models. Black hole masses were determined for more than 70 nearby galaxies, and upper limits on $M_\bullet$ were obtained for about 30 galaxies.

3.8.2 Examples for SMBHs in galaxies

Figure 3.41 shows an example for the kinematical method discussed in the previous section. A long-slit spectrum across the nucleus of the galaxy M84 clearly shows that, near the nucleus, both the rotational velocity (seen by the mean wavelength of the emission line) and the velocity dispersion

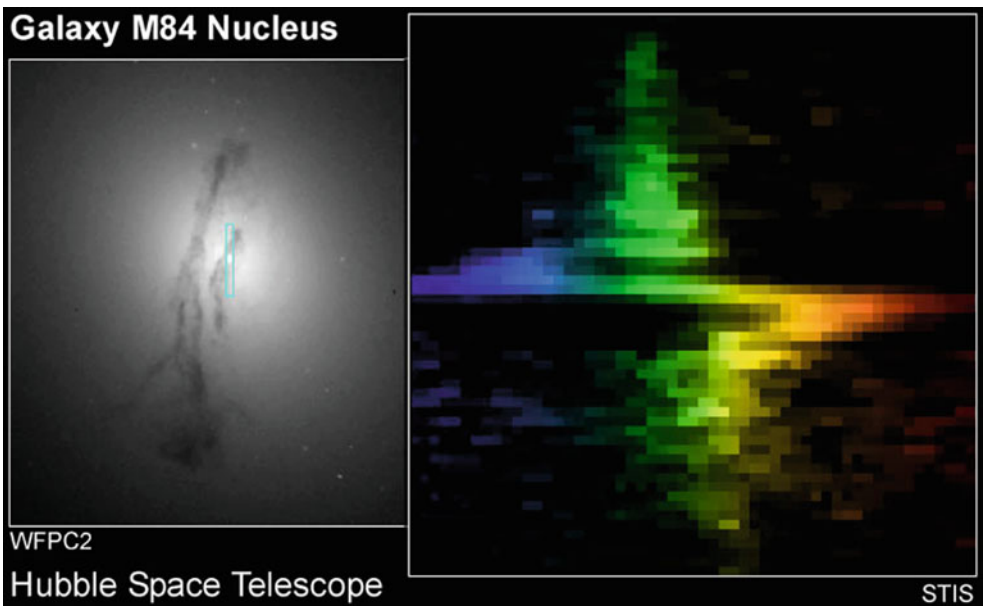


Fig. 3.41 An HST image of the nucleus of the galaxy M84 is shown in the *left-hand panel*. M84 is a member of the Virgo cluster, about 15 Mpc away from us. The *small rectangle* depicts the position of the slit used by the STIS (Space Telescope Imaging Spectrograph) instrument on-board the HST to obtain a spectrum of the central region. The spectral shape of five emission lines, as obtained from this long-slit spectrum, is shown in the *right-hand panel*; the position along the slit is plotted vertically, the relative wavelength change of the light (which is proportional to the radial velocity) horizontally, also illustrated by *colors*. Near the center of the galaxy the wavelength suddenly changes because the rotational velocity steeply increases inwards and then changes sign on the other side of the center. This shows the Kepler rotation in the central gravitational field of a SMBH, whose mass can be estimated as $M_\bullet \sim 3 \times 10^8 M_\odot$. Credit: Gary Bower, Richard Green (NOAO), the STIS Instrument Definition Team, and NASA/ESA

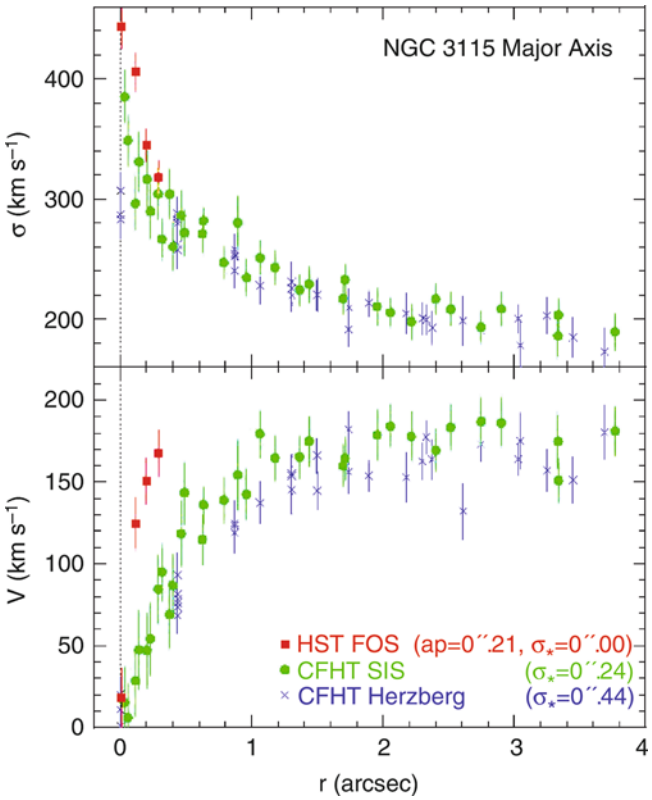


Fig. 3.42 Rotational velocity (*bottom*) and velocity dispersion (*top*) of stars, as functions of the distance from the center along the major axis of the galaxy NGC 3115. Colors of the symbols mark observations with different instruments. Results from CFHT data which have an angular resolution of 0.″44 are shown in *blue*. The SIS instrument at the CFHT uses active optics to achieve roughly twice this angular resolution; corresponding results are plotted in *green*. Finally, the *red symbols* show the result from HST observations using the Faint Object Spectrograph (FOS). As expected, with improved angular resolution an increase in the observed value of the velocity dispersion is detected towards the center. Even more dramatic is the impact of resolution on measurements of the rotational velocity. Due to projection effects, the measured central velocity dispersion is smaller than the real one; this effect can be corrected for. After correction, a central value of $\sigma \sim 600\,\mathrm{km/s}$ is found. This value is much higher than the escape velocity from the central star cluster if it were to consist solely of stars—it would dissolve within $\sim 2 \times 10^4$ yr. Therefore, an additional compact mass component of $M_\bullet \sim 10^9 M_\odot$ must exist. Source: J. Kormendy & L.C. Ho 2000, *Supermassive Black Holes in Inactive Galaxies*, astro-ph/0003268, p. 5, Fig. 2

(given by the width of the line) change; both increase dramatically towards the center. Figure 3.42 illustrates how strongly the measurability of the kinematical evidence for a SMBH depends on the achievable angular resolution of the observation. For this example of NGC 3115, observing with the resolution offered by space-based spectroscopy yields much higher measured velocities than is possible from the ground, due to the convolution with a larger point-spread function. Particularly interesting is the observation of the rotation curve very close to the center. Another impressive example is the central region of M87, the central galaxy of the Virgo cluster. The increase of the rotation curve and the broadening of the [OII]-line (a spectral line of singly-ionized oxygen) at $\lambda = 3727\,\text{Å}$ towards the center are displayed in Fig. 3.43 and argue very convincingly for a SMBH with $M_\bullet \approx 3 \times 10^9 M_\odot$.

The mapping of the Kepler rotation in the center of the Seyfert galaxy NGC 4258 is especially spectacular. This galaxy contains water masers—very compact sources whose position can be observed with very high precision using VLBI techniques (Fig. 3.44). In this case, the deviation from a Kepler rotation in the gravitational field of a point mass of $M_\bullet \sim 3.5 \times 10^7 M_\odot$ is much less than 1 %; the uncertainty in the estimated value of $M_\bullet$ is fully dominated by the uncertainty in the distance to this galaxy ($D \sim 7$ Mpc). The maser sources are embedded in an accretion disk having a thickness of less than 0.3 % of its radius, of which also a warping is detected. Changes in the radial velocities and the proper motions of these maser sources were measured, so that the model of a Kepler accretion disk could be confirmed in detail. Several more galaxies were discovered where central masers could be used for studying the dynamics in their centers.

Hence, there are three different probes of the gravitational potential in the center of galaxies: stars, gas, and masers. All three probes are employed for identifying a SMBH in galaxies, and to determine their masses.

All these observations are of course no proof of the existence of a SMBH in these galaxies because the sources from which we obtain the kinematic evidence are still too far away from the Schwarzschild radius. The conclusion of the presence of SMBHs is rather that of a missing alternative, as was already explained for the case of the GC (Sect. 2.6.3). We have no other plausible model for the mass concentrations detected. As for the case of the SMBH in the Milky Way, an ultra-compact star cluster might be postulated, but such a cluster would not be stable over a long period of time. Moreover, its luminosity would be observed in the NIR, but there is no known stellar population that could achieve the required M/L_{NIR}. Based on the existence of a SMBH in our Galaxy and in AGNs, the SMBH hypothesis is the only plausible explanation for these mass concentrations.

3.8.3 Correlation between SMBH mass and galaxy properties

Currently, strong indications for SMBHs have been found in the kinematics of stars or gas, resolving the sphere of influence of the black hole, in more than 70 nearby galaxies, and their masses have been estimated. This permits us to examine whether, and in what way, $M_\bullet$ is related to the properties of the host galaxy. In this way, a remarkable correlation was discovered: one finds that $M_\bullet$ is correlated with the absolute magnitude of the bulge component (or the

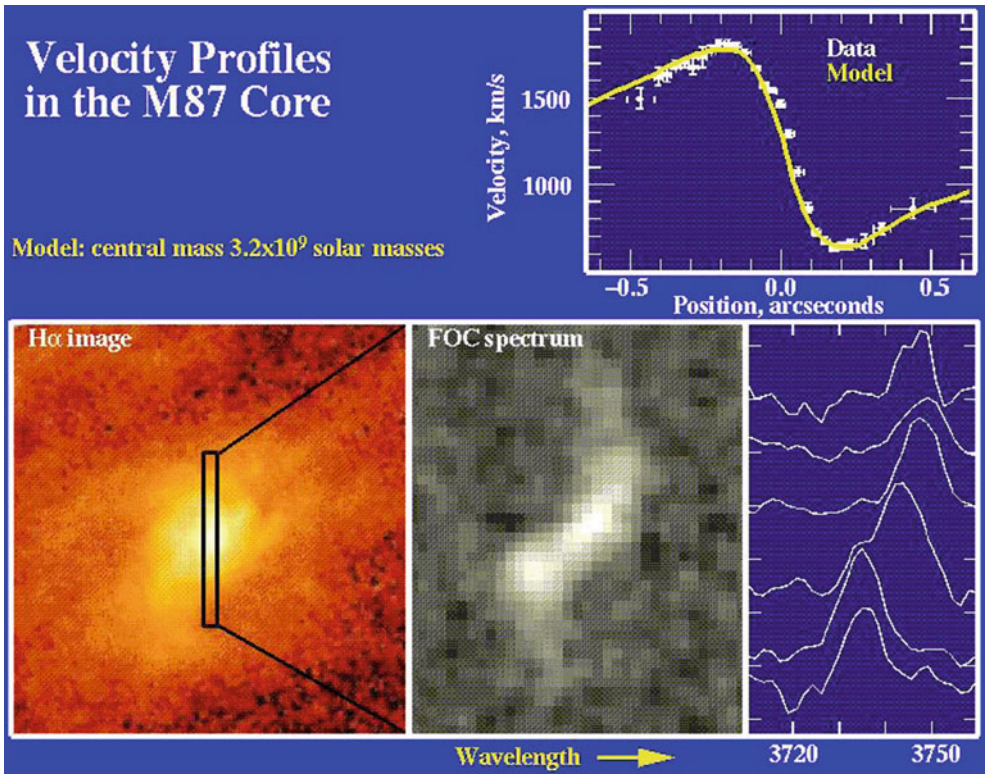


Fig. 3.43 M87 has long been one of the most promising candidates for harboring an SMBH in its center. In this figure, the position of the slit is shown superimposed on an Hα image of the galaxy (*lower left*) together with the spectrum of the [OII] line along this slit (*bottom, center*), and six spectra corresponding to six different positions along the slit, separated by $0\farcs14$ each (*lower right*). In the *upper right panel* the rotation curve extracted from the data using a kinematical model is displayed. These results show that a central mass concentration with $\sim 3 \times 10^9 M_\odot$ must be present, confined to a region less than 3 pc across—indeed leaving basically no alternative but a SMBH. Credit: STScI, NASA, ESA, W. Keel, and Macchetto et al. 1997, ApJ 489, 579, for providing the HST FOC data

spheroidal component) of the galaxy in which the SMBH is located (see Fig. 3.45, upper left panel). Here, the bulge component is either the bulge of a spiral or S0 galaxy or an elliptical galaxy as a whole. This correlation is described by

$$M_\bullet = 1.7 \times 10^9 M_\odot \left(\frac{L_V}{10^{11} L_{V\odot}} \right)^{1.11} , \qquad (3.46)$$

and indicated by the dotted line in the upper left panel of Fig. 3.45. The correlation is statistically highly significant, but the deviations of the data points from this power law are considerably larger than their error bars, with a scatter of about a factor 3 at high luminosities, increasing towards fainter galaxies. Instead of the bulge luminosity, one can also study the correlation of $M_\bullet$ with the mass of the bulge, which is plotted in the upper right panel of Fig. 3.45, and for which the best power-law fit

$$M_\bullet = 2.9 \times 10^8 M_\odot \left(\frac{M_{\rm bulge}}{10^{11} M_\odot} \right)^{1.05} \qquad (3.47)$$

is obtained. For the $M_\bullet(M_{\rm bulge})$ relation, the scatter is slightly smaller than around the $M_\bullet(L_V)$ relation. Given that the power-law index in (3.47) is almost unity, we can rewrite this relation in the form

$$M_\bullet \approx 3 \times 10^{-3} M_{\rm bulge} . \qquad (3.48)$$

Thus we find that the black hole mass is strongly correlated with the stellar properties of the host galaxy, and that the ratio of black hole mass and bulge mass is approximately 1/300. In other words, 0.3 % of the baryon mass that was used to make the stellar population in the bulge of these galaxies was transformed into a central black hole.

An even tighter correlation exists between $M_\bullet$ and the velocity dispersion in the bulge component, as can be seen in the lower panel of Fig. 3.45. This relation is best described by

$$\boxed{M_\bullet = 2.1 \times 10^8 M_\odot \left(\frac{\sigma_v}{200\,{\rm km/s}} \right)^{5.64}} . \qquad (3.49)$$

Fitting early- and late-type galaxies separately (shown by the red and blue lines in the bottom panel), the slope of the scaling relation becomes slightly flatter (5.2 and 5.06, respectively), with a normalization for the early-type galaxies being larger by about a factor 2 than that for late-type galaxies. Since the velocity dispersion in late-type galaxies is smaller than that for early-types, the difference in the normalization of the $M_\bullet(\sigma_v)$ relation between these two galaxy populations is responsible for the steeper slope of the combined power-law fit. The scatter of the $M_\bullet(\sigma_v)$ relation is smaller than those of the scaling relations with mass and luminosity, about a factor of $\sim$ 2.5, and the scatter decreases slightly with increasing σ_v.

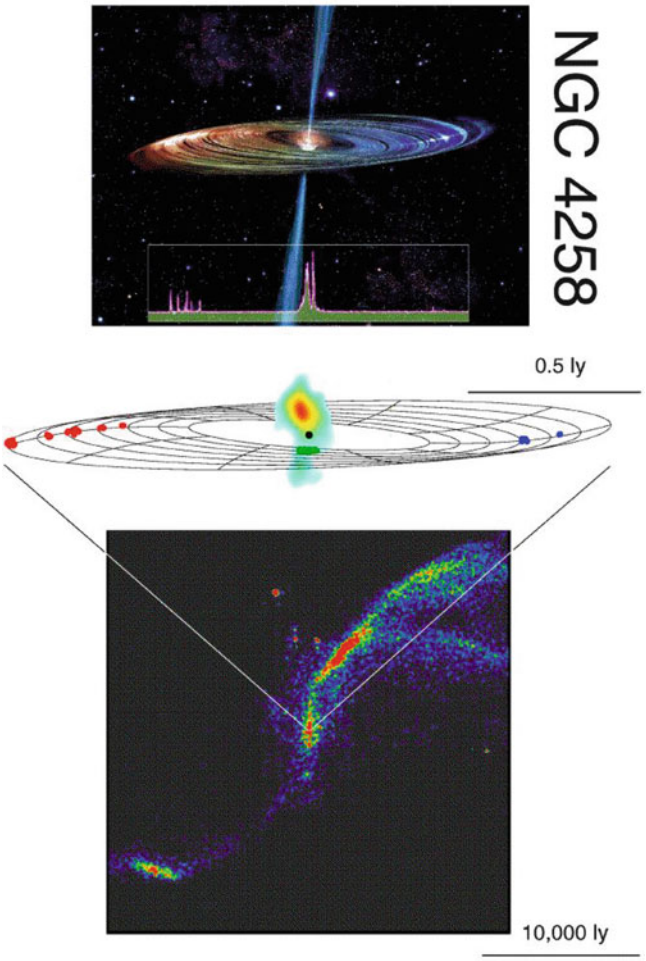


Fig. 3.44 The Seyfert galaxy NGC 4258 contains an accretion disk in its center in which several water masers are embedded. In the *top image*, an artist's impression of the hidden disk and the jet is displayed, together with the line spectrum of the maser sources. Their positions (*center image*) and velocities have been mapped by VLBI observations. From these measurements, the Kepler law for rotation in the gravitational field of a point mass of $M_\bullet \sim 35 \times 10^6 M_\odot$ in the center of this galaxy was verified. The best-fitting model of the central disk is also plotted. The *bottom image* is a 20 cm map showing the large-scale radio structure of the Seyfert galaxy. Credit: *Top*: M. Inoue (National Astronomical Observatory of Japan) & J. Kagaya (Hoshi No Techou). Center: Results from several groups, compiled by L. Greenhill, J. Herrnstein und J. Moran at CfA and the National Radio Astronomical Observatory. *Bottom*: C. De Pree, Agnes Scott College

Hence we conclude that galaxies with a bulge component host a supermassive black hole, whose mass is tightly correlated with the properties of the stellar component; in particular, the black hole mass amounts to about 0.3 % of the stellar mass in the bulge component.

Interestingly, the black hole mass at a fixed velocity dispersion is larger by a factor ~ 2 in early-type galaxies whose brightness profile shows a central core (see Sect. 3.2.2) than for those with a Sérsic light profile near the center.

The exact numerical coefficients in these scaling relations have been a matter of intense debate between different groups. However, these differences in the results can at least partially be traced back to different definitions of the velocity dispersion, especially concerning the choice of the spatial region across which it is measured.

There have been claims in the literature that even globular clusters contain a black hole; however, these claims are not undisputed. In addition, there may be objects that appear like globular clusters, but are in fact the stripped nucleus of a former dwarf galaxy. In this case, the presence of a central black hole is not unexpected, provided the scaling relation (3.49) holds down to very low velocity dispersion.

To date, the physical origin of this very close relation has not been understood in detail. The most obvious apparent explanation—that in the vicinity of a SMBH with a very large mass the stars are moving faster than around a smaller-mass SMBH—is not correct: the mass of the SMBH is significantly less than one percent of the mass of the bulge component. This is in contrast to the previously discussed case where the kinematics of the stars and gas were measured within the sphere of influence—but the size of this is *much* smaller than the bulge component itself. We can therefore disregard the contribution of the SMBH to the gravitational field in which the stars are orbiting, except in the very inner region. Instead, this correlation has to be linked to the fact that the spheroidal component of a galaxy evolves together with the SMBH. A better understanding of this relation can only be found from models of galaxy evolution. We will continue with this topic in Chap. 10.

3.9 Extragalactic distance determination

In Sect. 2.2 we discussed methods for distance determination within our own Galaxy. We will now proceed with the determination of distances to other galaxies. It should be noted that the Hubble law (1.2) specifies a relation between the redshift of an extragalactic object and its distance. The redshift z is easily measured from the shift in spectral lines. For this reason, the Hubble law (and its generalization—see Sect. 4.3.3) provides a simple method for determining distance. However, to apply this law, first the Hubble constant H_0 must be known, i.e., the Hubble law must be calibrated. Therefore, in order to determine the Hubble constant, distances have to be measured independently from redshift.

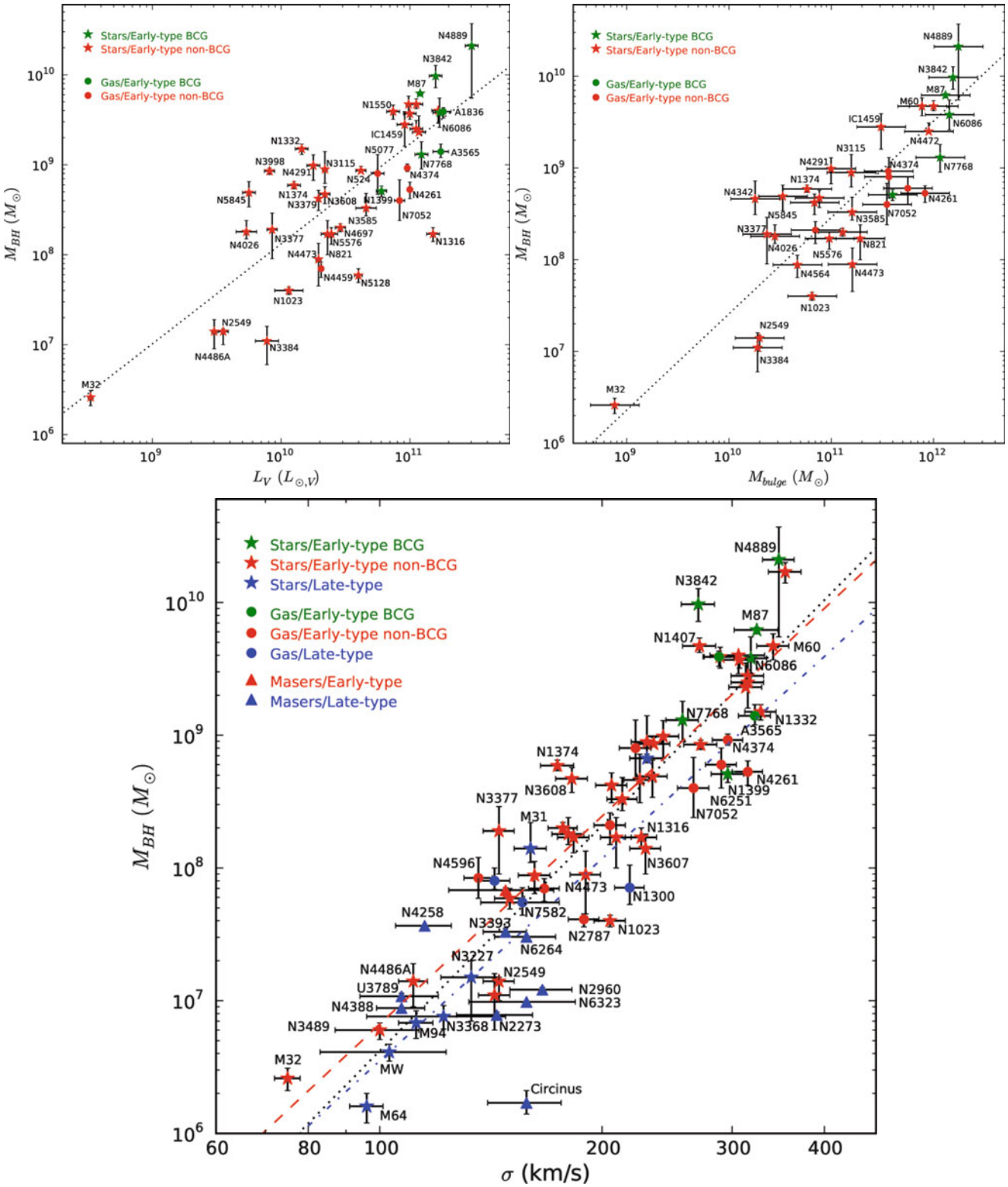


Fig. 3.45 Black hole mass scaling relations, based on measurements of $M_\bullet$ in 72 nearby galaxies. The *upper left* panel shows $M_\bullet$ as a function of the optical luminosity of the bulge component for early-type galaxies with reliable photometry. In the *upper right panel*, $M_\bullet$ is plotted as a function of the bulge stellar mass, as obtained from dynamical measurements. Finally, the *lower panel* shows $M_\bullet$ versus the velocity dispersion of the spheroidal component for the full sample of 72 galaxies. Symbols indicate the methods with which $M_\bullet$ was determined: *star-like symbols*—stellar dynamics; *circles*—gas dynamics; *triangles*—masers. The color of the symbols indicate the galaxy type: *green*—early type brightest cluster galaxy (BCG); *red*—other early-type galaxies; *blue*—late-type galaxies. The *lines* in the different panels correspond to power-law fits of the various scaling relations. Source: N.J. McConnell & C.-P. Ma 2013, *Revisiting the Scaling Relations of Black Hole Masses and Host Galaxy Properties*, ApJ 764, 184, Figs. 1, 2 & 3. ©AAS. Reproduced with permission

Peculiar motions. Furthermore, it has to be kept in mind that besides the general cosmic expansion, which is expressed in the Hubble law, objects also show *peculiar motion*, like the velocities of galaxies in clusters of galaxies or the motion of the Magellanic Clouds around our Milky Way. These peculiar velocities are induced by gravitational acceleration resulting from the locally inhomogeneous mass distribution in the Universe. For instance, our Galaxy is moving towards the Virgo cluster of galaxies, a dense accumulation of galaxies, due to the gravitational attraction caused by the cluster mass, and our neighboring galaxy M31 is actually approaching us because of mutual gravitational attraction. The measured redshift, and therefore the Doppler shift, is always a superposition of the cosmic expansion velocity and peculiar velocities.

CMB dipole anisotropy. The peculiar velocity of the Galaxy is very precisely known. The radiation of the cosmic microwave background is not completely isotropic but instead shows a dipole component. This component originates in the velocity of the Solar System relative to the rest frame in which the CMB appears isotropic (see Fig. 1.21). Due to the Doppler effect, the CMB appears hotter than average in the direction of our motion and cooler in the opposite direction. Analyzing this CMB dipole allows us to determine our peculiar velocity, which yields the result that the Sun moves at a velocity of $(368 \pm 2)\,\mathrm{km/s}$ relative to the CMB rest frame. Furthermore, the Local Group of galaxies (see Sect. 6.1) is moving at $v_{\mathrm{LG}} \approx 600\,\mathrm{km/s}$ relative to the CMB rest frame.

Distance ladder. For the redshift of a source to be dominated by the Hubble expansion, the cosmic expansion velocity $v = cz = H_0 D$ has to be much larger than typical peculiar velocities. This means that in order to determine H_0 we have to consider sources at large distances for the peculiar velocities to be negligible compared to $H_0\, D$.

Direct estimates of the distances of distant galaxies are very difficult to obtain. Traditionally one uses a *distance ladder*: at first, the *absolute distances* to nearby galaxies are measured directly. If methods to measure *relative distances* (that is, distance ratios) with sufficient precision are utilized, the distances to galaxies further away are then determined relative to those nearby. In this way, by means of relative methods, distances are estimated for galaxies that are sufficiently far away so that their redshifts are dominated by the Hubble flow.

3.9.1 Distance of the LMC

The distance of the Large Magellanic Cloud (LMC) can be estimated using various methods. For example, we can resolve and observe individual stars in the LMC, which forms the basis of the MACHO experiments (see Sect. 2.5.2). Because the metallicity of the LMC is significantly lower than that of the Milky Way, some of the methods discussed in Sect. 2.2 are only applicable after correcting for metallicity effects, e.g., the photometric distance determination or the period-luminosity relation for pulsating stars.

Supernova 1987A. Perhaps the most precise method of determining the distance to the LMC is a purely geometrical one. The supernova SN 1987A that exploded in 1987 in the LMC illuminates a nearly perfectly elliptical ring (see Fig. 3.46). This ring consists of material that was once ejected by the stellar winds of the progenitor star of the supernova and that is now radiatively excited by energetic photons from the supernova explosion. The corresponding recombination radiation is thus emitted only when photons from the SN hit the surrounding gas. Because the observed ring is almost certainly intrinsically circular and the observed ellipticity is caused only by its inclination with respect to the line-of-sight, the distance to SN 1987A can be derived from observations of the ring. First, the inclination angle is determined from its observed ellipticity. The gas in the ring is excited by photons from the SN a time R/c after the original explosion, where R is the radius of the ring. We do not observe the illumination of the ring instantaneously because light from the section of the ring closer to us reaches us earlier than light from the more distant part. Thus, its illumination was seen sequentially along the ring. Combining the time delay in the illumination between the nearest and farthest part of the ring with its inclination angle, we then obtain the physical diameter of the ring. When this is compared to the measured angular size of the major axis of $\sim 1\farcs7$, the ratio yields the distance to SN 1987A,

$$D_{\mathrm{SN1987A}} \approx 51.8\,\mathrm{kpc} \pm 6\% \,.$$

If we now assume the extent of the LMC along the line-of-sight to be small, this distance can be identified with the distance to the LMC. The value is also compatible with other distance estimates (e.g., as derived by using photometric methods based on the properties of main sequence stars—see Sect. 2.2.4).

Most recently, the distance to the LMC was determined by observing eclipsing binary systems with a long orbital period. Spectroscopy allowed the accurate determination of the orbits of these systems, which together with the measured angular separation yielded the distances to these binaries. This resulted in a distance to the LMC of

$$D_{\mathrm{EB}} = 50.0\,\mathrm{kpc} \pm 2.2\% \,.$$

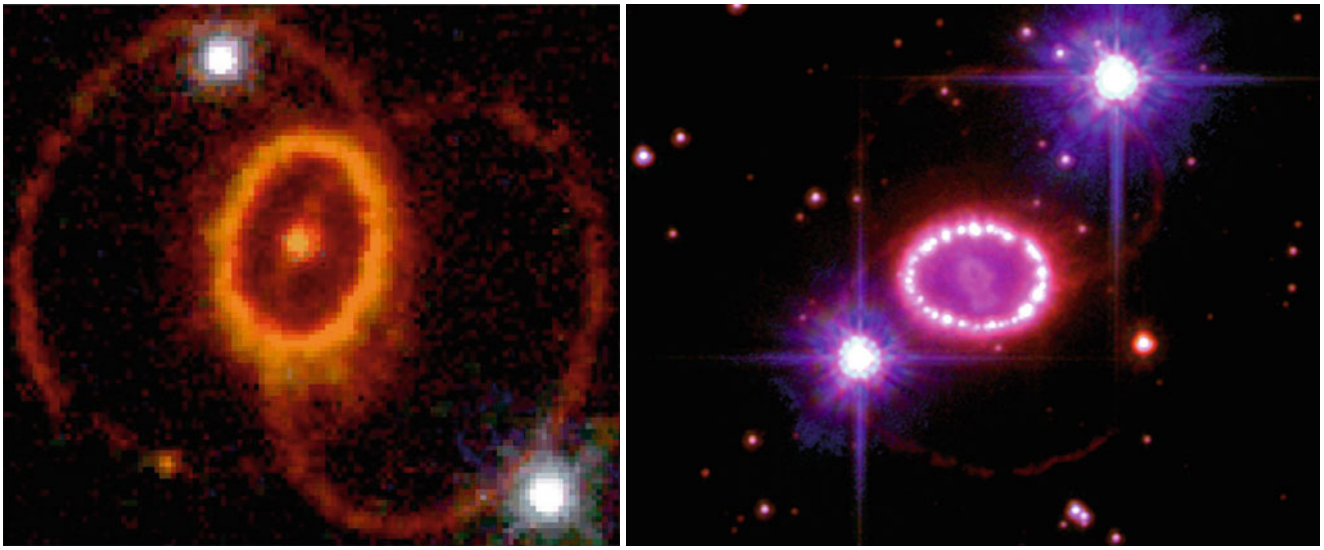


Fig. 3.46 *Left panel*: The ring around supernova 1987A in the LMC, as seen 7 years after the explosion, is illuminated by photons from the explosion which induce the radiation from the gas in the ring. It is inclined towards the line-of-sight; thus it appears to be elliptical. Lighting up of the ring was not instantaneous, due to the finite speed of light: those sections of the ring closer to us lit up earlier than the more distant parts. From the time shift in the onset of radiation across the ring, its diameter can be derived. Combining this with the measured angular diameter of the ring, the distance to SN 1987A—and thus the distance to the LMC—can be determined. *The picture on the right* shows an image of the ring, taken with the HST about 20 years after the original explosion (and with a different orientation of the telescope). In this later image, the ring is seen to host a large number of bright spots, which were not observed in the earlier image seen on the *left*. These bright spots correspond to gas concentrations in the inner regions of the ring, which were heated up, and thus exited to glow, by the supernova blast wave slamming into the ring. Since the blast wave propagates with a velocity much smaller than c, it took about 10 years before it reached the innermost parts of the ring and the first spots were seen. The material seen inside the ring is debris from the explosion, heated up by radioactive decays of nuclei which were formed during the supernova. Credit: NASA, STScI, ESA, P. Challis and R. Kirshner (Harvard-Smithsonian Center for Astrophysics)

3.9.2 The Cepheid distance

In Sect. 2.2.7, we discussed the period-luminosity relation of pulsating stars. Due to their high luminosity, Cepheids turn out to be particularly useful since they can be observed out to large distances.

For the period-luminosity relation of the Cepheids (also called the Leavitt law) to be a good distance measure, it must first be calibrated. This calibration has to be done with as large a sample of Cepheids as possible at a known distance. Cepheids in the LMC are well-suited for this purpose because we believe we know the distance to the LMC quite precisely, see above. Also, due to the relatively small extent of the LMC along the line-of-sight, all Cepheids in the LMC should be located at approximately the same distance. For this reason, the period-luminosity relation is calibrated using the Cepheids in the LMC. Due to the large number of Cepheids available for this purpose (several thousands, many of which were found in the microlens surveys discussed in Sect. 2.5.3), the resulting statistical errors are small. However, uncertainties remain in the form of systematic errors related to the metallicity dependence of the period-luminosity relation, as well as with regards to interstellar extinction. These effects can be corrected for since the color of Cepheids depends on the metallicity as well.

With the high angular resolution of the HST, individual Cepheids in galaxies are visible at distances up to that of the Virgo cluster of galaxies. In fact, determining the distance to Virgo as a central step in the determination of the Hubble constant was one of the major scientific aims of the HST. In the *Hubble Key Project*, the distances to numerous spiral galaxies in the Virgo cluster were determined by identifying Cepheids and measuring their periods.

Since most of the galaxies for which these Cepheid distances were determined have a metallicity that is comparable to that of the Milky Way, rather than the LMC, the aforementioned metallicity effects, as well as the absolute distance determination to the LMC, remained the main source of systematic uncertainty in the determination of galaxy distances in the Virgo cluster. Over the past few years, these uncertainties could be reduced substantially by measuring the trigonometric parallaxes of ten Galactic Cepheids at distances between 300 pc and 600 pc with the HST. This allowed a calibration of the period-luminosity relation with an accuracy of about 3 %. Since the number of these Galactic Cepheids is still rather small, and all but one have periods below 10 days, one can use the slope of the period-luminosity as obtained from LMC Cepheids, where the statistics is much better, but determine the amplitude of this relation from the Galactic Cepheids.

Another way to accurately calibrate Cepheid distances is provided by the maser galaxy NGC 4258 (Fig. 3.44) mentioned before. The dynamics of the maser source can be studied with great accuracy, due to the compact nature of these source, and these results can be interpreted straightforwardly, due to the simple orbital motion of the sources. In particular, with measurements of proper motions and acceleration of the maser sources in a Keplerian disk, the distance to NGC 4258 could be determined to be $D = 7.2 \pm 0.2$ Mpc. This distance estimate can then be used to calibrate the period-luminosity relation from Cepheids in this galaxy.

With these new results, the calibration of Cepheid distances has been considerably improved relative to that based on LMC Cepheids alone. It should also be noted that the new period-luminosity relation does not depend on the adopted distance to the LMC anymore. Conversely, with the newly calibrated period-luminosity relation, the distance to the LMC can be determined, yielding $D_{\mathrm{LMC}} = 47.9$ kpc, with an estimated error of $\sim 3\,\%$.

3.9.3 Tip of the Red Giant Branch

Another method which can be used to determine the distance to nearby galaxies is based on the color-magnitude relation of red giant branch stars. This method is based on the fact that red giant stars have a maximum luminosity, as is well understood from the theory of stellar evolution.[8] This maximum luminosity, which depends on the laws of nuclear physics, is almost independent of the chemical composition of a star. It can be identified by a clear discontinuity in the number of red giants in a galaxy as a function of magnitude. Relating the apparent magnitude of this tip of the red giant branch in a galaxy to the known absolute magnitude of this maximum luminosity, the distance to the galaxy can be determined.

Whereas red giants are less luminous that Cepheids, and thus cannot be observed to equally large distances, the tip of the red giant branch method can still be used for galaxies as far away as the Virgo cluster. Therefore, this method serves as a calibration for galaxy distances, independent of the Cepheid method.

[8] In red giants, nuclear burning of hydrogen occurs in a shell around the core which is formed by helium-rich gas. As shell burning proceeds, the helium core becomes more massive, as well as hotter, and the stellar luminosity increases. Once a certain threshold in the core temperature is reached, the central helium core ignites, and the stars quickly evolved to the horizontal branch. The threshold core temperature then corresponds to the maximum luminosity of a red giant.

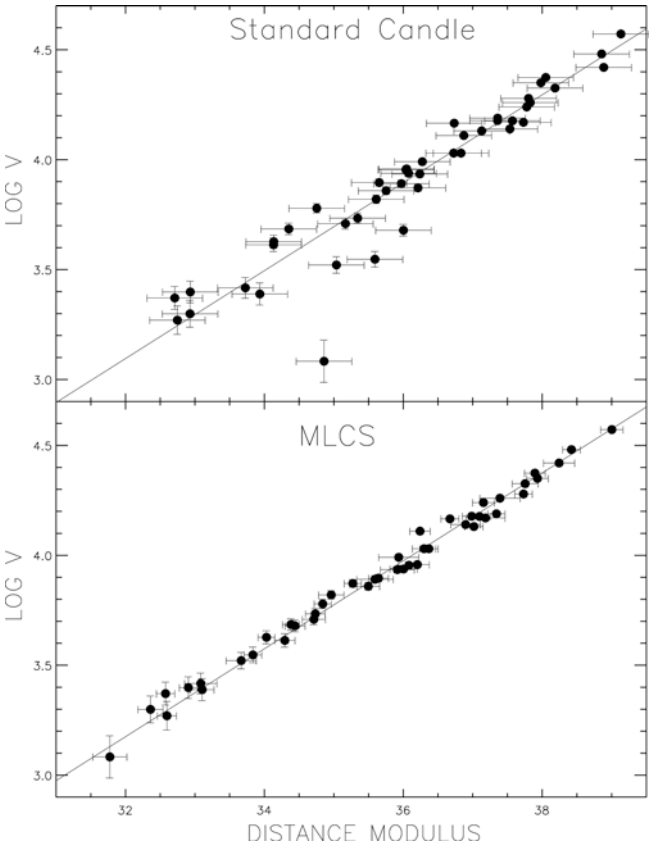


Fig. 3.47 The Hubble diagram for relatively nearby SNe Ia. Plotted is the measured expansion velocity cz as a function of the distance modulus for the individual supernovae. In the *top panel*, it is assumed that all sources have the same luminosity. If this was correct, all data points should be aligned along the *straight line*, as follows from the Hubble law. Obviously, the scatter is significant. In the *bottom panel*, the luminosities have been corrected by means of the so-called MLCS method in which the shape of the light curve and the colors of the SN are used to 'standardize' the luminosity (see text for more explanations). By this the deviations from the Hubble law become dramatically smaller—the dispersion is reduced from 0.42 to 0.15 mag. Source: A.V. Filippenko & A.G. Riess 2000, *Evidence from Type Ia Supernovae for an Accelerating Universe*, astro-ph/0008057, p. 5, Fig. 1

3.9.4 Supernovae Type Ia

As mentioned in Sect. 2.3.2, according to the (arguably) most plausible model, Type Ia supernovae are supposed to be the result of explosion processes of white dwarfs which cross a critical mass threshold by accretion of additional matter. This threshold should be identical for all SNe Ia, making it at least plausible that they all have the same luminosity. If this were the case, they would be ideal for standard candles: owing to their high luminosity, they can be detected and examined even at very large distances.

However, it turns out that SNe Ia are not really standard candles, since their maximum luminosity varies from object to object with a dispersion of about 0.4 mag in the blue band light. This is visible in the top panel of Fig. 3.47. If

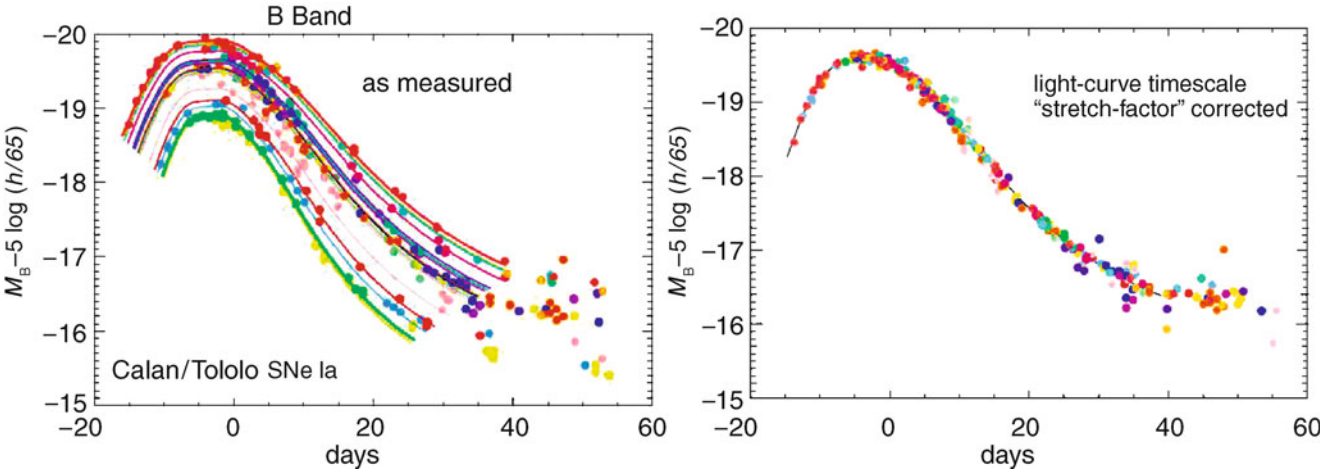


Fig. 3.48 *Left panel*: B-band light curves of different SNe Ia. One sees that the shape of the light curves and the maximum luminosity of the SNe Ia differ substantially among the sample. A transformation was found empirically with a single parameter described by the width of the light curve. By means of this transformation, the different light curves can all be made congruent, as displayed in the *right panel*. Credit: M. Hamuy, S. Perlmutter, Supernova Cosmology Project

SNe Ia were standard candles, the data points would all be located on a straight line, as described by the Hubble law. Clearly, deviations from the Hubble law can be seen, which are significantly larger than the photometric measurement errors.

It turns out that there is a strong correlation between the luminosity and the shape of the light curve of SNe Ia. Those of higher maximum luminosity show a slower decline in the light curve, as measured from its maximum. Furthermore, the observed flux is possibly affected by extinction in the host galaxy, in addition to the extinction in the Milky Way. With the resulting reddening of the spectral distribution, this effect can be derived from the observed colors of the SN. The combined analysis of these effects provides a possibility for deducing an empirical correction to the maximum luminosity from the observed light curves in several filters, accounting both for the relation of the width of the curve to the observed luminosity and for the extinction. This correction was calibrated on a sample of SNe Ia for which the distance to the host galaxies is very accurately known.[9] With this correction applied, the SNe Ia follow the Hubble law much more closely, as can be seen in the bottom panel of Fig. 3.47. A scatter of only $\sigma = 0.15$ mag around the Hubble relation remains. Figure 3.48 demonstrates the effect of this correction on the light curves of several SNe Ia which initially appear to have very different maximum luminosities and widths. After correction they become nearly identical. The left panel of Fig. 3.48 suggests that the light curves of SN Ia can basically be described by a one-parameter family of functions, and that this parameter can be deduced from the shape, in particular the width, of the light curves.

With this correction, SNe Ia become standardized candles, i.e., by observing the light curves in several bands their 'corrected' maximum luminosity can be determined. Since the observed flux of a source depends on its luminosity and its distance, once the luminosity is known and the flux measured, the distance to the SN Ia can be inferred. SNe Ia are visible out to very large distances, so that they also permit distance estimates at such large redshifts where the simple Hubble law (1.6) is no longer valid, but needs to be generalized based on a cosmological model (Sect. 4.3.3). We will see in Sect. 8.3 that these measurements belong to the most important pillars on which our standard model of cosmology rests.

3.9.5 Secondary distance indicators

The Virgo cluster, at a measured distance of about 16 Mpc, is not sufficiently far away from us to directly determine the Hubble constant from its distance and redshift, because peculiar velocities still contribute considerably to the measured redshift at this distance. To get to larger distances, a number of relative distance indicators are used. They are all based on measuring the distance *ratio* of galaxies. If the distance to one of the two is known, the distance to the other is then obtained from the ratio. By this procedure, distances to more remote galaxies can be measured. Below, we will review some of the most important secondary distance indicators.

Surface brightness fluctuations of galaxies. Another method of estimating distance ratios is surface brightness

[9] To calibrate the luminosity of SN Ia, the surveys for determining Cepheid distances were preferentially targeted towards those galaxies in which a SN Ia had been observed.

fluctuations. It is based on the fact that the number of bright stars per area element in a galaxy fluctuates—purely by Poisson noise: If N stars are expected in an area element, the relative fluctuations of the number of stars will be $\sqrt{N}/N = 1/\sqrt{N}$. These fluctuation in the number of stars are observed as fluctuations of the local surface brightness. To demonstrate that this effect can be used to estimate distances, we consider a solid angle $\mathrm{d}\omega$. The corresponding area element $\mathrm{d}A = D^2\,\mathrm{d}\omega$ depends quadratically on the distance D of the galaxy. If we now consider two galaxies at a radius from their center where their surface brightnesses are the same,[10] and assume that their stellar populations are comparable, then the galaxy with the larger distance from us will have a larger the number of stars N in this solid angle. Correspondingly, its relative fluctuations of the surface brightness will be smaller. By comparing the surface brightness fluctuations of different galaxies, one can therefore estimate relative distances. This method also has to be calibrated on the galaxies for which Cepheid or other primary distances are available.

Planetary nebulae. The brightness distribution of planetary nebulae in a galaxy seems to have an upper limit which is the nearly the same for each galaxy (see Fig. 3.49). If a sufficient number of planetary nebulae are observed and their brightnesses measured, it enables us to determine their luminosity function from which the maximum apparent magnitude is then derived. By calibration on galaxies of known Cepheid distance, the corresponding maximum absolute magnitude can be determined, which then allows the determination of the distance modulus for other galaxies, thus their distances.

Scaling relations. The scaling relations for galaxies—fundamental plane for ellipticals, Tully–Fisher relation for spirals (see Sect. 3.4)—can be calibrated on local groups of galaxies or on the Virgo cluster, the distances of which have been determined from Cepheids. Although the scatter of these scaling relations can be 15 % for individual galaxies, the statistical fluctuations are reduced when observing several galaxies at about the same distance (such as in clusters and groups). This enables us to estimate the distance ratio of two clusters of galaxies.

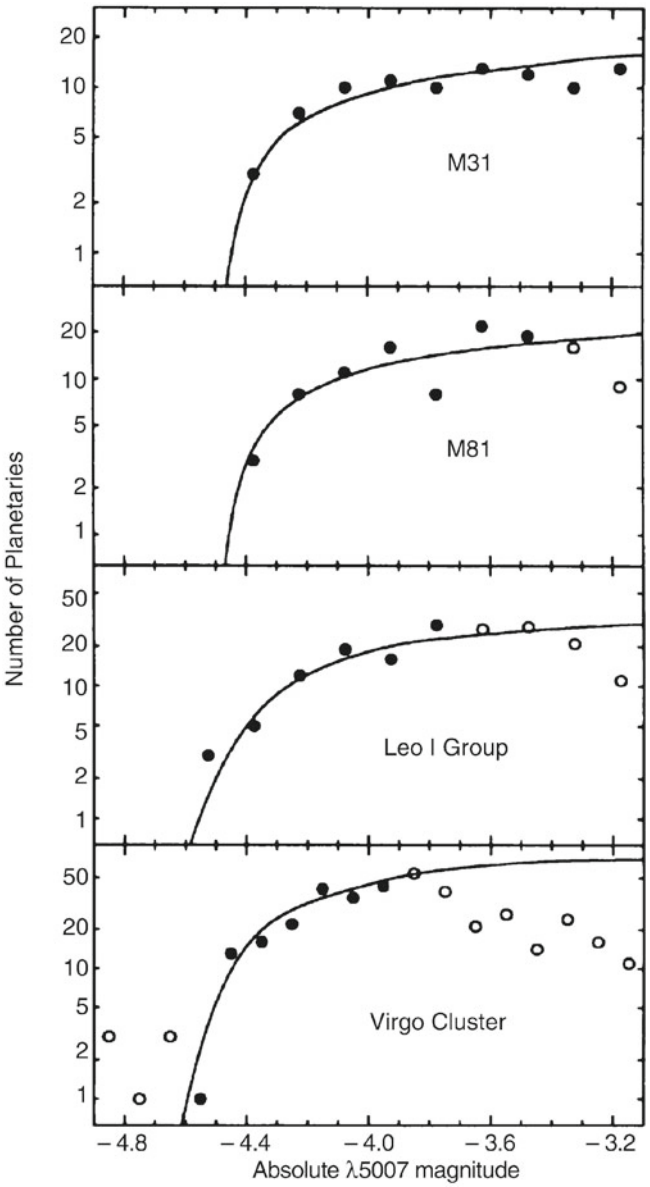


Fig. 3.49 Brightness distribution of planetary nebulae in Andromeda (M31), M81, three galaxies in the Leo I group, and six galaxies in the Virgo cluster. The plotted absolute magnitude was measured in the emission line of double-ionized oxygen at $\lambda = 5007$ Å in which a large fraction of the luminosity of a planetary nebula is emitted. This characteristic line emission is also used in the identification of such objects in other galaxies. In all cases, the distribution is described by a nearly identical luminosity function; it seems to be a universal function in galaxies. Therefore, the brightness distribution of planetary nebulae can be used to estimate the distance of a galaxy. In the fits shown, the data points marked by *open symbols* were disregarded: at these magnitudes, the distribution function is probably not complete. Source: G.H. Jacoby et al. 1992, *A critical review of selected techniques for measuring extragalactic distances*, PASP 104, 599, p. 635, Fig. 15

3.9.6 The Hubble Constant

Using the various methods described above, the Hubble Key Project aimed at determining the value of the Hubble constant. Cepheid distances to 18 galaxies in the range $3\,\mathrm{Mpc} \le D \le 25\,\mathrm{Mpc}$ were determined, which were then used to calibrate the Tully–Fischer relation for spirals,

[10] Recall that the surface brightness does not depend on distance, as long as we are considering objects in the nearby Universe, i.e., with redshifts $z \ll 1$.

the $D_n - \sigma$ relation for ellipticals, the peak luminosity of SN Ia, and the surface brightness fluctuation method. These secondary distance indicators were then applied to galaxies as much larger distances, such that their peculiar velocity is negligible compared to their radial velocity according to the Hubble law.

By combining the various methods, a distance to the Coma cluster of about 90 Mpc was obtained. Furthermore, using the SN Ia technique, distances of galaxies with $D \lesssim 400$ Mpc could be measured. The resulting Hubble constant, incorporating the new calibration of the period-luminosity relation from Cepheids in the Milky Way and in NGC 4258, reads

$$\boxed{H_0 = 74 \pm 3\,\mathrm{km/s/Mpc}}\,. \tag{3.50}$$

The error given here denotes the estimated systematic uncertainty in the determination of H_0, whereas the statistical uncertainty is smaller by a factor of two and thus subdominant. Thus, the dimensionless Hubble constant, defined in (1.7), is $h = 0.74 \pm 0.03$. A convenient way to memorize this is $h^2 \approx 1/2$.

Thus, the uncertainty about the value of the Hubble constant has finally shrunk to a mere 6 %—after decades of intense debates between two camps of scientists, where the first camp obtained values near 50 km/s/Mpc, and the other camp about twice this value, each with error bars that were very much smaller than the differences between their results.

We will see later that the Hubble constant can also be measured by completely different methods. The currently most accurate of these, based on tiny small-scale anisotropies of the cosmic microwave background (Sect. 8.7.1), results in a value which is in fairly good agreement with that in (3.50), and yields a smaller estimated error.

3.10 Luminosity function of galaxies

Definition of the luminosity function. The luminosity function specifies the way in which the members of a class of objects are distributed with respect to their luminosity. More precisely, the luminosity function is the number density of objects (here galaxies) of a specific luminosity. $\Phi(M)\,\mathrm{d}M$ is defined as the number density of galaxies with absolute magnitude in the interval $[M, M + \mathrm{d}M]$. The total density of galaxies is then

$$\nu = \int_{-\infty}^{\infty} \mathrm{d}M\ \Phi(M)\,. \tag{3.51}$$

Accordingly, $\Phi(L)\,\mathrm{d}L$ is defined as the number density of galaxies with a luminosity between L and $L + \mathrm{d}L$. It should be noted here explicitly that both definitions of the luminosity function are denoted by the same symbol, although they represent different mathematical functions, i.e., they describe different functional relations. It is therefore important (and in most cases not difficult) to deduce from the context which of these two functions is being referred to.

Problems in determining the luminosity function. At first sight, the task of determining the luminosity function of galaxies does not seem very difficult. The history of this topic shows, however, that we encounter a number of problems in practice. As a first step, the determination of galaxy luminosities is required, for which, besides measuring the flux, distance estimates are also necessary. For very distant galaxies redshift is a sufficiently reliable measure of distance, whereas for nearby galaxies the methods discussed in Sect. 3.9 have to be applied.

Another problem occurs for nearby galaxies, namely the large-scale structure of the galaxy distribution. To obtain a representative sample of galaxies, a sufficiently large volume has to be surveyed because the galaxy distribution is heavily structured on scales of $\sim 100h^{-1}$ Mpc and more. On the other hand, galaxies of particularly low luminosity can only be observed locally, so the determination of $\Phi(L)$ for small L always needs to refer to local galaxies. Finally, one has to deal with the so-called *Malmquist bias;* in a flux-limited sample luminous galaxies will always be overrepresented because they are visible at larger distances (and therefore are selected from a larger volume). A correction for this effect is always necessary, and was applied, e.g., to Fig. 3.7.

3.10.1 The Schechter luminosity function

The global galaxy distribution can be roughly approximated by the *Schechter luminosity function*

$$\boxed{\Phi(L) = \left(\frac{\Phi^*}{L^*}\right)\left(\frac{L}{L^*}\right)^{\alpha} \exp\left(-L/L^*\right)}\,, \tag{3.52}$$

where L^* is a characteristic luminosity above which the distribution decreases exponentially, α is the slope of the luminosity function for small L, and Φ^* specifies the normalization of the distribution. A schematic plot of this function, as well as a fit to early data, is shown in Fig. 3.50.

Expressed in magnitudes, this function appears much more complicated. Considering that an interval $\mathrm{d}L$ in luminosity corresponds to an interval $\mathrm{d}M$ in absolute magnitude, with $\mathrm{d}L/L = -0.4 \ln 10\,\mathrm{d}M$, and using $\Phi(L)\,\mathrm{d}L = \Phi(M)\,\mathrm{d}M$, i.e., the number of sources in these intervals are

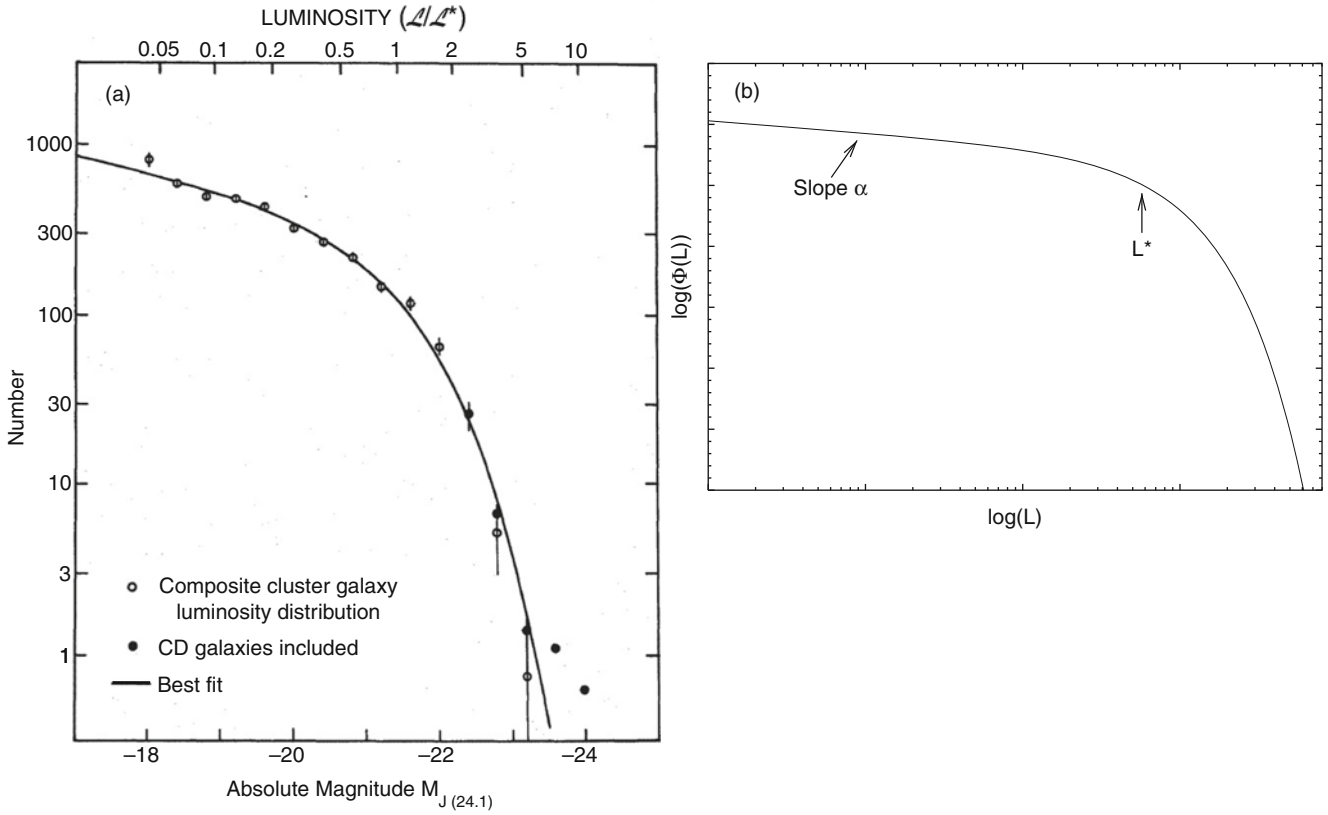


Fig. 3.50 *Left panel*: galaxy luminosity function as obtained from 13 clusters of galaxies. For the *solid circles*, cD galaxies have also been included. *Right panel*: a schematic plot of the Schechter function. Source (*left panel*): P. Schechter 1976, *An analytic expression for the luminosity function for galaxies*, ApJ 203, 297, p. 300, Fig. 2. ©AAS. Reproduced with permission

of course the same, we obtain

$$\Phi(M) = \Phi(L) \left|\frac{\mathrm{d}L}{\mathrm{d}M}\right| = \Phi(L)\, 0.4\, \ln 10\, L \tag{3.53}$$

$$= 0.921 \Phi^* 10^{0.4(\alpha+1)(M^*-M)} \exp\left(-10^{0.4(M^*-M)}\right) . \tag{3.54}$$

As mentioned above, the determination of the parameters entering the Schechter function is difficult; a characteristic set of parameters in the blue band is given as

$$\begin{aligned} \Phi^* &= 1.6 \times 10^{-2} h^3\, \mathrm{Mpc}^{-3} , \\ M_B^* &= -19.7 + 5 \log h \quad \text{or} \\ L_B^* &= 1.2 \times 10^{10} h^{-2} L_{\odot,\mathrm{B}} , \\ \alpha &= -1.07 . \end{aligned} \tag{3.55}$$

While the blue light of galaxies can be strongly affected by star formation, the luminosity function in the red bands measures the typical stellar distribution. In the K-band, we have

$$\begin{aligned} \Phi^* &= 1.6 \times 10^{-2} h^3\, \mathrm{Mpc}^{-3} , \\ M_K^* &= -23.1 + 5 \log h , \\ \alpha &= -0.9 . \end{aligned} \tag{3.56}$$

The total number density of galaxies is formally infinite if $\alpha \leq -1$, but the validity of the Schechter function does of course not extend to arbitrarily small L. The luminosity density[11]

$$l_{\mathrm{tot}} = \int_0^\infty \mathrm{d}L\; L\, \Phi(L) = \Phi^*\, L^*\, \Gamma(2+\alpha) \tag{3.58}$$

is finite for $\alpha \geq -2$. The integral in (3.58), for $\alpha \sim -1$, is dominated by $L \sim L^*$, and $n = \Phi^*$ is thus a good estimate for the mean density of L^*-galaxies.

[11] Here, $\Gamma(x)$ is the Gamma function, defined by

$$\Gamma(x) = \int_0^\infty \mathrm{d}y\; y^{(x-1)}\, \mathrm{e}^{-y} . \tag{3.57}$$

For positive integers, $\Gamma(n+1) = n!$. We have $\Gamma(0.7) \approx 1.30$, $\Gamma(1) = 1$, $\Gamma(1.3) \approx 0.90$. Since these values are all close to unity, $l_{\mathrm{tot}} \sim \Phi^* L^*$ is a good approximation for the luminosity density.

In fact, whereas the rise of the Schechter luminosity function towards small L may at first sight suggest that low-L galaxies are 'more important' than higher luminosity objects, this is clearly not the case. Quite the contrary: for $\alpha = -1$, 60 % of the whole luminosity of the galaxy population is emitted from objects with $0.22L^* \leq L \leq 1.6L^*$, and 90 % of the luminosity stems from galaxies with $0.1 \leq L/L^* \leq 2.3$. Hence, the total luminosity of galaxies stems from a fairly narrow range around $\sim L^*/2$. Since the luminosity, specifically in red and NIR bands, is almost proportional to the stellar mass, *most of the stars in the Universe live in* $\sim L^*$*-galaxies.* The parameters of the Schechter function then tell us that the mean number density of 'typical' galaxies (i.e., those with $L \sim L^*$) is about $2 \times 10^{-2} h^3\,\mathrm{Mpc}^{-3}$, meaning that the mean separation between two luminous galaxies is about $4h^{-1}$ Mpc.

3.10.2 More accurate luminosity and mass functions

With better statistics of galaxy surveys, it became clear that the luminosity function of galaxies deviates from the Schechter form. There is also no obvious reason why such a simple relation for describing the luminosity distribution of galaxies should exist. Whereas the Schechter function approximates the total galaxy distribution, each morphological type of galaxy has its own luminosity function, with a shape that can significantly deviate from a Schechter function—see Fig. 3.51. For instance, spirals are relatively narrowly distributed in L, whereas the distribution of ellipticals is much broader if we account for the full L-range, from giant ellipticals to dwarf ellipticals, although, if we just consider normal ellipticals, their luminosity range is comparable to that of spirals. Ellipticals dominate in particular at large L; the low end of the luminosity function is likewise dominated by dwarf ellipticals and irregular galaxies. In addition, the luminosity distribution of cluster and group galaxies differs from that of field galaxies. The fact that these populations add up to something as simple as (3.52) is a most likely a coincidence.

Indeed, the size and quality of the Sloan Digital Sky Survey and other redshift surveys allowed more robust conclusions about the luminosity function of galaxies. For the total population, a good fit is obtained by using a double-Schechter function of the form

$$\Phi(L) = \left[\left(\frac{\Phi_1^*}{L^*}\right)\left(\frac{L}{L^*}\right)^{\alpha_1} + \left(\frac{\Phi_2^*}{L^*}\right)\left(\frac{L}{L^*}\right)^{\alpha_2}\right] \times \exp\left(-L/L^*\right)\,, \tag{3.59}$$

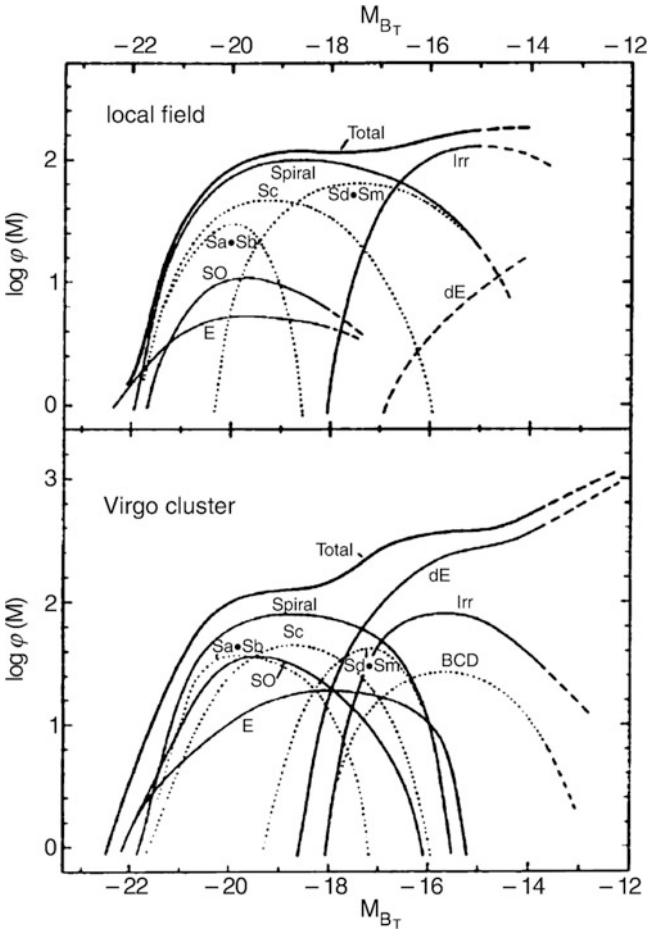


Fig. 3.51 The luminosity function for different Hubble types of field galaxies (*top*) and galaxies in the Virgo cluster of galaxies (*bottom*). *Dashed curves* denote extrapolations. In contrast to Fig. 3.50, the more luminous galaxies are plotted towards the *left*. The Schechter luminosity function of the total galaxy distribution is the sum of the luminosity functions of individual galaxy types which can deviate significantly from the Schechter function. One can see that in clusters the major contribution at faint magnitudes comes from the dwarf ellipticals (dEs), and that at the bright end ellipticals and S0s contribute much more strongly to the luminosity function than they do in the field. This trend is even more prominent in regular clusters of galaxies. Source: B. Binggeli et al. 1988, *The luminosity function of galaxies*, ARA&A 26, 509, Fig. 1, p. 542. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 26 ©1988 by Annual Reviews www.annualreviews.org

with two normalizations Φ_i^* and two slopes α_i, but the same cut-off luminosity L^*. This form allows a transition of the slope of the luminosity function, which for very small L is given by the more negative ones of the two α's. In Fig. 3.52, we show the r-band luminosity function of nearby galaxies as obtained from the SDSS, separated into galaxy types, together with a double-Schechter fit to the total galaxy population. At the luminous end of the distribution, early-type galaxies dominate the luminosity function, although not by a large factor. For faint galaxies, the situation is reversed,

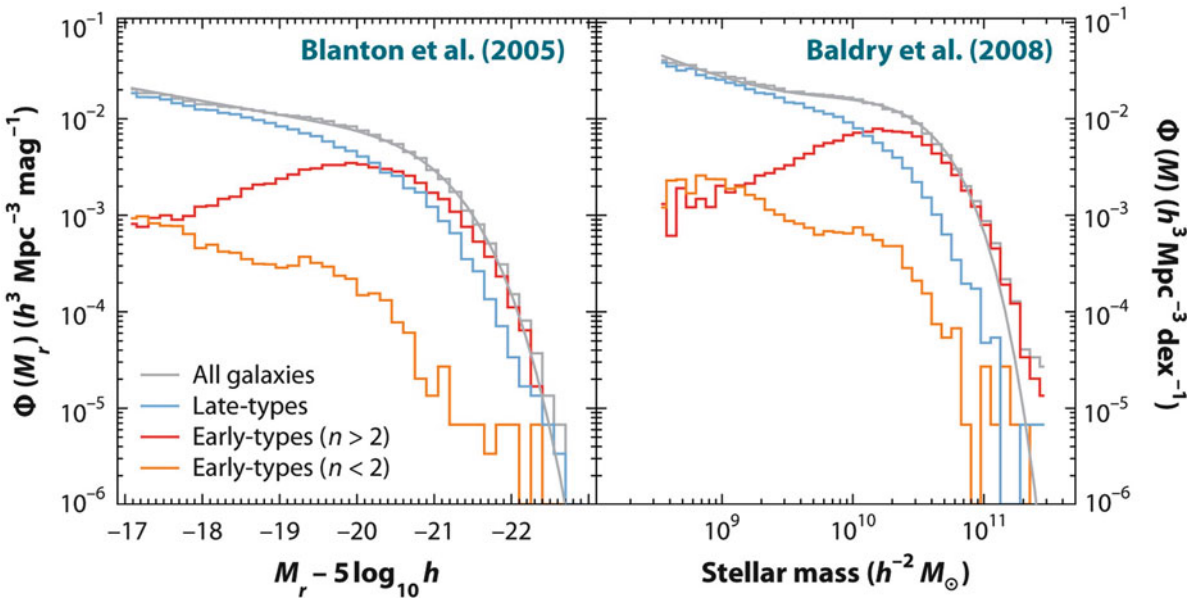


Fig. 3.52 *Left panel*: The luminosity function of galaxies, i.e., the number density of galaxies as a function of absolute r-band magnitude. The total luminosity function is shown as the *grey* histogram, with the *smooth curve* being a fit with a double-Schechter function (3.59). Also shown are the luminosity function of early-type galaxies, split according to the Sérsic index n into concentrated and less concentrated ones (*red* and *orange* histograms, respectively), and late-type galaxies shown in *blue*. The early-types with $n \leq 2$ are totally subdominant for all L, and contribute substantially to the early-type population only for very low luminosities, in agreement with what is seen in Fig. 3.39. *Right panel*: The stellar mass function of galaxies, with the same galaxy populations as in the *left-hand panel*. The total mass function is again fit with a double-Schechter function. Source: M.R. Blanton & J. Moustakas 2009, *Physical Properties and Environments of Nearby Galaxies*, ARA&A 47, 159, p. 166, Fig. 3. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 47 ©2009 by Annual Reviews www.annualreviews.org

with late-type galaxies being much more numerous than early types. This is due to the fact that the slope of the faint-end luminosity function is much steeper for late-type galaxies. If we consider instead the luminosity function in the near ultraviolet, it is totally dominated by late-type galaxies at all L.

Interestingly, the value of L^* in the double Schechter function is the same for the two components—one might have expected that a better fit could be obtained by the sum of two Schechter functions, with two different values of the cut-off luminosity L^*. This, however, is not the case. It thus seems that L^* corresponds to a characteristic luminosity of galaxies, whose value is fixed by the physics of galaxy formation and evolution. As we will show in Chap. 10, this is indeed the case.

The right-hand panel of Fig. 3.52 displays the corresponding mass function of galaxies, obtained from the luminosity function using the appropriate M/L for the stellar population. Here we see that the dominance of early-type galaxies at the high stellar mass end of the distribution is even stronger, since they have a higher M/L than late types.

The mass functions of individual galaxy types can be used to estimate where most of the stellar mass is located. Curiously, about one third of the stellar mass is contained in disks, one third in ellipticals, and one third in bulges and bars.

3.11 Galaxies as gravitational lenses

In Sect. 2.5 the gravitational lens effect was discussed, where we concentrated on the deflection of light by point masses. The lensing effect by stars leads to image separations too small to be resolved by any existing telescope. Since the separation angle is proportional to the square root of the lens mass (2.82), the angular separation of the images will be about a million times larger if a galaxy acts as a gravitational lens. In this case it should be observable, as was predicted in 1937 by Fritz Zwicky. Indeed, multiple images of very distant sources have been found, together with the galaxy responsible for the image splitting. In this section we will first describe this effect by continuing the discussion we began in Sect. 2.5.1. Examples of the lens effect and its various applications will then be discussed.

3.11.1 The gravitational lens effect—Part II

The geometry of a typical gravitational lens system is sketched in Fig. 2.30 and again in Fig. 3.53. The physical description of such a lens system for an arbitrary mass distribution of the deflector is obtained from the following considerations.

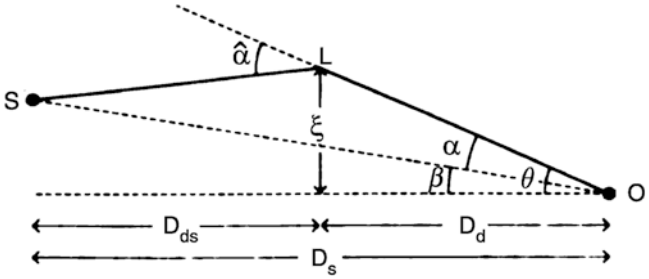


Fig. 3.53 As a reminder, another sketch of the lens geometry. Source: R.D. Blandford & R. Narayan 1992, *Cosmological applications of gravitational lensing*, ARA&A 30, 311, Fig. 5, p. 318. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 30 ©1992 by Annual Reviews www.annualreviews.org

If the gravitational field is weak (which is the case in all situations considered here), the gravitational effects can be linearized.[12] Hence, the deflection angle of a lens that consists of several mass components can be described by a linear superposition of the deflection angles of the individual components,

$$\hat{\boldsymbol{\alpha}} = \sum_i \hat{\boldsymbol{\alpha}}_i \ . \tag{3.60}$$

We assume that the deflecting mass has a small extent along the line-of-sight, as compared to the distances between observer and lens (D_d) and between lens and source (D_ds). All mass elements can then be assumed to be located at the same distance D_d. This physical situation is called a *geometrically thin lens*. If a galaxy acts as the lens, this condition is certainly fulfilled—the extent of galaxies is typically $\sim 100h^{-1}\,\mathrm{kpc}$ while the distances of lens and source are typically $\sim$ Gpc. We can therefore write (3.60) as a superposition of Einstein angles of the form (2.74),

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \sum_i \frac{4Gm_i}{c^2}\frac{\boldsymbol{\xi}-\boldsymbol{\xi}_i}{|\boldsymbol{\xi}-\boldsymbol{\xi}_i|^2} \ , \tag{3.61}$$

where $\boldsymbol{\xi}_i$ is the projected position vector of the mass element m_i, and $\boldsymbol{\xi}$ describes the position of the light ray in the lens plane, also called the impact vector.

[12] To characterize the strength of a gravitational field, we refer to the gravitational potential Φ. The ratio Φ/c^2 is dimensionless and therefore well suited to distinguishing between strong and weak gravitational fields. For weak fields, $|\Phi|/c^2 \ll 1$. Another possible way to quantify the field strength is to apply the virial theorem: if a mass distribution is in virial equilibrium, then $v^2 \sim |\Phi|$, and weak fields are therefore characterized by $v^2/c^2 \ll 1$. Because the typical velocities in galaxies are $\sim 200\,\mathrm{km/s}$, for galaxies $|\Phi|/c^2 \lesssim 10^{-6}$. The typical velocities of galaxies in a cluster of galaxies are $\sim 1000\,\mathrm{km/s}$, so that in clusters $|\Phi|/c^2 \lesssim 10^{-5}$. Thus the gravitational fields occurring are weak in both cases.

For a continuous mass distribution we can imagine subdividing the lens into mass elements of mass $\mathrm{d}m = \Sigma(\boldsymbol{\xi})\,\mathrm{d}^2\xi$, where $\Sigma(\boldsymbol{\xi})$ describes the *surface mass density* of the lens at the position $\boldsymbol{\xi}$, obtained by projecting the spatial (three-dimensional) mass density ρ along the line-of-sight to the lens. With this definition the deflection angle (3.61) can be transformed into an integral,

$$\boxed{\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \frac{4G}{c^2}\int \mathrm{d}^2\xi'\ \Sigma(\boldsymbol{\xi}')\frac{\boldsymbol{\xi}-\boldsymbol{\xi}'}{|\boldsymbol{\xi}-\boldsymbol{\xi}'|^2}} \ . \tag{3.62}$$

This deflection angle is then inserted into the lens equation (2.78),

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \frac{D_\mathrm{ds}}{D_\mathrm{s}}\hat{\boldsymbol{\alpha}}(D_\mathrm{d}\boldsymbol{\theta}) \ , \tag{3.63}$$

where $\boldsymbol{\xi} = D_\mathrm{d}\boldsymbol{\theta}$ describes the relation between the position $\boldsymbol{\xi}$ of the light ray in the lens plane and its apparent direction $\boldsymbol{\theta}$. We define the scaled deflection angle as in (2.79),

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) = \frac{D_\mathrm{ds}}{D_\mathrm{s}}\hat{\boldsymbol{\alpha}}(D_\mathrm{d}\boldsymbol{\theta}) \ ,$$

so that the lens equation (3.63) can be written in the simple form (see Fig. 3.53)

$$\boxed{\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})} \ . \tag{3.64}$$

A more convenient way to write the scaled deflection is as follows,

$$\boldsymbol{\alpha}(\boldsymbol{\theta}) = \frac{1}{\pi}\int \mathrm{d}^2\theta'\ \kappa(\boldsymbol{\theta}')\frac{\boldsymbol{\theta}-\boldsymbol{\theta}'}{|\boldsymbol{\theta}-\boldsymbol{\theta}'|^2} \ , \tag{3.65}$$

where

$$\kappa(\boldsymbol{\theta}) = \frac{\Sigma(D_\mathrm{d}\boldsymbol{\theta})}{\Sigma_\mathrm{cr}} \tag{3.66}$$

is the *dimensionless surface mass density*, and the so-called *critical surface mass density*

$$\Sigma_\mathrm{cr} = \frac{c^2\,D_\mathrm{s}}{4\pi G\,D_\mathrm{d}\,D_\mathrm{ds}} \tag{3.67}$$

depends only on the distances to the lens and to the source. Although Σ_cr incorporates a combination of cosmological distances, it is of a rather 'human' order of magnitude,

$$\Sigma_\mathrm{cr} \approx 0.35\left(\frac{D_\mathrm{d}\,D_\mathrm{ds}}{D_\mathrm{s}\ 1\,\mathrm{Gpc}}\right)^{-1}\ \mathrm{g\,cm^{-2}} \ .$$

A source is visible at several positions $\boldsymbol{\theta}$ on the sphere, or multiply imaged, if the lens equation (3.64) has several solutions $\boldsymbol{\theta}$ for a given source position $\boldsymbol{\beta}$. A more detailed analysis of the properties of this lens equation yields the following general result:

If $\Sigma \geq \Sigma_{\rm cr}$ in at least one point of the lens, then source positions $\boldsymbol{\beta}$ exist such that a source at $\boldsymbol{\beta}$ has multiple images. It immediately follows that κ is a good measure for the strength of the lens. A mass distribution with $\kappa \ll 1$ at all points is a weak lens, unable to produce multiple images, whereas one with $\kappa \gtrsim 1$ for certain regions of $\boldsymbol{\theta}$ is a strong lens.

For sources that are small compared to the characteristic scales of the lens, the magnification μ of an image, caused by the differential light deflection, is given by (2.86), i.e.,

$$\boxed{\mu = \left| \det \left(\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}} \right) \right|^{-1}} \,. \tag{3.68}$$

The importance of the gravitational lens effect for extragalactic astronomy stems from the fact that gravitational light deflection is independent of the nature and the state of the deflecting matter. Therefore, it is equally sensitive to both dark and baryonic matter and independent of whether or not the mass distribution is in a state of equilibrium. The lens effect is thus particularly suitable for probing matter distributions, without requiring any further assumptions about the state of equilibrium or the relation between dark and luminous matter.

3.11.2 Simple models

Axially symmetric mass distributions. The simplest models for gravitational lenses are those which are axially symmetric, for which $\Sigma(\boldsymbol{\xi}) = \Sigma(\xi)$, where $\xi = |\boldsymbol{\xi}|$ denotes the distance of a point from the center of the lens in the lens plane. In this case, the deflection angle is directed radially inwards, and we obtain

$$\hat{\alpha} = \frac{4GM(\xi)}{c^2\,\xi} = \frac{4G}{c^2\,\xi}\, 2\pi \int_0^{\xi} \mathrm{d}\xi'\; \xi'\, \Sigma(\xi') \,, \tag{3.69}$$

where $M(\xi)$ is the mass within radius ξ. Accordingly, for the scaled deflection angle we have

$$\alpha(\theta) = \frac{m(\theta)}{\theta} := \frac{1}{\theta}\, 2 \int_0^{\theta} \mathrm{d}\theta'\; \theta'\, \kappa(\theta') \,, \tag{3.70}$$

where, in the last step, $m(\theta)$ was defined as the dimensionless mass within θ. Since $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are collinear, the lens equation becomes one-dimensional because only the radial coordinate needs to be considered,

$$\beta = \theta - \alpha(\theta) = \theta - \frac{m(\theta)}{\theta} \,. \tag{3.71}$$

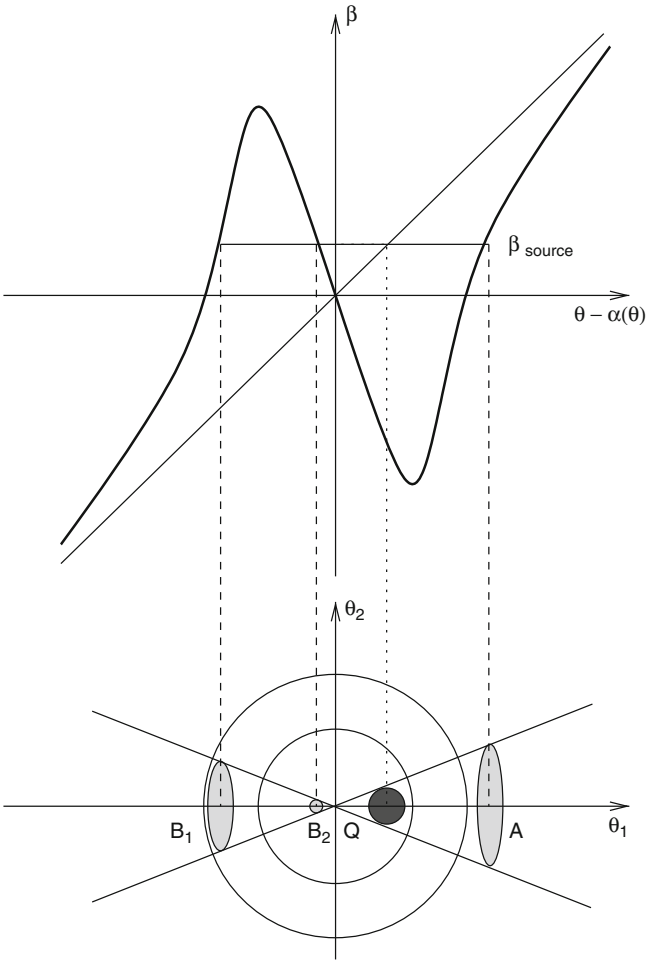


Fig. 3.54 Sketch of an axially symmetric lens. In the top panel, $\theta - \alpha(\theta)$ is plotted as a function of the angular separation θ from the center of the lens, together with the straight line $\beta = \theta$. The three intersection points of the *horizontal line* at fixed β with the *curve* $\theta - \alpha(\theta)$ are the three solutions of the lens equation. The *bottom image* indicates the positions and sizes of the images on the observer's sky. Here, Q is the unlensed source (which is not visible itself in the case of light deflection, of course!), and A, B1, B2 are the observed images of the source. The sizes of the images, and thus their fluxes, differ considerably; the *inner image* B2 is particularly weak in the case depicted here. The flux of B2 relative to that of image A depends strongly on the core radius of the lens; it can be so low as to render the third image unobservable. In the special case of a singular isothermal sphere, the innermost image is in fact absent. Adapted from P. Young et al. 1980, *The double quasar Q0957 + 561 A,B—A gravitational lens image formed by a galaxy at* $z = 0.39$, ApJ 241, 507, Fig. 6

An illustration of this one-dimensional lens mapping is shown in Fig. 3.54.

Example: Point-mass lens. For a point mass M, the dimensionless mass is independent of θ,

$$m(\theta) = \frac{4GM}{c^2}\, \frac{D_{\rm ds}}{D_{\rm d}\, D_{\rm s}} \,,$$

reproducing the lens equation from Sect. 2.5.1 for a point-mass lens.

Example: Isothermal sphere. We saw in Sect. 2.4.2 that the rotation curve of our Milky Way is flat for large radii, and we know from Sect. 3.3.4 that the rotation curves of other spiral galaxies are flat as well. This indicates that the mass of a galaxy increases proportional to r, thus $\rho(r) \propto r^{-2}$, or more precisely,

$$\rho(r) = \frac{\sigma_v^2}{2\pi G r^2} \,. \tag{3.72}$$

Here, σ_v is the one-dimensional velocity dispersion of stars in the potential of the mass distribution if the distribution of stellar orbits is isotropic. In principle, σ_v is therefore measurable spectroscopically from the line width. The mass distribution described by (3.72) is called a *singular isothermal sphere* (SIS). Because this mass model is of significant importance not only for the analysis of the lens effect, we will discuss its properties in a bit more detail.

The density (3.72) diverges for $r \to 0$ as $\rho \propto r^{-2}$, so that the mass model cannot be applied up to the very center of a galaxy. However, the steep central increase of the rotation curve shows that the core region of the mass distribution, in which the density profile will deviate considerably from the r^{-2}-law, must be small for galaxies. Furthermore, the mass diverges for large r such that $M(r) \propto r$. The mass profile thus has to be cut off at some radius in order to get a finite total mass. This cut-off radius is probably very large ($\gtrsim$ 100 kpc for L^*-galaxies) because the rotation curves are flat to at least the outermost point at which they are observable.

The SIS is an appropriate simple model for gravitational lenses over a wide range in radius since it seems to reproduce the basic properties of lens systems (such as image separation) quite well. The surface mass density is obtained from the projection of (3.72) along the line-of-sight,

$$\Sigma(\xi) = \frac{\sigma_v^2}{2G\xi} \,, \tag{3.73}$$

which yields the projected mass $M(\xi)$ within radius ξ

$$M(\xi) = 2\pi \int_0^\xi \mathrm{d}\xi'\, \xi'\, \Sigma(\xi') = \frac{\pi \sigma_v^2 \xi}{G} \,. \tag{3.74}$$

With (3.69) the deflection angle can be obtained,

$$\hat{\alpha}(\xi) = 4\pi \left(\frac{\sigma_v}{c}\right)^2 , \quad \boxed{\alpha(\theta) = 4\pi \left(\frac{\sigma_v}{c}\right)^2 \left(\frac{D_{\mathrm{ds}}}{D_{\mathrm{s}}}\right) \equiv \theta_{\mathrm{E}}} \,. \tag{3.75}$$

Thus the deflection angle for an SIS is constant and equals θ_{E}, and it depends quadratically on σ_v. θ_{E} is called the *Einstein angle* of the SIS. The characteristic scale of the Einstein angle is

$$\theta_{\mathrm{E}} = 1\farcs15 \left(\frac{\sigma_v}{200\,\mathrm{km/s}}\right)^2 \left(\frac{D_{\mathrm{ds}}}{D_{\mathrm{s}}}\right) \,, \tag{3.76}$$

from which we conclude that the angular scale of the lens effect in galaxies is about an arcsecond for massive galaxies. The lens equation (3.71) for an SIS is

$$\beta = \theta - \theta_{\mathrm{E}} \frac{\theta}{|\theta|} \,, \tag{3.77}$$

where we took into account the fact that the deflection angle is negative for $\theta < 0$ since it is always directed inwards.

Solution of the lens equation for the singular isothermal sphere. If $|\beta| < \theta_{\mathrm{E}}$, two solutions of the lens equation exist,

$$\theta_1 = \beta + \theta_{\mathrm{E}} \quad , \quad \theta_2 = \beta - \theta_{\mathrm{E}} \,. \tag{3.78}$$

Without loss of generality, we assume $\beta \geq 0$; then $\theta_1 > \theta_{\mathrm{E}} > 0$ and $0 > \theta_2 > -\theta_{\mathrm{E}}$: one image of the source is located on either side of the lens center, and the separation of the images is

$$\boxed{\Delta\theta = \theta_1 - \theta_2 = 2\theta_{\mathrm{E}} = 2\farcs3 \left(\frac{\sigma_v}{200\,\mathrm{km/s}}\right)^2 \left(\frac{D_{\mathrm{ds}}}{D_{\mathrm{s}}}\right)} \,. \tag{3.79}$$

Thus, the angular separation of the images does not depend on the position of the source. For massive galaxies acting as lenses it is of the order of somewhat more than 1 arcsec. For $\beta > \theta_{\mathrm{E}}$ only one image of the source exists, at θ_1, meaning that it is located on the same side of the center of the lens as the unlensed source.

For the magnification, we find

$$\mu(\theta) = \frac{|\theta/\theta_{\mathrm{E}}|}{||\theta/\theta_{\mathrm{E}}| - 1|} \,. \tag{3.80}$$

If $\theta \approx \theta_{\mathrm{E}}$, μ is very large. Such solutions of the lens equation exist for $|\beta| \ll \theta_{\mathrm{E}}$, so that sources close to the center of the source plane may be highly magnified. If $\beta = 0$, the image of the source will be a ring of radius $\theta = \theta_{\mathrm{E}}$, a so-called *Einstein ring*.

More realistic models. Mass distributions occurring in nature are not expected to be truly symmetric. The ellipticity of the mass distribution or external shear forces (caused, for example, by the tidal gravitational field of neighboring galaxies) will disturb the symmetry. The lensing properties of the galaxy will change by this symmetry breaking. For example, more than two images may be generated. Figure 3.55 illustrates the lens properties of such elliptical mass distributions. One can see, for example, that pairs of

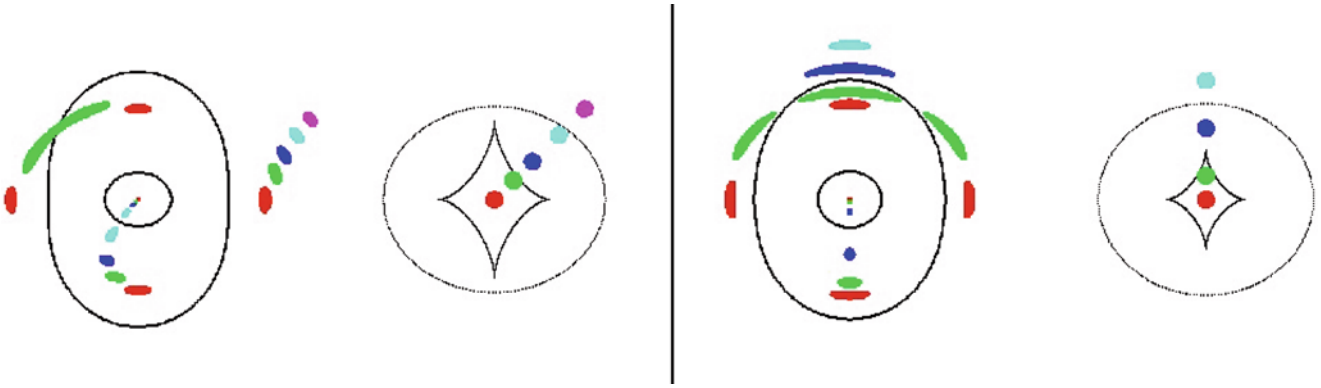


Fig. 3.55 Geometry of an 'elliptical' lens, whereby it is of little importance whether the surface mass density Σ is constant on ellipses (i.e., the mass distribution has elliptical isodensity contours) or whether an originally spherical mass distribution is distorted by an external tidal field. On the *right-hand side in both panels*, several different source positions in the source plane are displayed, each corresponding to a *different color*. The origin in the source plane is chosen as the intersection point of the line connecting the center of symmetry in the lens and the observer with the source plane (see also Fig. 2.31). Depending on the position of the source, 1, 3, or 5 images may appear in the lens plane (i.e., the observer's sky); they are shown on the *left-hand side of each panel*. The *curves* in the lens plane are the *critical curves*, the location of all points for which $\mu \to \infty$. The *curves* in the source plane (i.e., on the *right-hand side* of each panel) are *caustics*, obtained by mapping the *critical curves* onto the source plane using the lens equation. Obviously, the number of images of a source depends on the source location relative to the location of the caustics. Strongly elongated images of a source occur close to the critical curves. Source: R. Narayan & M. Bartelmann 1996, *Lectures on Gravitational Lensing*, astro-ph/9606001

images, which are both heavily magnified, may be observed with a separation significantly smaller than the Einstein radius of the lens. Nevertheless, the characteristic image separation is still of the order of magnitude given by (3.79).

3.11.3 Examples for gravitational lenses

Currently, about 200 gravitational lens systems are known in which a galaxy acts as the lens. Many of them were discovered serendipitously, but most were found in systematic searches for lens systems. Amongst the most important lens surveys are: (1) *The Cosmic Lens All-Sky Survey (CLASS).* About 15 000 radio sources with a flat radio spectrum (these often contain compact radio components, see Sect. 5.1.3) were scanned for multiple components, using the VLA. Possible multiple image candidate were then studied in more detail. From this survey, 22 lens systems were found. These numbers immediately show that strong lensing is a rather rare phenomenon, with roughly 1 out of 1000 distant sources being lensed by a foreground galaxy. (2) *The SDSS Quasar Lens Search (SQLS).* The multicolor image data from SDSS were used to study images of quasars from the spectroscopic SDSS survey, to search for indications of multiple images. In this survey, 28 lens systems were found. Due to the resolution of the SDSS imaging data, this survey preferentially selected lenses with large image separations. (3) *The Sloan Lens Advanced Camera for Surveys (SLACS).* If a galaxy lies directly behind a closer galaxy, then the resulting spectrum will be a superposition of the spectra of the two galaxies. In SLACS, the galaxy spectra of the SDSS were searched for indications of the presence of two different redshifts. Candidate systems were then imaged with the HST, to find evidence for multiple images or (partial) Einstein rings. The SLACS survey yielded 85 strong lensing systems.

The different search strategies for lenses all have their merits. For example, radio lenses, as found by CLASS, can often be studied with much higher angular resolution, due to the availability of Very Long Baseline Interferometry (VLBI). However, many of these radio sources are very faint in the optical, and determining their redshift from optical spectroscopy can be highly challenging. Indeed, the source redshift is known only for half of the CLASS lenses. This problem is absent in the SLACS survey, since the redshifts of both (foreground and background) galaxies were determined at the stage were candidate systems were identified. Furthermore, the redshift distribution of the lenses and sources are quite different; of the three surveys mentioned, SLACS has the lowest lens and source redshifts, due to the limiting magnitude of the spectroscopic galaxy survey.

Most lens galaxies are ellipticals; in fact, spirals occur in only $\sim$ 10 % of all lens systems. The reason for that can be traced back to the fact the massive ellipticals are more abundant than spirals, as can be seen in the right-hand panel of Fig. 3.52. Since the mass does not only determine the image separation that a lens can generate, but also the effective area of the sky in which a background source must be located in order to be multiply imaged, it turns out that ellipticals dominate the lensing probability distribution.[13]

[13] We have seen that an isothermal sphere can multiply image a source if its position on the sky lies within θ_E of the center of the lens galaxy. The corresponding area within which sources are multiply images is thus $\pi\theta_E^2 \propto \sigma_v^4$. According to the Tully–Fischer relation, or the Faber–

QSO 0957+561, the first double quasar. The first lens system was discovered in 1979 by Walsh, Carswell & Weymann when the optical identification of a radio source showed two point-like optical sources (see Fig. 3.56). Both could be identified as quasars located at the same redshift of $z_s = 1.41$ and having very similar spectra (see Fig. 3.57). Deep optical images of the field show an elliptical galaxy situated between the two quasar images, at a redshift of $z_d = 0.36$. The galaxy is so massive and so close to image B of the source that it *has to* produce a lens effect. However, the observed image separation of $\Delta\theta = 6''.1$ is considerably larger than expected from the lens effect by a single galaxy (3.79). The explanation for this is that the lens galaxy is located in a cluster of galaxies; the additional lens effect of the cluster adds to that of the galaxy, boosting the image separation to a large value. The lens system QSO 0957+561 was observed in all wavelength ranges, from the radio to the X-ray. The two images of the quasar are very similar at all λ, including the VLBI structure (Fig. 3.57)—as would be expected since the lens effect is independent of the wavelength, i.e., achromatic.

QSO PG1115+080. In 1980, the so-called triple quasar was discovered, composed of three optical quasars at a maximum angular separation of just below $3''$. Component A is significantly brighter than the other two images (B, C; see Fig. 3.58, left). In high-resolution images it was found that the brightest image is in fact a double image: A is split into A1 and A2. The angular separation of the two roughly equally bright images is $\sim 0''.5$, which is considerably smaller than all other angular separations in this system. The four quasar images have a redshift of $z_s = 1.72$, and the lens is located at $z_d = 0.31$. The image configuration is one of those that are expected for an elliptical lens, see Fig. 3.55.

With the NIR camera NICMOS on-board HST, not only the quasar images and the lens galaxy were observed, but also a nearly complete Einstein ring (Fig. 3.58, right). The source of this ring is the host galaxy of the quasar (see Sect. 5.4.5) which is substantially redder than the active galactic nucleus itself.

From the image configuration in such a quadruple system, the mass of the lens within the images can be estimated very accurately. The four images of the lens system trace a circle around the center of the lens galaxy, the radius of which can be identified with the Einstein radius of the lens. From this, the mass of the lens within the Einstein radius follows immediately because the Einstein radius is obtained from the lens equation (3.71) by setting $\beta = 0$. Therefore, the Einstein radius is the solution of the equation

$$\theta = \alpha(\theta) = \frac{m(\theta)}{\theta} \,,$$

or

$$m(\theta_E) = \frac{4GM(\theta_E)}{c^2} \frac{D_{ds}}{D_d\, D_s} = \theta_E^2 \,.$$

This equation is best written as

$$\boxed{M(\theta_E) = \pi\, (D_d \theta_E)^2\, \Sigma_{cr}} \,, \tag{3.81}$$

which is readily interpreted:

> The mass within θ_E of a lens follows from the fact that the mean surface mass density within θ_E equals the critical surface mass density Σ_{cr}. A more accurate determination of lens masses is possible by means of detailed lens models. For quadruple image systems, the masses can be derived with a precision of a few percent—these are the most precise mass determinations in (extragalactic) astronomy.

QSO 2237+0305: The Einstein Cross. A spectroscopic survey of galaxies found several unusual emission lines in the nucleus of a nearby spiral galaxy which cannot originate from the galaxy itself. Instead, they are emitted by a background quasar at redshift $z_s = 1.7$ situated exactly behind this spiral. High-resolution images show four point sources situated around the nucleus of this galaxy, with an image separation of $\Delta\theta \approx 1''.8$ (Fig. 3.59). The spectroscopic analysis of these point sources revealed that all four are images of the same quasar (Fig. 3.60).

The images in this system are positioned nearly symmetrically around the lens center; this is also a typical lens configuration which may be caused by an elliptical lens (see Fig. 3.55). The Einstein radius of this lens is $\theta_E \approx 0''.9$, and we can determine the mass within this radius with a precision of $\sim 3\,\%$.

Einstein rings. More examples of Einstein rings are displayed in Figs. 3.61 and 3.62. The first of these is a radio galaxy, with its two radio components being multiply imaged by a lens galaxy—one of the two radio sources is imaged into four components, the other mapped into a double image. In the NIR the radio galaxy is visible as a complete Einstein ring. This example shows very clearly that the appearance of

Jackson relation, $\sigma^4 \propto L$, so the lensing probability of a galaxy is roughly proportional to its luminosity. Since we have seen before that the luminosity of the galaxy population is fully dominated by galaxies with $L \sim L^*/2$, we do not expect to find many lens systems with very small or very large image separation, in agreement with observational results.

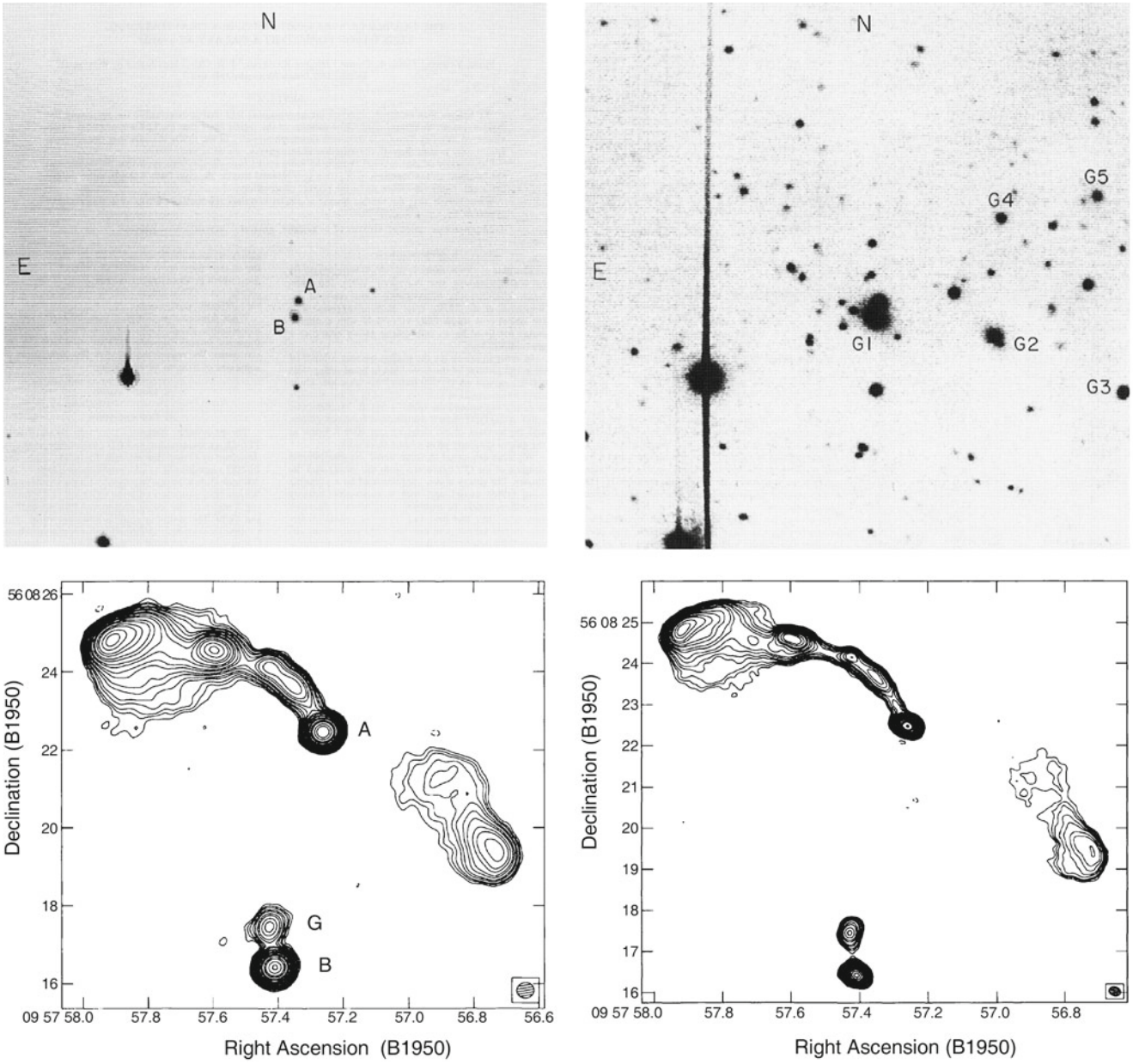


Fig. 3.56 *Top:* optical images of the double quasar QSO 0957+561. The image on the *top left* has a short exposure time; here, the two point-like images A & B of the quasar are clearly visible. In contrast, the image on the *top right* has a longer exposure time, showing the lens galaxy G1 between the two quasar images. Several other galaxies (G2–G5) are visible as well. The lens galaxy is a member of a cluster of galaxies at $z_d = 0.36$. *Bottom:* two radio maps of QSO 0957+561, observed with the VLA at 6 cm (*left*) and 3.6 cm (*right*), respectively. The two images of the quasar are denoted by A & B; G is the radio emission of the lens galaxy. The quasar has a radio jet, which is a common property of many quasars (see Sect. 5.1.3). On small angular scales, the jet can be observed by VLBI techniques in both images (see Fig. 3.57). On large scales only a single image of the jet exists, seen in image A; this property should be compared with Fig. 3.55 where it was demonstrated that the number of images of a source (component) depends on its position in the source plane. Source: *Top*: P. Young et al. 1980, *The double quasar Q0957 + 561 A,B—A gravitational lens image formed by a galaxy at* $z = 0.39$, ApJ 241, 507, p. 508, 509, Fig. 1a,b. ©AAS. Reproduced with permission. *Bottom*: M. Harvanek et al. 1997, *High Dynamic Range VLA Observations of the Gravitationally Lensed Quasar 0957+561*, AJ 114, 2240, p. 2242, Fig. 1. ©AAS. Reproduced with permission

the images of a source depends on the source size: to obtain an Einstein ring a sufficiently extended source is needed.

At radio wavelengths, the quasar MG 1654+13 consists of a compact central source and two radio lobes. As we will discuss in Sect. 5.1.3, this is a very typical radio morphology for quasars. One of the two lobes has a ring-shaped structure, which prior to this observation had never been observed before. An optical image of the field shows the optical quasar

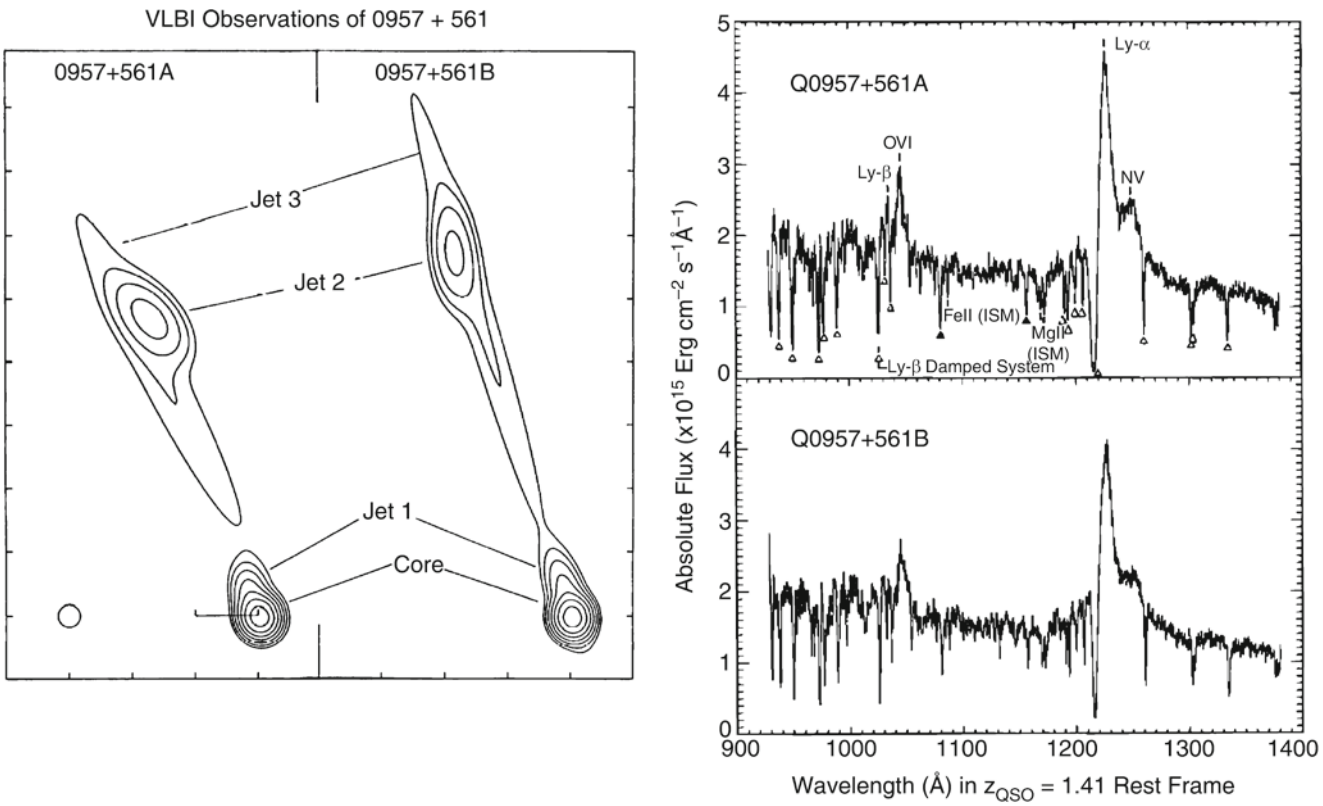


Fig. 3.57 *Left:* milliarcsecond structure of the two images of the quasar QSO 0957+561, a VLBI map at 13 cm wavelength. Both quasar images show a core-jet structure, and it is clearly seen that they are mirror-symmetric, as predicted by lens models. *Right*: spectra of the two quasar images QSO 0957+561A,B, observed by the Faint Object Camera (FOC) on-board HST. The similarity of the spectra, in particular the identical redshift, is a clear indicator of a common source of the two quasar images. The broad Lyα-line, in the wings of which an NV-line is visible, is virtually always the strongest emission line in quasars. Source: *Left*: M. Gorenstein et al. 1988, *VLBI observations of the gravitational lens system 0957+561—Structure and relative magnification of the A and B images*, ApJ 334, 42, p. 53, Fig. 5. ©AAS. Reproduced with permission. *Right*: A.G. Michalitsianos et al. 1997, *Ly alpha Absorption-Line Systems in the Gravitational Lens Q0957+561*, ApJ 474, 598, p. 599, Fig. 1. ©AAS. Reproduced with permission

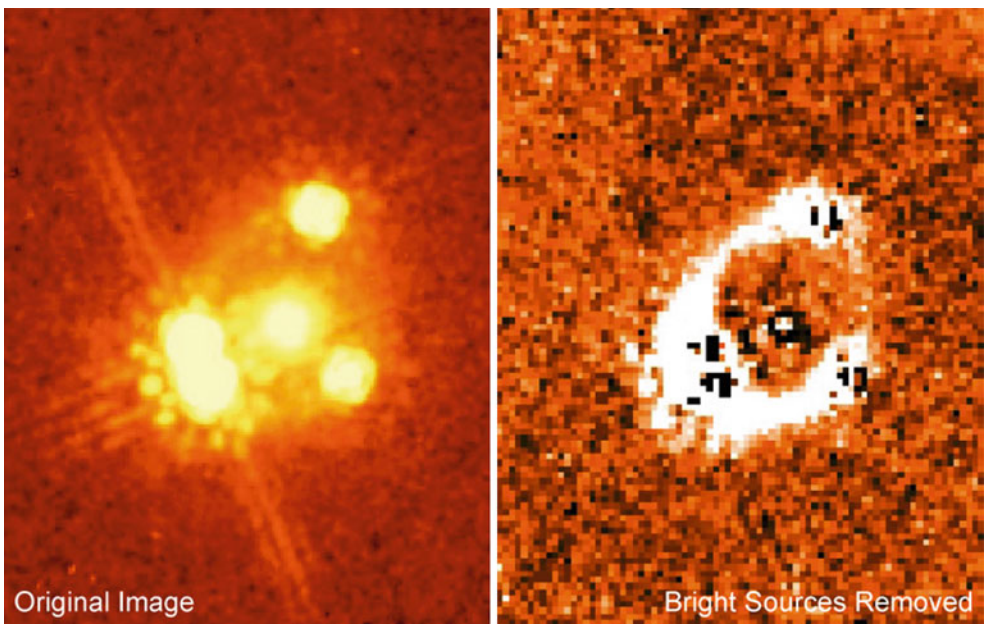


Fig. 3.58 A NIR image of QSO 1115+080 is shown *on the left*, as observed with the NICMOS camera on-board HST. The double structure of image A (the *left* of the QSO images) is clearly visible, although the image separation of the two A components is less than 0.″5. The lens galaxy, located in the 'middle' of the QSO images, has a much redder spectral energy distribution than the quasar images. In the *right-hand panel*, the quasar images and the lens galaxy have been subtracted. What remains is a nearly closed ring; the light of the galaxy which hosts the active galactic nucleus is imaged into an Einstein ring. Credit: C. Impey (University of Arizona) & NASA

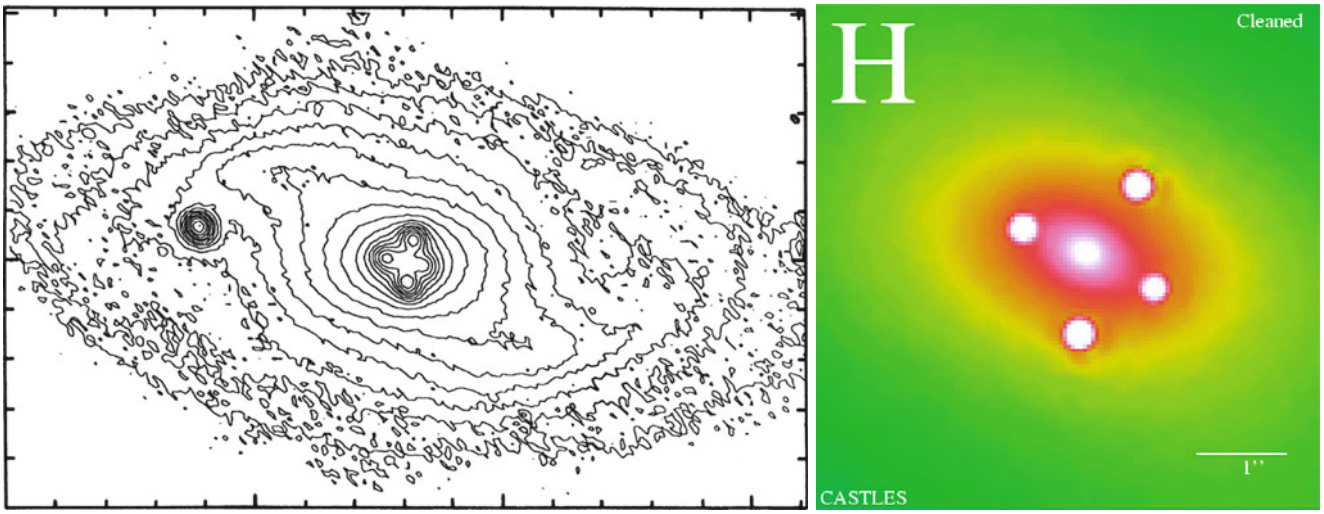


Fig. 3.59 *Left*: in the center of a nearby spiral galaxy, four point-like sources were found whose spectra show strong emission lines. This image from the CFHT clearly shows the bar structure in the core of the lens galaxy. An HST/NICMOS image of the center of QSO 2237+0305 is shown *on the right*. The central source is not a fifth quasar image but rather the bright nucleus of the lens galaxy. Credit: *Left*: H.K.C. Yee 1988, *High-resolution imaging of the gravitational lens system candidate 2237+030*, AJ 95, 1331, p. 1332, Fig. 4. ©AAS. Reproduced with permission. *Right*: CASTLES-Collaboration, C.S. Kochanek

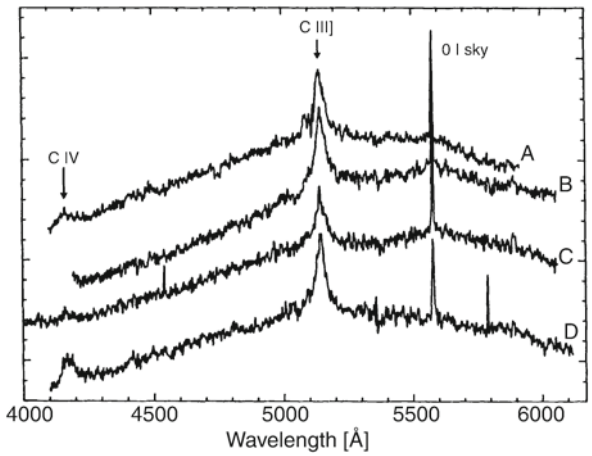


Fig. 3.60 Spectra of the four images of the quasar 2237+0305, observed with the CFHT. As is clearly visible, the spectral properties of these four images are very similar; this is the final proof that we are dealing with a lens system here. Measuring the individual spectra of these four very closely spaced sources is extremely difficult and can only be performed under optimum observing conditions. Source: G. Adam et al. 1989, *Observations of the Einstein Cross 2237+030 with the TIGER Integral Field Spectrograph*, A&A 208, L15, p. L17, Fig. 6. ©ESO. Reproduced with permission

at the position of the compact radio component and, in addition, a bright elliptical galaxy right in the center of the ring-shaped radio lobe. This galaxy has a significantly lower redshift than the quasar and hence is the gravitational lens responsible for imaging the lobe into an Einstein ring. The SLACS survey has found a large number of Einstein rings, as shown in Fig. 3.63.

3.11.4 Applications of the lens effect

Mass determination. As mentioned previously, the mass within a system of multiple images can be determined directly, sometimes very precisely. Even without a specific model, an estimate of the Einstein radius from the location of the multiple images immediately yields a mass estimate from (3.81). Its accuracy depends on the detailed image configuration but can be substantially better than 10 % for quadruple image systems. Once the mass distribution is quantitatively modeled, such as it reproduces the observed image positions, the mass within the Einstein radius can be determined with very high accuracy.

Since the length scale in the lens plane (at given angular scale) and $\Sigma_{\rm cr}$ depend on H_0, these mass estimates scale with H_0. For instance, for QSO 2237+0305, a mass within $0\farcs9$ of $(1.08 \pm 0.02) h^{-1} \times 10^{10} M_\odot$ is derived. An even more precise determination of the mass was obtained for the lens galaxy of the Einstein ring in the system MG 1654+13 (Fig. 3.62). The dependence on the other cosmological parameters is comparatively weak, especially at low redshifts of the source and the lens. Most lens galaxies are early-type galaxies (ellipticals); from the determination of their mass it is concluded that ellipticals contain dark matter, as spirals do. For example, the fraction of dark matter inside the Einstein ring 1938+666 (Fig. 3.61) is 0.20 if a Salpeter initial mass function is assumed, but increases to 0.55 if a (more realistic) IMF is used which flattens for masses below $\sim 1 M_\odot$.

Similar results have been obtained for large number of lens galaxies. In the lens system 1933+503, a three-component radio source is lensed into a total of ten images

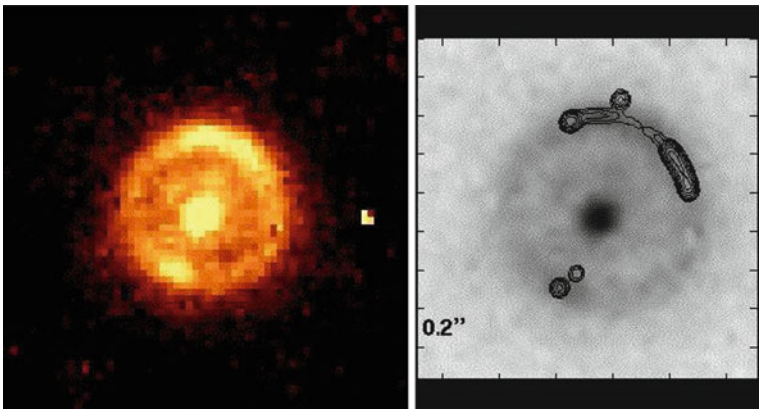


Fig. 3.61 The radio source B1938+666 with $z_s = 2.059$ is seen to be multiply imaged (contours in the *right-hand part*); here, the radio source consists of two components, one of which is imaged fourfold, the other twofold. A NIR image taken with the NICMOS camera on-board the HST (*left-hand part*, also shown on the right in gray-scale) shows the lens galaxy ($z_d = 0.88$) in the center of an Einstein ring that originates from the stellar light of the host galaxy of the active galactic nucleus. Credit: L.J. King, based on data from King et al. 1998, *A complete infrared Einstein ring in the gravitational lens system B1938+666*, MNRAS 295, L41

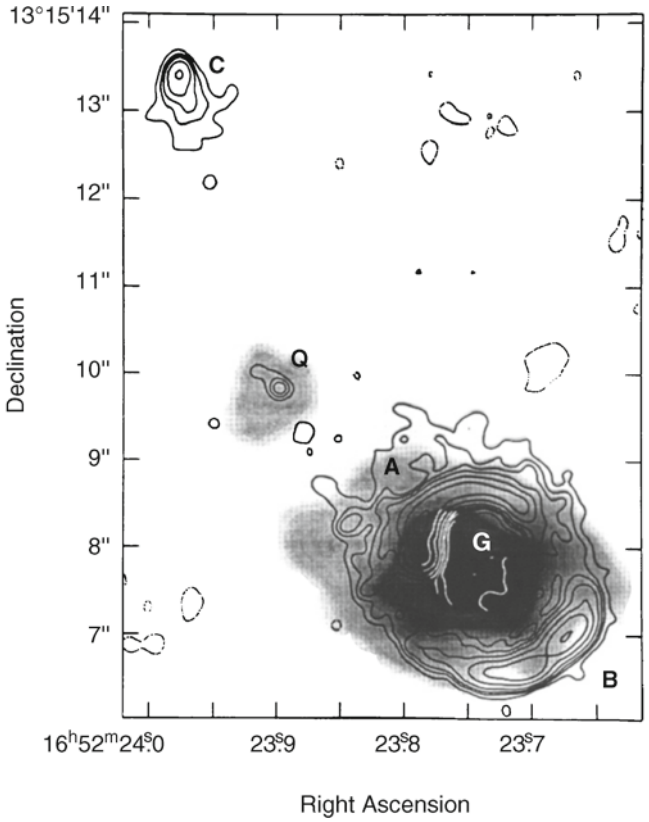


Fig. 3.62 The quasar MG 1654+13 shows, in addition to the compact radio core (Q), two radio lobes; the northern lobe is denoted by C, whereas the southern lobe is imaged into a ring. An optical image is displayed in *gray-scales*, showing not only the quasar at Q ($z_s = 1.72$) but also a massive foreground galaxy (G) at $z_d = 0.25$ that is responsible for the lensing of the lobe into an Einstein ring. The mass of this galaxy within the ring can be derived with a precision of $\sim 1\,\%$. Credit: G. Langston, based on data from Langston et al. 1989, *MG 1654+1346 - an Einstein Ring image of a quasar radio lobe*, AJ 97, 1283

by a spiral lens galaxy, where two of the radio components have four images, the third one has two. Using the observed rotation curve of the spiral, a clear decomposition of the baryonic matter (located in a disk) and the dark matter (distributed in an extended halo) became feasible. The dark matter fraction projected inside the effective radius in this lens galaxy is about 40 % for a realistic initial mass function. Comparing the light distribution of the lens, and translating this into a stellar mass, using stellar population synthesis, it is found that a bottom-heavy IMF (like Salpeter) is strongly disfavored compared to those with a flattening at low masses, and analogous results were found for other spiral lens galaxies as well. At least for one early-type lens galaxy, a bottom-heavy IMF like the Salpeter function is actually ruled out, as otherwise the stellar mass would exceed the lensing mass.

Mass profile and dark matter fraction. Whereas one can determine the mass within the Einstein radius with high accuracy, in a typical lens configuration one cannot say much about the density profile. There are special lens systems where this becomes possible, namely those where the images span a large range in separation from the lens center. But even in those systems, conclusions about the slope of the profile are not necessarily very robust. An exception to this occurs for lens systems where two sources at different redshift are lensed by the same galaxy; in this case, one has two Einstein radii (one corresponding to each source redshift), and one can determine the masses at two different radii.

However, if it is possible to combine the mass estimate within the Einstein radius with a different mass estimate at a different radius, we can obtain information about the density profile. Another mass estimate is obtained from the stellar kinematics in the (early-type) lens galaxies, studied via spectroscopy. The velocity dispersion of stars, which

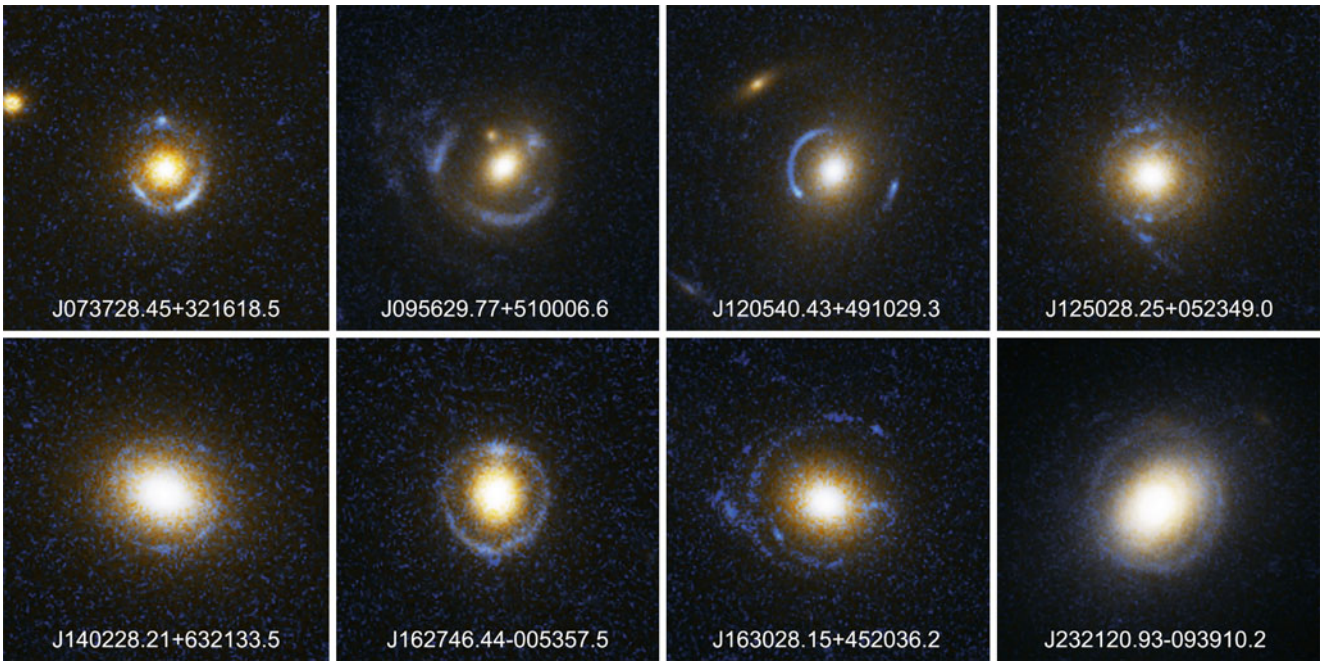


Fig. 3.63 Eight strong lens systems from the SLACS survey. In contrast to multiply imaged quasars, the images of the lensed source are extended, and often highly elongated or even mapped onto a full Einstein ring. Such extended images can probe the gravitational potential of the lens at far more locations than a few point-like images, and thus potentially provide more information about the mass distribution. Credit: NASA, ESA, and the SLACS Survey team: A. Bolton (Harvard/ Smithsonian), S. Burles (MIT), L. Koopmans (Kapteyn), T. Treu (UCSB), and L. Moustakas (JPL/Caltech)

determines the absorption line width, depends on the local gravitational potential via the virial theorem, and is typically estimated at the effective radius. Thus, if the effective radius is significantly different from the Einstein radius, which is the case in most lens systems, the slope of the mass profile can be estimated. It turns out that the isothermal profile, $\rho \propto r^{-2}$, is a very good description for most lens galaxies, with only small variations of the slope being observed.

This is a very surprising result! To see why, let us recall that the brightness profile of ellipticals is approximated by a de Vaucouleurs profile, which differs substantially from a projected isothermal profile. Thus, the first conclusion from this is that the mass profile does not follow the light profile; hence, beside the stars, there must be an additional mass component in these galaxies. From the spectral energy distribution of the stellar population in the lens galaxies, one can estimate the mass-to-light ratio of the stellar population, hence the stellar mass within the Einstein radius. Comparing with the mass determined from lensing, one finds that about half the mass within the Einstein radius is stellar; the other half is dark matter. In agreement with what was said before, the fraction of dark matter inside the Einstein radius varies between ~ 30 and $\sim 70\,\%$, with more massive lenses having a higher dark matter fraction. As we will see in Sect. 7.6.1, the mass profile of the dark matter is predicted to differ substantially from an isothermal profile, at least at small scales. However, the results about the mass profile tell us that the distributions of stars and of dark matter conspire in such a way that the sum of them is approximately isothermal. Needless to say that this results is an important constraint for the theory of galaxy formation and evolution.

Mass fundamental plane. One finds from the large sample of SLACS lenses that the mass-to-light ratio of lens galaxies increases with mass, in concordance with what was discussed in relation with the fundamental plane (Sect. 3.4.3). In fact, lensing accurately measures the mean surface mass density within the Einstein radius—see (3.81). For the SLACS lenses, the typical Einstein radius is about $0.6R_{\rm e}$. Together with the fact that the slope of the mass profile is isothermal with good accuracy, the mean surface mass density within half the effective radius can be determined. This allows us to write a 'fundamental plane'-relation in terms of the surface mass density, instead of the surface brightness. From the virial theorem, one would conclude

$$\sigma_0^2 \propto M/R \propto \Sigma_{\rm e2}\, R_{\rm e}\ ,$$

where we specialized $R = R_{\rm e}/2$, and $\Sigma_{\rm e2}$ is the mean surface mass density within half the effective radius. In other words, this implies a relation of the form

$$R_{\rm e} \propto \sigma_0^a\, \Sigma_{\rm e2}^b\ , \tag{3.82}$$

with $a = 2$ and $b = -1$. Indeed, the SLACS lenses define such a mass-based fundamental plane, with $a \approx 2$ and $b \approx -1$, and their dispersion about this mass-based fundamental plane is even slightly smaller than that around the standard fundamental plane (3.30), based on luminosity. This shows that the tilt of the fundamental plane is indeed due to a varying mass-to-light ratio as a function of galaxy mass, as described by (3.32).

Shape of the mass distribution. From modeling gravitational lens systems, the ellipticity of the mass distribution and the orientation of the major axis can be determined.[14] It is interesting to study whether ellipticity and orientation of the mass agree with that of the light distribution. Indeed, this is the case: The orientation of the mass distribution agrees with that of the light distribution, with a dispersion of the difference between the two position angles of $\sim 10°$. The same holds for the ellipticity, or axis ratio; in early-type lens galaxies, the axis ratios of mass and light agree to within 10 %.

These are no trivial statements, given that about half the mass inside the Einstein radius is dark. The shape of the dark matter distribution must be quite similar to, but not identical as that of the stars. Thus, whereas the radial density profile of dark matter and stars are quite different, their shapes are similar.

Environmental effects. Detailed lens models show that the light deflection of most gravitational lenses is affected by an external tidal field. This is due to the fact that lens galaxies are often members of galaxy groups which contribute to the light deflection as well. In some cases the members of the group were identified. Mass properties of the corresponding group can be derived from the strength of this external influence.

Determination of the Hubble constant. The light travel times along the different paths (according to the multiple images) are not the same. On the one hand the paths have different geometrical lengths, and on the other hand the light rays traverse different depths of the gravitational potential of the lens, resulting in a (general relativistic) time dilation effect. The difference in the light travel times Δt is measurable because luminosity variations of the source are observed at different times in the individual images. Δt can be measured from this difference in arrival time, called the time delay.

It is easy to see that Δt depends on the Hubble constant, or in other words, on the size of the Universe. If a universe is twice the size of our own, Δt would be twice as large as well—see Fig. 3.64. Thus if the mass distribution of the lens can be modeled sufficiently well, by modeling the geometry

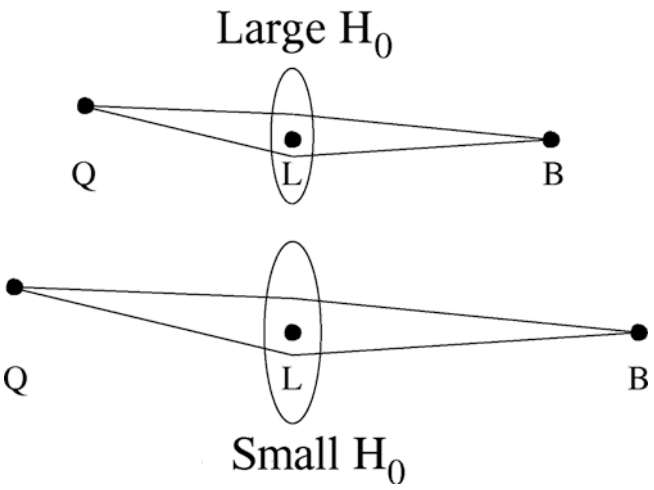


Fig. 3.64 Lens geometry in two universes with different Hubble constant. All observables are dimensionless—angular separations, flux ratios, redshifts—except for the difference in the light travel time. This is larger in the universe at the *bottom* than in the one at the *top*; hence, $\Delta t \propto H_0^{-1}$. If the time delay Δt can be measured, and if one has a good model for the mass distribution of the lens, then the Hubble constant can be derived from measuring Δt. Source: R. Narayan & M. Bartelmann 1996, *Lectures on Gravitational Lensing*, astro-ph/9606001

of the image configuration, then the Hubble constant can be derived from measuring the difference in the light travel time. To date, Δt has been measured in about 20 lens systems (see Fig. 3.65 for an example). Based on 'plausible' lens models we can derive values for the Hubble constant. Early results obtained with that method often yielded rather small values of H_0, probably because of employing too simple mass models for the lens. The main difficulty here is that the mass distribution in lens galaxies cannot unambiguously be derived from the positions of the multiple images.

However, much more detailed models are feasible for lens systems where an extended source component is lensed, e.g., into an Einstein ring, in addition to multiple images of a compact component. Recently, results from two such detailed modeling efforts became available, resulting in $H_0 \approx 71$ and $79\,\mathrm{km\,s^{-1}Mpc^{-1}}$, respectively, with an estimated error of about 5 %. These measurements are in good agreement with those from the distance ladder—see (3.50). We should note, however, that the determination of H_0 from time delay lenses is affected by the so-called mass-sheet degeneracy (see Problem 3.5), which may lead to an increased error budget. In Sect. 6.4.4 we will discuss the value of H_0 determinations from lens time delays in a slightly different context.

The ISM in lens galaxies. Since the same source is seen along different sight lines passing through the lens galaxy, the comparison of the colors and spectra of the individual images provides information on reddening and on dust extinction in the ISM of the lens galaxy. From such investigations it was shown that the extinction in ellipticals is in

[14] As it turns out, these parameters are more accurately determined from lens models than the slope of the mass profile.

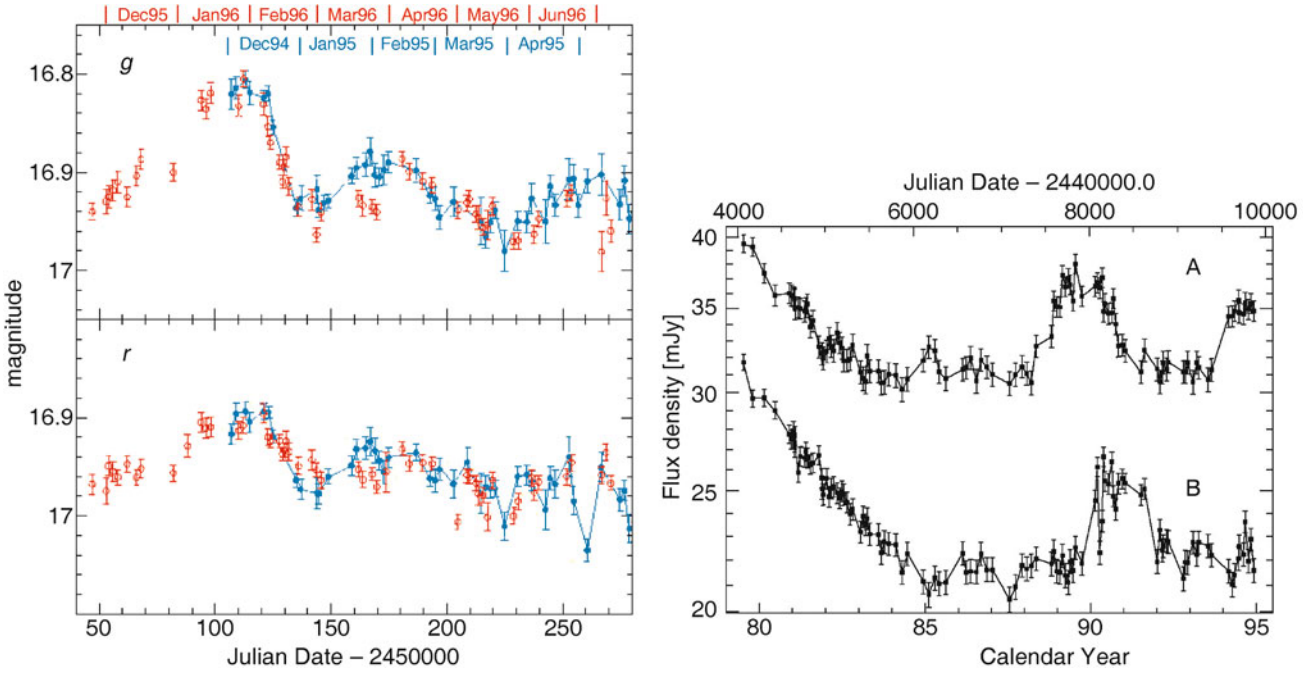


Fig. 3.65 *Left*: optical light curves of the double quasar 0957+561 in two broad-band filters. The light curve of image A is displayed in *red* and that of image B in *blue*, where the latter is shifted in time by 417 days. With this shift, the two light curves are made to coincide—this light travel time difference of 417 days is determined with an accuracy of $\sim \pm 3$ days. *Right*: radio light curves of QSO 0957+561A,B at 6 cm. From these radio measurements Δt can also be measured, and the corresponding value is compatible with that obtained from optical data.

Source: *Left*: T. Kundić et al. 1997, *A Robust Determination of the Time Delay in 0957+561A, B and a Measurement of the Global Value of Hubble's Constant*, ApJ 482, 75, p. 79, Fig. 3. ©AAS. Reproduced with permission. *Right*: D.B. Haarsma et al. 1997, *The 6 Centimeter Light Curves of B0957+561, 1979–1994: New Features and Implications for the Time Delay*, ApJ 479, 102, p. 104, Fig. 1. ©AAS. Reproduced with permission

fact very low, as is to be expected from the small amount of interstellar medium they contain, whereas the extinction is considerably higher for spirals. These analyses also enable us to study the relation between extinction and reddening, and from this to search for deviations from the Galactic reddening law (2.21)—see Fig. 2.7. In fact, the constant of proportionality R_V is different in other galaxies, indicating a different composition of the dust, e.g., with respect to the chemical composition and to the size distribution of the dust grains.

3.12 Problems

3.1. Central surface brightness of disk galaxies. Assume the validity of Freeman's law, and consider a spiral galaxy with central surface brightness $\mu_0 = 21.5\,\mathrm{mag/arcsec^2}$ at the distance of the Virgo cluster, i.e., $D = 16\,\mathrm{Mpc}$.

1. With the absolute magnitude of the Sun in the B-band of $M_{\odot,\mathrm{B}} = 5.54$, calculate the central surface brightness of the galaxy in Solar luminosities per pc^2.
2. The disk is seen to have an exponential surface brightness, with angular scale length of $\theta_R = 50''$. What is the total luminosity of the disk?

3.2. Properties of the Salpeter IMF. Let us assume that the stellar mass function has the same shape as the Salpeter IMF (3.36), with $m_\mathrm{L} = 0.1 M_\odot$ and $m_\mathrm{U} = 70 M_\odot$. We define m_{m50} such that half of the stellar mass is contained in stars with mass below m_{m50}, the other half in stars with $m > m_{m50}$. Similarly, we define m_{L50} such that half of the luminosity from the stellar population is due to stars with $m < m_{L50}$. Calculate the masses m_{m50} and m_{L50}, the latter by assuming that the luminosity of stars scales with mass as $L \propto m^3$. From a comparison of m_{m50} and m_{L50}, draw conclusions about the relative importance of low- and high-mass stars for the mass budget and the luminosity of the stellar population.

3.3. Observable supernova rate. The rate of Type Ia supernovae explosions is about $3 \times 10^{-5}\,\mathrm{Mpc^{-3}\,yr^{-1}}$. Assume that a photometric supernova survey is carried out with a sensitivity which allows the detection of these sources out to a distance of 500 Mpc. How many square degrees of the sky need to be surveyed in order to find 10 SNe Ia per year?

3.4. Obtaining the luminosity function of galaxies. A galaxy survey is carried out over a solid angle ω, and only objects with distance $\leq D_\mathrm{lim}$ shall be considered. The galaxy survey is flux-limited, which means that only sources with flux above a threshold, $S \geq S_\mathrm{min}$, can be observed.

1. Show that the total volume in which galaxies are considered for the survey is $V_{\rm tot} = D_{\rm lim}^3\,\omega/3$.
2. Calculate the volume $V_{\rm max}(L)$ within which we can observe galaxies with luminosity L.
3. Let $N(L)$ be the number of galaxies found with luminosity smaller than L. Show that the luminosity function is then determined as

$$\Phi(L) = \frac{1}{V_{\rm max}(L)}\,\frac{{\rm d}N(L)}{{\rm d}L}\,. \tag{3.83}$$

3.5. Mass-sheet degeneracy. For a given gravitational lens system, suppose you have a perfect model: a surface mass density profile $\kappa(\boldsymbol{\theta})$ such that the corresponding scaled deflection angle $\boldsymbol{\alpha}(\boldsymbol{\theta})$, inserted into the lens equation (3.64), yields a solution for all image positions $\boldsymbol{\theta}_i$, i.e., there exist a source position $\boldsymbol{\beta}$ such that $\boldsymbol{\beta} = \boldsymbol{\theta}_i - \boldsymbol{\alpha}(\boldsymbol{\theta}_i)$ for all images i.

1. Consider now a family of lens models, described by the surface mass density

$$\kappa_\lambda(\boldsymbol{\theta}) = (1-\lambda) + \lambda\kappa(\boldsymbol{\theta})\,. \tag{3.84}$$

Thus, $\kappa_\lambda(\boldsymbol{\theta})$ is obtained by scaling the original mass distribution by a factor λ, and adding a constant surface mass density of amplitude $(1-\lambda)$. Calculate the scaled deflection angle $\boldsymbol{\alpha}_\lambda(\boldsymbol{\theta})$ corresponding to the surface mass density κ_λ.

2. Derive the lens equation corresponding to the new surface mass density (3.84) and show that there exists a source position $\boldsymbol{\beta}_\lambda = \boldsymbol{\beta}/\lambda$ such that the image positions $\boldsymbol{\theta}_i$ all satisfy the new lens equation. Hence, the new mass distribution κ_λ describes the image position equally well as the original distribution κ, for all $\lambda \neq 0$. This implies that from image positions alone, one can not distinguish between κ and κ_λ.
3. Calculate the magnification $\mu_\lambda(\boldsymbol{\theta})$ for the new mass distribution in terms of the magnification $\mu(\boldsymbol{\theta})$ of the original mass distribution. Show that one cannot distinguish between these mass distributions from considering the flux ratios (which is the same as the magnification ratios) of the images.

Cosmology I: Homogeneous isotropic world models

4

We will now begin to consider the Universe as a whole. Individual objects such as galaxies and stars will no longer be the subject of discussion, but instead we will turn our attention to the space and time in which these objects are embedded. These considerations will then lead to a world model, the model of our cosmos. We need such a model also to interpret the observations of distant objects, i.e., those with a redshift for which the local Hubble law (1.2) ceases to be valid.

This chapter will deal with aspects of homogeneous cosmology. As we will see, the Universe can, to first approximation, be considered as being homogeneous. At first sight this fact obviously seems to contradict observations because the world around us is highly inhomogeneous and structured. Thus the assumption of homogeneity is certainly not valid on small scales. But observations are compatible with the assumption that the Universe is homogeneous when averaged over large spatial scales. Aspects of inhomogeneous cosmology, and thus the formation and evolution of structures in the Universe, will be considered later in Chap. 7.

4.1 Introduction and fundamental observations

Cosmology is a very special science indeed. To be able to appreciate its peculiar role we should recall the typical way of establishing knowledge in natural sciences. It normally starts with the observation of some regular patterns, for instance the observation that the height h a stone falls through is related quadratically to the time t it takes to fall, $h = (g/2)t^2$. This relation is then also found for other objects and observed at different places on Earth. Therefore, this relation is formulated as the 'law' of free fall. The constant of proportionality $g/2$ in this law is always the same. This law of physics is tested by its prediction of how an object falls, and wherever this prediction is tested it is confirmed—disregarding the resistance of air in this simple example, of course.

Relations become physical laws if the predictions they make are confirmed again and again; the validity of such a law is considered more secure the more diverse the tests have been. The law of free fall was tested only on the surface of the Earth and it is only valid there with this constant of proportionality.[1] In contrast to this, Newton's law of gravity contains the law of free fall as a special case, but it also describes the free fall on the surface of the Moon, and the motion of planets around the Sun. If only a single stone was available, we would not know whether the law of free fall is a property of this particular stone or whether it is valid more generally.

In some ways, cosmology corresponds to the latter example: we have only one single Universe available for observation. Relations that are found in our cosmos cannot be verified in other universes. Thus it is not possible to consider any property of our Universe as 'typical'—we have no statistics on which we could base a statement like this. Despite this special situation, enormous progress has been made in understanding our Universe, as we will describe here and in subsequent chapters.

Cosmological observations are difficult in general, simply because the majority of the Universe (and with it most of the sources it contains) is very far away from us. Distant sources are very dim. This explains why our knowledge of the Universe runs in parallel with the development of large telescopes and sensitive detectors. Much of today's knowledge of the distant Universe became available only with the new generation of optical telescopes of the 8-m class, as well as new and powerful telescopes in other wavelength regimes.

The most important aspect of cosmological observations is the finite speed of light. We observe a source at distance D in an evolutionary state at which it was $\Delta t = (D/c)$ younger than today. Thus we can observe the current state of the Universe only very locally. Another consequence of this effect, however, is of even greater importance: due to the finite speed of light, it is possible to look back into the past.

[1] Strictly speaking, the constant of proportionality g depends slightly on the location.

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_4,

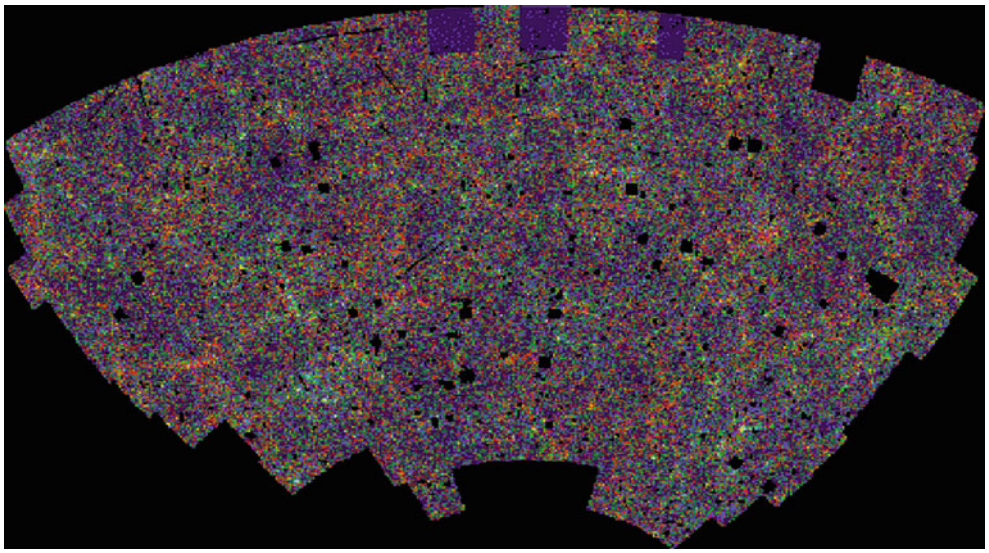


Fig. 4.1 The APM-survey: galaxy distribution in a $\sim 100 \times 50\,\text{degree}^2$ field around the South Galactic Pole. The intensities of the pixels are scaled with the number of galaxies per pixel, i.e., the projected galaxy number density on the sphere. The 'holes' are regions around bright stars, globular clusters etc., that were not surveyed. Credit: S. Maddox, W. Sutherland, G. Efstathiou & J. Loveday, with follow-up by G. Dalton, and Astrophysics Dept., Oxford University

At a distance of ten billion light years we observe galaxies in an evolutionary state when the Universe had only a third of its current age. Although we cannot observe the past of our own Milky Way, we can study that of other galaxies. If we are able to identify among them the ones that will form objects similar to our Galaxy in the course of cosmic evolution, we will be able to learn a great deal about the typical evolutionary history of such spirals.

The finite speed of light in a Euclidean space, in which we are located at the origin $\boldsymbol{r} = 0$ today ($t = t_0$), implies that we can only observe points in spacetime for which $|\boldsymbol{r}| = c(t_0 - t)$; an arbitrary point $(\boldsymbol{r}, t)$ in spacetime is not observable. The set of points in spacetime which satisfy the relation $|\boldsymbol{r}| = c(t_0 - t)$ is called our *backward light cone*.

The fact that our astronomical observations are restricted to sources which are located on our backward light cone implies that our possibilities to observe the Universe are fundamentally limited. If somewhere in spacetime there would be a highly unusual event, we will not be able to observe it unless it happens to lie on our backward light cone. Only if the Universe has an essentially 'simple' structure will we be able to understand it, by combining astronomical observations with theoretical modeling. Luckily, our Universe seems to be basically simple in this sense.

4.1.1 Fundamental cosmological observations

We will begin with a short list of key observations that have proven to be of particular importance for cosmology. Using these observational facts we will then be able to draw a number of immediate conclusions; other observations will be explained later in the context of a cosmological model.

1. The sky is dark at night (Olbers' paradox).
2. Averaged over large angular scales, faint galaxies (e.g., those with $R > 20$) are uniformly distributed on the sky (see Fig. 4.1).
3. With the exception of a very few very nearby galaxies (e.g., Andromeda = M31), a redshift is observed in the spectra of galaxies—most galaxies are moving away from us, and their escape velocity increases linearly with distance (Hubble law; see Fig. 1.13).
4. In nearly all cosmic objects (e.g., gas nebulae, main sequence stars), the mass fraction of helium is 25–30 %.
5. The oldest star clusters in our Galaxy have an age of $\sim 12\,\text{Gyr} = 12 \times 10^9\,\text{yr}$ (see Fig. 4.2).
6. A microwave radiation (cosmic microwave background radiation, CMB) is observed, reaching us from all directions. This radiation is isotropic except for very small, but immensely important, fluctuations with relative amplitude $\sim 10^{-5}$ (see Fig. 1.21).
7. The spectrum of the CMB corresponds, within the very small error bars that were obtained by the measurements with COBE, to that of a perfect blackbody, i.e., a Planck radiation of a temperature of $T_0 = 2.728 \pm 0.004\,\text{K}$—see Fig. 4.3.
8. The number counts of radio sources at high Galactic latitude does *not* follow the simple law $N(> S) \propto S^{-3/2}$ (see Fig. 4.4).

4.1.2 Simple conclusions

We will next draw a number of simple conclusions from the observational facts listed above. These will then serve as a motivation and guideline for developing the cosmological model. We will start with the assumption of an infinite,

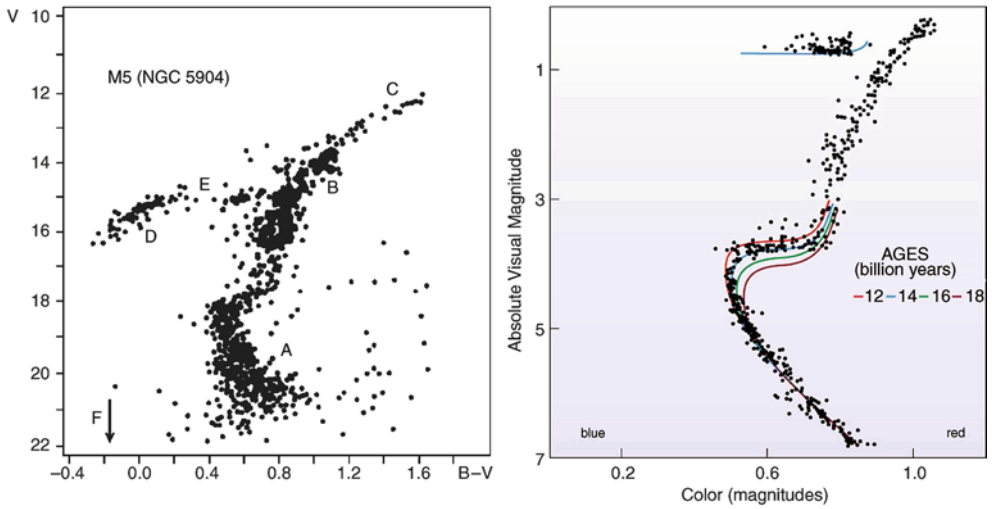


Fig. 4.2 *Left panel*: Color-magnitude diagram of the globular cluster M 5. The different sections in this diagram are labeled. A: main sequence; B: red giant branch; C: point of helium flash; D: horizontal branch; E: Schwarzschild-gap in the horizontal branch; F: white dwarfs, below the *arrow*. At the point where the main sequence turns over to the red giant branch (called the 'turn-off point'), stars have a mass corresponding to a main-sequence lifetime which is equal to the age of the globular cluster (see Appendix B.3). Therefore, the age of the cluster can be determined from the position of the turn-off point by comparing it with models of stellar evolution. *Right panel*: Isochrones, i.e., curves connecting the stellar evolutionary position in the color-magnitude diagram of stars of equal age, are plotted for different ages and compared to the stars of the globular cluster 47 Tucanae. Such analyses reveal that the oldest globular clusters in our Milky Way are about 12 billion years old, where different authors obtain slightly differing results—details of stellar evolution may play a role here. The age thus obtained also depends on the distance of the cluster. A revision of these distances by the Hipparcos satellite led to a decrease of the estimated ages by about two billion years. Credit: M5: ©Leos Ondra; 47 Tuc: J.E. Hesser, W.E. Harris, D.A. Vandenberg, J.W.B. Allwright, P. Scott & P.B. Stetson 1987, *A CCD color-magnitude study of 47 Tucanae*, PASP 99, 739

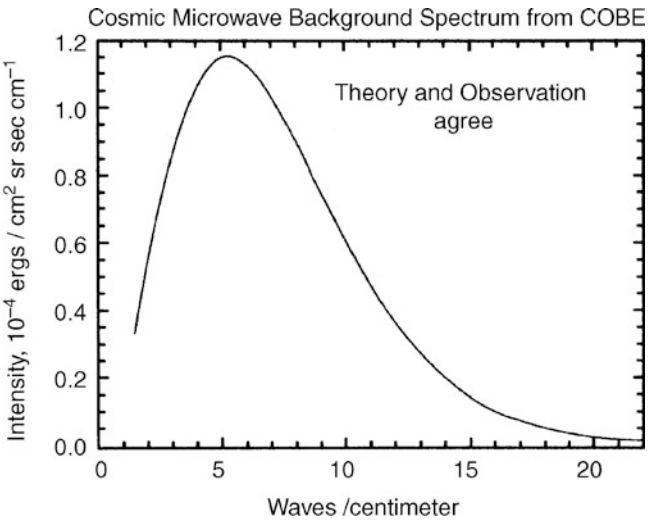


Fig. 4.3 CMB spectrum, plotted as intensity vs. frequency, measured in waves per centimeter. The *solid line* shows the expected spectrum of a blackbody of temperature $T = 2.728\,\mathrm{K}$. The error bars of the data, observed by the FIRAS instrument on-board COBE, are so small that the data points with error bars cannot be distinguished from the theoretical curve. Credit: COBE, NASA. We acknowledge the use of the Legacy Archive for Microwave Data Analysis (LAMBDA). Support for LAMBDA is provided by the NASA Office for Space Science

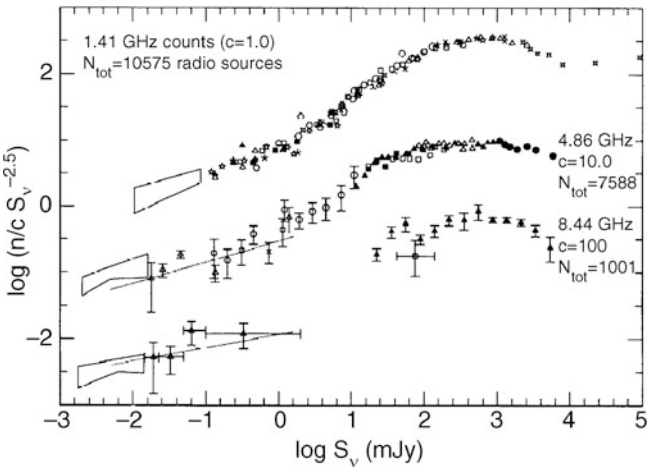


Fig. 4.4 Number counts of radio sources as a function of their flux, normalized by the Euclidean expectation $N(S) \propto S^{-5/2}$, corresponding to the integrated counts $N(>S) \propto S^{-3/2}$. Counts are displayed for three different frequencies; they clearly deviate from the Euclidean expectation. Source: R.A. Windhorst et al. 1993, *Microjansky source counts and spectral indices at 8.44 GHz*, ApJ 405, 498, p. 508, Fig. 3. ©AAS. Reproduced with permission

on-average homogeneous, Euclidean, static universe, and show that this assumption is in direct contradiction to observations (1) and (8).

Olbers' paradox (1): We can show that the night sky would be bright in such a universe—uncomfortably bright, in fact. Let n_* be the mean number density of stars, constant in space and time according to the assumptions, and let R_* be their mean radius. A spherical shell of radius r and thickness dr

around us contains $n_* \, \mathrm{d}V = 4\pi r^2 \, \mathrm{d}r \, n_*$ stars. Each of these stars subtends a solid angle of $\pi R_*^2 / r^2$ on our sky, so the stars in the shell cover a total solid angle of

$$\mathrm{d}\omega = 4\pi r^2 \, \mathrm{d}r \, n_* \, \frac{R_*^2 \pi}{r^2} = 4\pi^2 \, n_* \, R_*^2 \, \mathrm{d}r \; . \tag{4.1}$$

We see that this solid angle is independent of the radius r of the spherical shell because the solid angle covered by a single star $\propto r^{-2}$ just compensates the volume of the shell $\propto r^2$. To compute the total solid angle of all stars in a static Euclidean universe, (4.1) has to be integrated over all distances r, but the integral

$$\omega = \int_0^\infty \mathrm{d}r \, \frac{\mathrm{d}\omega}{\mathrm{d}r} = 4\pi^2 \, n_* \, R_*^2 \int_0^\infty \mathrm{d}r$$

diverges. Formally, this means that the stars cover an infinite solid angle, which of course makes no sense physically. The reason for this divergence is that we disregarded the effect of overlapping stellar disks on the sphere. However, these considerations demonstrate that the sky would be completely filled with stellar disks, i.e., from any direction, along any line-of-sight, light from a stellar surface would reach us. Since the specific intensity I_ν is independent of distance—the surface brightness of the Sun as observed from Earth is the same as seen by an observer who is much closer to the Solar surface—the sky would have a temperature of $\sim 10^4$ K; fortunately, this is not the case!

Source counts (8): Consider now a population of sources with a luminosity function that is constant in space and time, i.e., let $n(>L)$ be the spatial number density of sources with luminosity larger than L. A spherical shell of radius r and thickness $\mathrm{d}r$ around us contains $4\pi r^2 \, \mathrm{d}r \, n(>L)$ sources with luminosity larger than L. Because the observed flux S is related to the luminosity via $L = 4\pi \, r^2 \, S$, the number of sources with flux $> S$ in this spherical shell is given as $\mathrm{d}N(>S) = 4\pi r^2 \, \mathrm{d}r \, n(> 4\pi \, r^2 \, S)$, and the total number of sources with flux $> S$ results from integration over the radii of the spherical shells,

$$N(>S) = \int_0^\infty \mathrm{d}r \; 4\pi \, r^2 \, n(> 4\pi \, r^2 \, S) \; .$$

Changing the integration variable to $L = 4\pi \, r^2 \, S$, or $r = \sqrt{L/(4\pi S)}$, with $\mathrm{d}r = \mathrm{d}L/(2\sqrt{4\pi L S})$, yields

$$\begin{aligned} N(>S) &= \int_0^\infty \frac{\mathrm{d}L}{2\sqrt{4\pi L S}} \, \frac{L}{S} \, n(>L) \\ &\propto S^{-3/2} \int_0^\infty \mathrm{d}L \, \sqrt{L} \, n(>L) \; . \end{aligned} \tag{4.2}$$

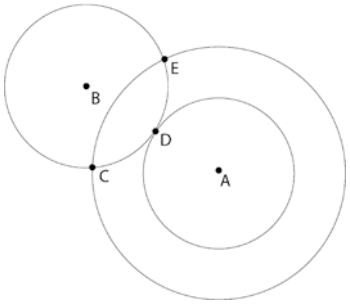


Fig. 4.5 Homogeneity follows from the isotropy around two points. If the Universe is isotropic around observer B, the densities at C, D, and E are equal. Drawing spheres of different radii around observer A, it is seen that the region within the spherical shell around A has to be homogeneous. By varying the radius of the shell, we can conclude the whole Universe must be homogeneous. Credit: J.A. Peacock 1999, *Cosmological Physics*, Cambridge University Press

From this result we deduce that the source counts in such a universe is $N(>S) \propto S^{-3/2}$, independent of the luminosity function. This is in contradiction to the observations.

From these two contradictions—Olbers' paradox and the non-Euclidean source counts—we conclude that at least one of the assumptions must be wrong. Our Universe cannot be all four of Euclidean, homogeneous, infinite, and static. The Hubble flow, i.e., the redshift of galaxies, indicates that the assumption of a static Universe is wrong.

The **age of globular clusters (5)** requires that the Universe is at least 12 Gyr old because it cannot be younger than the oldest objects it contains. Interestingly, the age estimates for globular clusters yield values which are very close to the *Hubble time* $H_0^{-1} = 9.78 \, h^{-1}$ Gyr. This similarity suggests that the Hubble expansion may be directly linked to the evolution of the Universe.

The apparently isotropic **distribution of galaxies (2)**, when averaged over large scales, and the **CMB isotropy (6)** suggest that the Universe around us is isotropic on large angular scales. Therefore we will first consider a world model that describes the Universe around us as isotropic. If we assume, in addition, that our place in the cosmos is not privileged over any other place, then the assumption of isotropy around us implies that the Universe appears isotropic as seen from any other place. The homogeneity of the Universe follows immediately from the isotropy around every location, as explained in Fig. 4.5. The combined assumption of homogeneity and isotropy of the Universe is also known as the *cosmological principle*. We will see that a world model based on the cosmological principle in fact provides an excellent description of numerous observational facts.

However, homogeneity is in principle unobservable because observations of distant objects show those at an earlier epoch. If the Universe evolves in time, as the

aforementioned observations suggest, evolutionary effects cannot directly be separated from spatial variations.

The assumption of homogeneity of course breaks down on small scales. We observe structures in the Universe, like galaxies and clusters of galaxies, and even accumulations of clusters of galaxies, so-called superclusters. Structures have been found in redshift surveys that extend over $\sim 100\,h^{-1}\,\mathrm{Mpc}$. However, we have no indication of the existence of structures in the Universe with scales $\gg 100\,h^{-1}\,\mathrm{Mpc}$. This length-scale can be compared to a characteristic length of the Universe, which is obtained from the Hubble constant. If H_0^{-1} specifies the characteristic age of our Universe, then light will travel a distance c/H_0 in this time. With this, we have obtained in problem 1.1 the *Hubble radius* as a characteristic length-scale of the Universe (or more precisely, of the observable Universe),

$$\boxed{R_\mathrm{H} := \frac{c}{H_0} = 2998\,h^{-1}\,\mathrm{Mpc} \;:\; \text{Hubble radius}} \,. \tag{4.3}$$

The Hubble volume $\sim R_\mathrm{H}^3$ can contain a very large number of structures of size $\sim 100\,h^{-1}\,\mathrm{Mpc}$, so that it still makes sense to assume an on-average homogeneous cosmological model. Superposed on this homogeneous universe we then have density fluctuations that are identified with the observed large-scale structures; these will be discussed in detail in Chap. 7. To a first approximation we can neglect these density perturbations in a description of the Universe as a whole. We will therefore consider world models that are based on the cosmological principle, i.e., in which the universe looks the same for all observers (or, in other words, if observed from any point).

Homogeneous and isotropic world models are the simplest cosmological solutions of the equations of General Relativity (GR). We will examine how far such simple models are compatible with observations. As we shall see, the application of the cosmological principle results in the observational facts which were mentioned in Sect. 4.1.1.

4.2 An expanding universe

Gravitation is the fundamental force in the Universe. Only gravitational forces and electromagnetic forces can act over large distance. Since cosmic matter is electrically neutral on average, electromagnetic forces do not play any significant role on large scales, so that gravity has to be considered as the driving force in cosmology. The laws of gravity are described by the theory of General Relativity, formulated by A. Einstein in 1915. It contains Newton's theory of gravitation as a special case for weak gravitational fields and small spatial scales. Newton's theory of gravitation has been proven to be eminently successful, e.g., in describing the motion of planets. Thus it is tempting to try to design a cosmological model based on Newtonian gravity. We will proceed to do that as a first step because not only is this Newtonian cosmology very useful from a didactic point of view, but one can also argue why the Newtonian cosmos correctly describes the major aspects of a relativistic cosmology.

4.2.1 Newtonian cosmology

The description of a gravitational system necessitates the application of GR if the length-scales in the system are comparable to the radius of curvature of spacetime; this is certainly the case in our Universe. Even if we cannot explain at this point what exactly the 'curvature radius of the Universe' is, it should be plausible that it is of the same order of magnitude as the Hubble radius R_H. We will discuss this more thoroughly further below. Despite this fact, one can expect that a Newtonian description is essentially correct: in a homogeneous universe, any small spatial region is characteristic for the whole universe. If the evolution of a small region in space is known, we also know the history of the whole universe, due to homogeneity. However, on small scales, the Newtonian approach is justified. We will therefore, based on the cosmological principle, first consider spatially homogeneous and isotropic world models in the framework of Newtonian gravity.

4.2.2 Kinematics of the Universe

Comoving coordinates. We consider a homogeneous sphere which may be radially expanding (or contracting); however, we require that the density $\rho(t)$ remains spatially homogeneous. The density may vary in time due to expansion or contraction. We choose a point $t = t_0$ in time and introduce a coordinate system $\boldsymbol{x}$ at this instant with the origin coinciding with the center of the sphere. A particle in the sphere which is located at position $\boldsymbol{x}$ at time t_0 will be located at some other time t at the position $\boldsymbol{r}(t)$ which results from the expansion of the sphere. Since the expansion is radial or, in other words, the velocity vector of a particle at position $\boldsymbol{r}(t)$ is parallel to $\boldsymbol{r}$, the direction of $\boldsymbol{r}(t)$ is constant. Because $\boldsymbol{r}(t_0) = \boldsymbol{x}$, this means that

$$\boldsymbol{r}(t) = a(t)\,\boldsymbol{x} \,. \tag{4.4}$$

Since $\boldsymbol{x}$ and $\boldsymbol{r}$ both have the dimension of a length, the function $a(t)$ is dimensionless; it can depend only on time. Although requiring radial expansion alone could make a depend on $|\boldsymbol{x}|$ as well, the requirement that the density remains homogeneous implies that a must be spatially constant. The function $a(t)$ is called the *cosmic scale factor*; due

to $\boldsymbol{r}(t_0) = \boldsymbol{x}$, it obeys

$$a(t_0) = 1 \ . \tag{4.5}$$

The value of t_0 is arbitrary; we choose $t_0 =$ today. Particles (or observers) which move according to (4.4) are called *comoving particles (observers)*, and $\boldsymbol{x}$ is the comoving coordinate. The world line $(\boldsymbol{r}, t)$ of a comoving observer is unambiguously determined by $\boldsymbol{x}$, $(\boldsymbol{r}, t) = [a(t)\boldsymbol{x}, t]$.

Expansion rate. The velocity of such a comoving particle is obtained from the time derivative of its position,

$$\boldsymbol{v}(\boldsymbol{r}, t) = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{r}(t) = \frac{\mathrm{d}a}{\mathrm{d}t}\,\boldsymbol{x} \equiv \dot{a}\,\boldsymbol{x} = \frac{\dot{a}}{a}\,\boldsymbol{r} \equiv H(t)\,\boldsymbol{r} \ , \tag{4.6}$$

where in the last step we defined the *expansion rate*

$$\boxed{H(t) := \frac{\dot{a}}{a}} \ . \tag{4.7}$$

The choice of this notation is not accidental, since H is closely related to the Hubble constant. To see this, we consider the relative velocity vector of two comoving particles at positions $\boldsymbol{r}$ and $\boldsymbol{r} + \Delta\boldsymbol{r}$, which follows directly from (4.6):

$$\Delta\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{r} + \Delta\boldsymbol{r}, t) - \boldsymbol{v}(\boldsymbol{r}, t) = H(t)\,\Delta\boldsymbol{r} \ . \tag{4.8}$$

Hence, the relative velocity is proportional to the separation vector, so that the relative velocity is purely radial. Furthermore, the constant of proportionality $H(t)$ depends only on time but not on the position of the two particles. Obviously, (4.8) is very similar to the Hubble law

$$v = H_0\,D \ , \tag{4.9}$$

in which v is the radial velocity of a source at distance D from us. Therefore, setting $t = t_0$ and $H_0 \equiv H(t_0)$, (4.8) is simply the Hubble law, in other words, (4.8) is a generalization of (4.9) for arbitrary time. It expresses the fact that any observer expanding with the sphere will observe an isotropic velocity field that follows the Hubble law. Since we are observing an expansion today—sources are moving away from us—we have $H_0 > 0$, and $\dot{a}(t_0) > 0$.

The kinematics of comoving observers in an expanding universe is analogous to that of raisins in a yeast dough. Once in the oven, the dough expands, and accordingly the positions of the raisins change. All raisins move away from all other ones, and the mutual radial velocity is proportional to the separation between any pair of raisins—i.e., their motion follows the Hubble law (4.8), with an expansion rate $H(t)$ which depends on the quality of the yeast and the temperature of the oven. The spatial position of each raisin at the time the oven is started uniquely identifies a raisin, and can be taken as its comoving coordinate $\boldsymbol{x}$, measured relative to the center of the dough. The spatial position $\boldsymbol{r}(t)$ at some later time t is then given by (4.4), where $a(t)$ denotes the linear size of the dough at time t relative to the size when the oven was started.

4.2.3 Dynamics of the expansion

The above discussion describes the kinematics of the expansion. However, to obtain the behavior of the function $a(t)$ in time, and thus also the motion of comoving observers and the time evolution of the density of the sphere, it is necessary to consider the dynamics. The evolution of the expansion rate is determined by self-gravity of the sphere, from which it is expected that it will cause a deceleration of the expansion.

Equation of motion. We therefore consider a spherical surface of radius x at time t_0 and, accordingly, a radius $r(t) = a(t)\,x$ at arbitrary time t. The mass $M(x)$ enclosed in this comoving surface is constant in time, and is given by

$$\begin{aligned} M(x) &= \frac{4\pi}{3}\,\rho_0\,x^3 = \frac{4\pi}{3}\,\rho(t)\,r^3(t) \\ &= \frac{4\pi}{3}\,\rho(t)\,a^3(t)\,x^3 \ , \end{aligned} \tag{4.10}$$

where ρ_0 must be identified with the mass density of the universe today ($t = t_0$). The density is a function of time and, due to mass conservation, it is inversely proportional to the volume of the sphere,

$$\rho(t) = \rho_0\,a^{-3}(t) \ . \tag{4.11}$$

The gravitational acceleration of a particle on the spherical surface is $GM(x)/r^2$, directed towards the center. This then yields the *equation of motion* of the particle,

$$\ddot{r}(t) \equiv \frac{\mathrm{d}^2 r}{\mathrm{d}t^2} = -\frac{G\,M(x)}{r^2} = -\frac{4\pi G}{3}\,\frac{\rho_0\,x^3}{r^2} \ , \tag{4.12}$$

or, after substituting $r(t) = x\,a(t)$, an equation for a,

$$\ddot{a}(t) = \frac{\ddot{r}(t)}{x} = -\frac{4\pi G}{3}\,\frac{\rho_0}{a^2(t)} = -\frac{4\pi G}{3}\,\rho(t)\,a(t) \ . \tag{4.13}$$

It is important to note that this equation of motion does not dependent on x. The dynamics of the expansion, described by $a(t)$, is determined solely by the matter density.

'Conservation of energy'. Another way to describe the dynamics of the expanding shell is based on the law of energy conservation: the sum of kinetic and potential energy is constant in time. This conservation of energy is derived directly from (4.13). To do this, (4.13) is multiplied by $2\dot{a}$, and the resulting equation can be integrated with respect to time since $\mathrm{d}(\dot{a}^2)/\mathrm{d}t = 2\dot{a}\ddot{a}$, and $\mathrm{d}(-1/a)/\mathrm{d}t = \dot{a}/a^2$:

$$\dot{a}^2 = \frac{8\pi G}{3}\rho_0 \frac{1}{a} - Kc^2 = \frac{8\pi G}{3}\rho(t)\, a^2(t) - Kc^2 \; ; \quad (4.14)$$

here, Kc^2 is a constant of integration that will be interpreted later. After multiplication with $x^2/2$, (4.14) can be written as

$$\frac{v^2(t)}{2} - \frac{G\,M}{r(t)} = -Kc^2\,\frac{x^2}{2} \; ,$$

which is interpreted such that the kinetic + potential energy (per unit mass) of a particle is a constant on the spherical surface. Thus (4.14) in fact describes the conservation of energy. The latter equation also immediately suggests an interpretation of the integration constant: K is proportional to the total energy of a comoving particle, and thus the history of the expansion depends on K. The sign of K characterizes the qualitative behavior of the cosmic expansion history.

- If $K < 0$, the right-hand side of (4.14) is always positive. Since $\mathrm{d}a/\mathrm{d}t > 0$ today, $\mathrm{d}a/\mathrm{d}t$ remains positive for all times or, in other words, the universe will expand forever.
- If $K = 0$, the right-hand side of (4.14) is always positive, i.e., $\mathrm{d}a/\mathrm{d}t > 0$ for all times, and the universe will also expand forever, but in a way that $\mathrm{d}a/\mathrm{d}t \to 0$ for $t \to \infty$—the asymptotic expansion velocity for $t \to \infty$ is zero.
- If $K > 0$, the right-hand side of (4.14) vanishes if $a = a_{\max} = (8\pi G\rho_0)/(3Kc^2)$. For this value of a, $\mathrm{d}a/\mathrm{d}t = 0$, and the expansion will come to a halt. After that, the expansion will turn into a contraction, and such a universe will re-collapse.

In the special case of $K = 0$, which separates eternally expanding world models from those that will re-collapse in the future, the universe has a current density called *critical density* which can be inferred from (4.14) by setting $t = t_0$ and $H_0 = \dot{a}(t_0)$:

$$\rho_{\mathrm{cr}} := \frac{3H_0^2}{8\pi G} = 1.88 \times 10^{-29}\, h^2\, \mathrm{g/cm^3} \; . \quad (4.15)$$

Obviously, ρ_{cr} is a characteristic density of the current universe. As in many situations in physics, it is useful to express physical quantities in terms of dimensionless parameters, for instance the current cosmological density. We therefore define the *density parameter*

$$\boxed{\Omega_0 := \frac{\rho_0}{\rho_{\mathrm{cr}}}} \; , \quad (4.16)$$

where $K > 0$ corresponds to $\Omega_0 > 1$, and $K < 0$ corresponds to $\Omega_0 < 1$. Thus, Ω_0 is one of the central cosmological parameters. Its accurate determination was possible only quite recently, and we shall discuss this in detail later. However, we should mention here that matter which is visible as stars contributes only a small fraction to the density of our Universe, $\Omega_* \lesssim 0.01$. But, as we already discussed in the context of rotation curves of spiral galaxies and the mass determination of elliptical galaxies from the gravitational lensing effect, we find clear indications of the presence of dark matter which can in principle dominate the value of Ω_0. We will see that this is indeed the case.

4.2.4 Modifications due to General Relativity

The Newtonian approach contains nearly all essential aspects of homogeneous and isotropic world models, otherwise we would not have discussed it in detail. Most of the above equations are also valid in relativistic cosmology, although the interpretation needs to be altered. In particular, the image of an expanding sphere needs to be revised—this picture implies that a 'center' of the universe exists. Such a picture implicitly contradicts the cosmological principle in which no point is singled out over others—our Universe neither has a center, nor is it expanding away from a privileged point. However, the image of a sphere does not show up in any of the relevant equations: (4.11) for the evolution of the cosmological density and (4.13) and (4.14) for the evolution of the scale factor $a(t)$ contain no quantities that refer to a sphere.

General Relativity modifies the Newtonian model in several respects:

- We know from the theory of Special Relativity that mass and energy are equivalent, according to Einstein's famous relation $E = m\,c^2$. This implies that it is not only the matter density that contributes to the equations of motion. For example, a radiation field like the CMB has an energy density and, due to the equivalence above, this has to enter the expansion equations. We will see below that such a radiation field can be characterized as matter with pressure. The pressure will then explicitly appear in the equation of motion for $a(t)$.
- The field equation of GR as originally formulated by Einstein did not permit a solution which corresponds to a homogeneous, isotropic, and static cosmos. But since Einstein, like most of his contemporaries, believed the Universe to be static, he modified his field equations by introducing an additional term, the cosmological constant.
- The interpretation of the expansion is changed completely: it is not the particles or the observers that are expanding away from each other, nor is the Universe an expanding sphere. Instead, it is space itself that expands. In particular, the redshift is no Doppler redshift, but is itself a property of expanding spacetimes. However, we may still visualize redshift locally as being due to the Doppler effect without making a substantial conceptual error.[2]

[2] Returning to the picture of the raisins in a yeast dough above: On the one hand, the raisins have a velocity relative to each other, and thus

In the following, we will explain the first two aspects in more detail.

First law of thermodynamics. When air is compressed, for instance when pumping up a tire, it heats up. The temperature increases and accordingly so does the thermal energy of the air. In the language of thermodynamics, this fact is described by the first law: the change in internal energy $\mathrm{d}U$ through an (adiabatic) change in volume $\mathrm{d}V$ equals the work $\mathrm{d}U = -P\,\mathrm{d}V$, where P is the pressure in the gas. From the equations of GR as applied to a homogeneous isotropic cosmos, a relation is derived which reads

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(c^2\,\rho\,a^3\right) = -P\,\frac{\mathrm{d}a^3}{\mathrm{d}t}\,, \tag{4.17}$$

in full analogy to this law. Here, $\rho\,c^2$ is the energy density, i.e., for 'normal' matter, ρ is the mass density, and P is the pressure of the matter. If we now consider a constant comoving volume element V_x, then its physical volume $V = a^3(t)V_x$ will change due to expansion. Thus, $a^3 = V/V_x$ is the volume, and $c^2\,\rho\,a^3$ the energy contained in the volume, each divided by V_x. Taken together, (4.17) corresponds to the first law of thermodynamics in an expanding universe.

The Friedmann–Lemaître expansion equations. Next, we will present equations for the scale factor $a(t)$ which follow from GR for a homogeneous isotropic universe. Afterwards, we will derive these equations from the relations stated above—as we shall see, the modifications by GR are in fact only minor, as expected from the argument that a small section of a homogeneous universe characterizes the cosmos as a whole. The field equations of GR yield the equations of motion

$$\boxed{\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{Kc^2}{a^2} + \frac{\Lambda}{3}} \tag{4.18}$$

and

$$\boxed{\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) + \frac{\Lambda}{3}}\,, \tag{4.19}$$

where Λ is the aforementioned cosmological constant introduced by Einstein.[3] Compared to (4.13) and (4.14), these two equations have been changed in two places. First, the cosmological constant occurs in both equations, and second, the equation of motion (4.19) now contains a pressure term. The pair of (4.18) and (4.19) are called the *Friedmann equations.*

The cosmological constant. When Einstein introduced the Λ-term into his equations, he did this solely for the purpose of obtaining a static solution for the resulting expansion equations. We can easily see that (4.18) and (4.19), without the Λ-term, have no solution for $\dot{a} \equiv 0$. However, if the Λ-term is included, such a solution can be found (which is irrelevant, however, as we now know that the Universe is expanding). Einstein had no real physical interpretation for this constant, and after the expansion of the Universe was discovered he discarded it again. But with the genie out of the bottle, the cosmological constant remained in the minds of cosmologists, and their attitude towards Λ has changed frequently in the past 90 years. Around the turn of the millennium, observations were made which strongly suggest a non-vanishing cosmological constant, and the evidence has been further strengthened since, as will be detailed in Chap. 8. Today we know that our Universe has a non-zero cosmological constant, or at least something very similar to it.

But the physical interpretation of the cosmological constant has also been modified. In quantum mechanics even completely empty space, the so-called vacuum, may have a finite energy density, the vacuum energy density. For physical measurements not involving gravity, the value of this vacuum energy density is of no relevance since those measurements are only sensitive to energy *differences*. For example, the energy of a photon that is emitted in an atomic transition equals the energy difference between the two corresponding states in the atom. Thus the absolute energy of a state is measurable only up to a constant. Only in gravity does the absolute energy become important, because $E = m\,c^2$ implies that it corresponds to a mass.

It is now found that the cosmological constant is equivalent to a finite vacuum energy density—the equations of GR, and thus also the expansion equations, are not affected by this new interpretation. We will explain this fact in the following.

4.2.5 The components of matter in the Universe

Starting from the equation of energy conservation (4.14), we will now derive the relativistically correct expansion equations (4.18) and (4.19). The only change with respect to the Newtonian approach in Sect. 4.2.3 will be that we introduce other forms of matter. The essential components of our Universe can be described as pressure-free matter, radiation, and vacuum energy.

are 'moving'. On the other hand, they are stuck in the dough, and thus have no (peculiar) velocity—they are comoving with the dough. Their mutual relative velocity thus results solely from the expansion of the dough, which can be considered an analog to the expanding spacetime.

[3] In the original notation, the Λ used here is denoted by Λc^2; for notational simplicity, we absorb the c^2 into the definition of Λ.

Pressure-free matter. The pressure in a gas is determined by the thermal motion of its constituents. At room temperature, molecules in the air move at a speed comparable to the speed of sound, $c_s \sim 300\,\mathrm{m/s}$. For such a gas, $P \sim \rho c_s^2 \ll \rho c^2$, so that its pressure is of course gravitationally completely insignificant. In cosmology, a substance with $P \ll \rho c^2$ is denoted as (pressure-free) matter, also called cosmological dust.[4] We approximate $P_m = 0$, where the index 'm' stands for matter. The constituents of the (pressure-free) matter move with velocities much smaller than c.

Radiation. If this condition is no longer satisfied, thus if the thermal velocities are no longer negligible compared to the speed of light, then the pressure will also no longer be small compared to ρc^2. In the limiting case that the thermal velocity equals the speed of light, we denote this component as 'radiation'. One example of course is electromagnetic radiation, in particular the CMB photons. Another example would be other particles of vanishing rest mass. Even particles of finite mass can have a thermal velocity very close to c if the thermal energy of the particles is much larger than the rest mass energy, i.e., $k_B T \gg mc^2$. In these cases, the pressure is related to the density via the equation of state for radiation,

$$P_r = \frac{1}{3}\rho_r c^2 \ . \tag{4.20}$$

The pressure of radiation. Pressure is defined as the momentum transfer onto a perfectly reflecting wall per unit time and per unit area. Consider an isotropic distribution of photons (or another kind of particle) moving with the speed of light. The momentum of a photon is given in terms of its energy as $p = E/c = h_P \nu/c$, where h_P is the Planck constant. Consider now an area element $\mathrm{d}A$ of the wall; the momentum transferred to it per unit time is given by the momentum transfer per photon, times the number of photons hitting the area $\mathrm{d}A$ per unit time. We will assume for the moment that all photons have the same frequency. If θ denotes the direction of a photon relative to the normal of the wall, the momentum component perpendicular to the wall before scattering is $p_\perp = p\cos\theta$, and after scattering $p_\perp = -p\cos\theta$; the two other momentum components are unchanged by the reflection. Thus, the momentum transfer per photon scattering is $\Delta p = 2p\cos\theta$. The number of photons scattering per unit time within the area $\mathrm{d}A$ is given by the number density of photons, n_γ times the area element $\mathrm{d}A$, times the thickness of the layer from which photons arrive at the wall per unit time. The latter is given by $c\cos\theta$, since only the perpendicular velocity component brings them closer to the wall. Putting these terms together, we find for the momentum transfer to the wall per unit time per unit area the expression

$$P_r(\theta) = 2\,\frac{h_P \nu}{c}\,\cos\theta\, n_\gamma\, c\,\cos\theta \ .$$

Averaging this expression over a half-sphere (only photons moving towards the wall can hit it) then yields

$$P_r = \frac{1}{3} h_P \nu n_\gamma = \frac{1}{3} u_\gamma \ ,$$

where $u_\gamma = \rho_r c^2$ is the energy density of the photons. Since this final expression does not depend on the photon frequency, the assumption of a mono-chromatic distribution is not important, and the result applies to any frequency distribution.

Vacuum energy. The equation of state for vacuum energy takes a very unusual form which results from the first law of thermodynamics. Because the energy density ρ_v of the vacuum is constant in space and time, (4.17) immediately yields the relation

$$P_v = -\rho_v c^2 \ . \tag{4.21}$$

Thus the vacuum energy has a negative pressure. This unusual form of an equation of state can also be made plausible as follows: consider the change of a volume V that contains only vacuum. Since the internal energy is $U \propto V$, and thus a growth by $\mathrm{d}V$ implies an increase in U, the first law $\mathrm{d}U = -P\,\mathrm{d}V$ demands that P be negative.

4.2.6 "Derivation" of the expansion equation

Beginning with the equation of energy conservation (4.14), we are now able to derive the expansion equations (4.18) and (4.19). To achieve this, we differentiate both sides of (4.14) with respect to t and obtain

$$2\,\dot{a}\,\ddot{a} = \frac{8\pi G}{3}\left(\dot{\rho}\,a^2 + 2\,a\,\dot{a}\,\rho\right) \ .$$

Next, we carry out the differentiation in (4.17), thereby obtaining $\dot{\rho}a^3 + 3\rho a^2\dot{a} = -3Pa^2\dot{a}/c^2$. This relation is then used to replace the term containing $\dot{\rho}$ in the previous equation, yielding

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right) \ . \tag{4.22}$$

This derivation therefore reveals that the pressure term in the equation of motion results from the combination of energy conservation and the first law of thermodynamics. However, we point out that the first law in the form (4.17) is based explicitly on the equivalence of mass and energy, resulting from Special Relativity. When assuming this equivalence, we indeed obtain the Friedmann equations from Newtonian cosmology, as expected from the discussion at the beginning of Sect. 4.2.1.

Next we consider the three aforementioned components of the cosmos and write the density and pressure as the sum of dust, radiation, and vacuum energy,

$$\rho = \rho_m + \rho_r + \rho_v = \rho_{m+r} + \rho_v \ , \quad P = P_r + P_v \ ,$$

[4] The notation 'dust' should not be confused with the dust that is responsible for the extinction and reddening of light—'dust' in cosmology only denotes matter with $P = 0$.

where $\rho_{\mathrm{m+r}}$ combines the density in matter and radiation. In the second equation, the pressureless nature of matter, $P_{\mathrm{m}} = 0$, was used so that $P_{\mathrm{m+r}} = P_{\mathrm{r}}$. By inserting the first of these equations into (4.14), we indeed obtain the first Friedmann equation (4.18) if the density ρ there is identified with $\rho_{\mathrm{m+r}}$ (the density in 'normal matter'), and if

$$\rho_{\mathrm{v}} = \frac{\Lambda}{8\pi G} \,. \tag{4.23}$$

Furthermore, we insert the above decomposition of density and pressure into the equation of motion (4.22) and immediately obtain (4.19) if we identify ρ and P with $\rho_{\mathrm{m+r}}$ and $P_{\mathrm{m+r}} = P_{\mathrm{r}}$, respectively. Hence, this approach yields both Friedmann equations; the density and the pressure in the Friedmann equations refer to normal matter, i.e., all matter except the contribution by Λ. Alternatively, the Λ-terms in the Friedmann equations may be discarded if instead the vacuum energy density and its pressure are explicitly included in P and ρ.

4.2.7 Discussion of the expansion equations

Following the 'derivation' of the expansion equations, we will now discuss their consequences. First we consider the density evolution of the various cosmic components resulting from (4.17). For pressure-free matter, we immediately obtain $\rho_{\mathrm{m}} \propto a^{-3}$ which is in agreement with (4.11). Inserting the equation of state (4.20) for radiation into (4.17) yields the behavior $\rho_{\mathrm{r}} \propto a^{-4}$; the vacuum energy density is a constant in time. Hence

$$\begin{aligned} &\rho_{\mathrm{m}}(t) = \rho_{\mathrm{m},0}\, a^{-3}(t) \;;\quad \rho_{\mathrm{r}}(t) = \rho_{\mathrm{r},0}\, a^{-4}(t) \;; \\ &\rho_{\mathrm{v}}(t) = \rho_{\mathrm{v}} = \text{const.} \,, \end{aligned} \tag{4.24}$$

where the index '0' indicates the current time, $t = t_0$. The physical origin of the a^{-4} dependence of the radiation density is seen as follows: as for matter, the number density of photons changes $\propto a^{-3}$ because the number of photons in a comoving volume is unchanged. However, photons are redshifted by the cosmic expansion. Their wavelength λ changes proportional to a (see Sect. 4.3.2). Since the energy of a photon is $E = h_{\mathrm{P}}\,\nu$ and $\nu = c/\lambda$, the energy of a photon changes as a^{-1} due to cosmic expansion so that the photon energy density changes $\propto a^{-4}$.

Analogous to (4.16), we define the dimensionless density parameters for matter, radiation, and vacuum,

$$\boxed{\Omega_{\mathrm{m}} = \frac{\rho_{\mathrm{m},0}}{\rho_{\mathrm{cr}}} \;;\quad \Omega_{\mathrm{r}} = \frac{\rho_{\mathrm{r},0}}{\rho_{\mathrm{cr}}} \;;\quad \Omega_{\Lambda} = \frac{\rho_{\mathrm{v}}}{\rho_{\mathrm{cr}}} = \frac{\Lambda}{3H_0^2}} \,, \tag{4.25}$$

so that $\Omega_0 = \Omega_{\mathrm{m}} + \Omega_{\mathrm{r}} + \Omega_{\Lambda}$.[5]

By now we know the current composition of our Universe quite well. The matter density of galaxies (including their dark halos) corresponds to $\Omega_{\mathrm{m}} \gtrsim 0.02$, depending on the—largely unknown—extent of their dark halos. This value therefore provides a lower limit for Ω_{m}. Studies of galaxy clusters, which will be discussed in Chap. 6, yield a lower limit of $\Omega_{\mathrm{m}} \gtrsim 0.1$. Finally, we will show in Chap. 8 that $\Omega_{\mathrm{m}} \sim 0.3$.

In comparison to matter, the radiation energy density today is much smaller. The energy density of the photons in the Universe is dominated by that of the cosmic background radiation. This is even more so the case in the early Universe before the first stars have produced additional radiation. Since the CMB has a Planck spectrum of temperature 2.73 K, we know its energy density from the Stefan–Boltzmann law,

$$\begin{aligned} \boxed{\rho_{\mathrm{CMB}} = a_{\mathrm{SB}}\, T^4 \equiv \left(\frac{\pi^2\, k_{\mathrm{B}}^4}{15\hbar^3\, c^3}\right) T^4} \\ \simeq 4.5 \times 10^{-34} \left(\frac{T}{2.73\,\mathrm{K}}\right)^4 \frac{\mathrm{g}}{\mathrm{cm}^3} \,, \end{aligned} \tag{4.26}$$

where in the final step we inserted the CMB temperature; here, $\hbar = h_{\mathrm{P}}/(2\pi)$ is the reduced Planck constant. This energy density corresponds to a density parameter of

$$\Omega_{\mathrm{CMB}} \simeq 2.4 \times 10^{-5} h^{-2} \,. \tag{4.27}$$

As will be explained below, the photons are not the only contributors to the radiation energy density. In addition, there are neutrinos from the early cosmic epoch which add to the density parameter of radiation, which then becomes

$$\Omega_{\mathrm{r}} \simeq 1.68\,\Omega_{\mathrm{CMB}} \sim 4.2 \times 10^{-5} h^{-2} \,, \tag{4.28}$$

so that today, the energy density of radiation in the Universe can be neglected when compared to that of matter. However, (4.24) reveal that the ratio between matter and radiation density was different at earlier epochs since ρ_{r} evolves faster with a than ρ_{m},

$$\frac{\rho_{\mathrm{r}}(t)}{\rho_{\mathrm{m}}(t)} = \frac{\rho_{\mathrm{r},0}}{\rho_{\mathrm{m},0}}\,\frac{1}{a(t)} = \frac{\Omega_{\mathrm{r}}}{\Omega_{\mathrm{m}}}\,\frac{1}{a(t)} \,. \tag{4.29}$$

Thus radiation and dust had the same energy density at an epoch when the scale factor was

[5] In the literature, different definitions for Ω_0 are used. Often the notation Ω_0 is used for Ω_{m}.

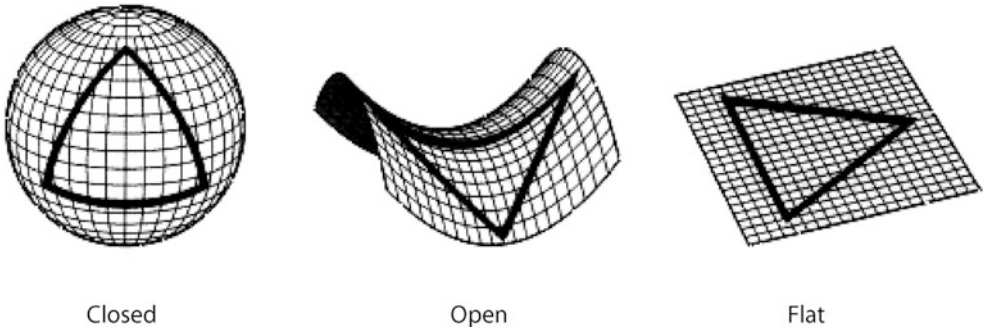


Fig. 4.6 Two-dimensional analogies for the three possible curvatures of space. In a universe with positive curvature ($K > 0$) the sum of the angles in a triangle is larger than 180°, in a universe of negative curvature it is smaller than 180°, and in a flat universe the sum of angles is exactly 180°. Adopted from J.A. Peacock 1999, *Cosmological Physics*, Cambridge University Press

$$\boxed{a_{\rm eq} = \frac{\Omega_{\rm r}}{\Omega_{\rm m}} = 4.2 \times 10^{-5} \left(\Omega_{\rm m} h^2\right)^{-1}} \,. \tag{4.30}$$

This value of the scale factor and the corresponding epoch in cosmic history play a very important role in structure evolution in the Universe, as we will see in Chap. 7.

With $\rho = \rho_{\rm m+r} = \rho_{\rm m,0}\, a^{-3} + \rho_{\rm r,0}\, a^{-4}$ and (4.25), the expansion equation (4.18) can be written as

$$H^2(t) = H_0^2 \left[\frac{\Omega_{\rm r}}{a^4(t)} + \frac{\Omega_{\rm m}}{a^3(t)} - \frac{Kc^2}{H_0^2\, a^2(t)} + \Omega_\Lambda \right] . \tag{4.31}$$

Evaluating this equation at the present epoch, with $H(t_0) = H_0$ and $a(t_0) = 1$, yields the value of the integration constant K,

$$\boxed{\begin{aligned} K &= \left(\frac{H_0}{c}\right)^2 (\Omega_0 - 1) = \left(\frac{H_0}{c}\right)^2 (\Omega_{\rm m} + \Omega_\Lambda + \Omega_{\rm r} - 1) \\ &\approx \left(\frac{H_0}{c}\right)^2 (\Omega_{\rm m} + \Omega_\Lambda - 1) \,. \end{aligned}} \tag{4.32}$$

Hence the constant K is obtained from the density parameters, mainly those of matter and vacuum since $\Omega_{\rm r} \ll \Omega_{\rm m}$, and has the dimension of (length)$^{-2}$. In the context of GR, K is interpreted as the curvature scalar of the universe today, or more precisely, the homogeneous, isotropic three-dimensional space at time $t = t_0$ has a curvature K. Depending on the sign of K, we can distinguish the following cases:

- If $K = 0$, the three-dimensional space for any fixed time t is Euclidean, i.e., flat.
- If $K > 0$, $1/\sqrt{K}$ can be interpreted as the curvature radius of the spherical 3-space—the two-dimensional analogy would be the surface of a sphere. As already speculated in Sect. 4.2.1, the order of magnitude of the curvature radius is c/H_0 according to (4.32).
- If $K < 0$, the space is called hyperbolic—the two-dimensional analogy would be the surface of a saddle (see Fig. 4.6).

Hence GR provides a relation between the curvature of space and the density of the universe. In fact, this is the central aspect of GR which links the geometry of spacetime to its matter content. However, Einstein's theory makes no statement about the topology of spacetime and, in particular, says nothing about the topology of the universe.[6] If the universe has a simple topology, it is finite in the case of $K > 0$, whereas it is infinite if $K \leq 0$. However, in both cases it has no boundary (compare: the surface of a sphere is a finite space without boundaries).

With (4.31) and (4.32), we finally obtain the expansion equation in the form

$$\boxed{\begin{aligned} \left(\frac{\dot a}{a}\right)^2 &= H^2(t) \\ &= H_0^2 \left[\frac{\Omega_{\rm r}}{a^4(t)} + \frac{\Omega_{\rm m}}{a^3(t)} + \frac{(1-\Omega_{\rm m}-\Omega_\Lambda)}{a^2(t)} + \Omega_\Lambda \right] , \\ &\equiv H_0^2 E^2(t) \end{aligned}} \tag{4.33}$$

where in the final step we defined the dimensionless Hubble function $E(t) = H(t)/H_0$ for later purposes.

4.3 Consequences of the Friedmann expansion

The cosmic expansion equations imply a number of immediate consequences, some of which will be discussed next. In particular, we will first demonstrate that the early Universe

[6] The surface of a cylinder is also considered a flat space, like a plane, because the sum of angles in a triangle on a cylinder is also 180°. But the surface of a cylinder obviously has a topology different from a plane; in particular, closed straight lines do exist—walking on a cylinder in a direction perpendicular to its axis, one will return to the starting point after a finite amount of time.

must have evolved out of a very dense and hot state called the *Big Bang*. We will then link the scale factor a to an observable, the redshift, and explain what the term 'distance' means in cosmology.

4.3.1 The necessity of a Big Bang

The terms on the right-hand side of (4.33) each have a different dependence on a:

- For very small a, the first term dominates and the universe is dominated by radiation then.
- For slightly larger $a \gtrsim a_{\rm eq}$, the pressureless matter (dust) term dominates.
- If $K \neq 0$, the third term, also called the curvature term, can dominate for larger a.
- For even larger a, the cosmological constant dominates if it is different from zero.

The differential equation (4.33) in general cannot be solved analytically. However, its numerical solution for $a(t)$ poses no problems. Nevertheless, we can analyze the qualitative behavior of the function $a(t)$ and thereby understand the essential aspects of the expansion history. From the Hubble law, we conclude that $\dot{a}(t_0) > 0$, i.e., a is currently an increasing function of time. Equation (4.33) shows that $\dot{a}(t) > 0$ for all times, unless the right-hand side of (4.33) vanishes for some value of a: the sign of $\dot{a}$ can only switch when the right-hand side of (4.33) is zero. If $H^2 = 0$ for a value of $a > 1$, the expansion will come to a halt and the Universe will recollapse afterwards. On the other hand, if $H^2 = 0$ for a value $a = a_{\rm min}$ with $0 < a_{\rm min} < 1$, then the sign of $\dot{a}$ switches at $a_{\rm min}$. At this epoch, a collapsing Universe changes into an expanding one.

Classification of model. Which of these alternatives describes our Universe depends on the density parameters. We find the following classification (also see Fig. 4.7 and problem 4.3):

- If $\Lambda = 0$, then $H^2 > 0$ for all $a \leq 1$, whereas the behavior for $a > 1$ depends on $\Omega_{\rm m}$:
  - if $\Omega_{\rm m} \leq 1$ (or $K \leq 0$, respectively), $H^2 > 0$ for all a: the universe will expand for all times. This behavior is expected from the Newtonian approach because if $K \leq 0$, the kinetic energy in any spherical shell is larger than the modulus of the potential energy, i.e., the expansion velocity exceeds the escape velocity and the expansion will never come to a halt.
  - If $\Omega_{\rm m} > 1$ ($K > 0$), H^2 will vanish for $a = a_{\rm max} = \Omega_{\rm m}/(\Omega_{\rm m} - 1)$. The universe will have its maximum expansion when the scale factor is $a_{\rm max}$ and will recollapse thereafter. In Newtonian terms, the total energy of any spherical shell is negative, so that it is gravitationally bound.

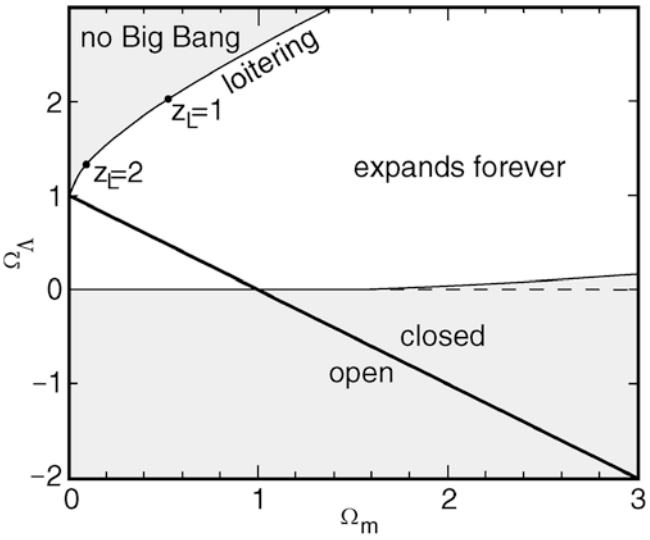


Fig. 4.7 Classification of cosmological models. The *straight solid line* connects flat models (i.e., those without spatial curvature, $\Omega_{\rm m} + \Omega_\Lambda = 1$) and separates open ($K < 0$) and closed ($K > 0$) models. The nearly horizontal curve separates models that will expand forever from those that will recollapse in the distant future. Models in the upper left corner have an expansion history where a has never been close to zero and thus did not experience a Big Bang. In those models, a maximum redshift for sources exists, which is indicated for two cases. Since we know that $\Omega_{\rm m} > 0.1$, and sources at redshift > 6 have been observed, these models can be excluded. Adopted from J.A. Peacock 1999, *Cosmological Physics*, Cambridge University Press

  We thus have reobtained the classification discussed before in Sect. 4.2.3, which is valid for $\Lambda = 0$, for which the qualitative behavior of the expansion depends only on the sign of K.
- In the presence of a cosmological constant $\Lambda > 0$, the discussion becomes more complicated; in particular, the geometry of the universe, i.e., the sign of K, is not sufficient to predict the qualitative expansion behavior. For example, there are models with positive curvature (indicated as 'closed' in Fig. 4.7) which expand forever. One finds for $\Lambda \geq 0$:
  - If $\Omega_{\rm m} < 1$, the universe will expand for all $a > 1$.
  - However, for $\Omega_{\rm m} > 1$ the future behavior of $a(t)$ depends on Ω_Λ: if Ω_Λ is sufficiently small, a value $a_{\rm max}$ exists at which the expansion comes to a halt and reverses. In contrast, if Ω_Λ is large enough the universe will expand forever.
  - If $\Omega_\Lambda < 1$, then $H^2 > 0$ for all $a \leq 1$.
  - However, if $\Omega_\Lambda > 1$, it is in principle possible that $H^2 = 0$ for an $a = a_{\rm min} < 1$. Such models, in which a minimum value for a existed in the past (so-called bouncing models), can be excluded by observations (see Sect. 4.3.2).

With the exception of the last case, which can be excluded, we come to the conclusion that a must have attained the

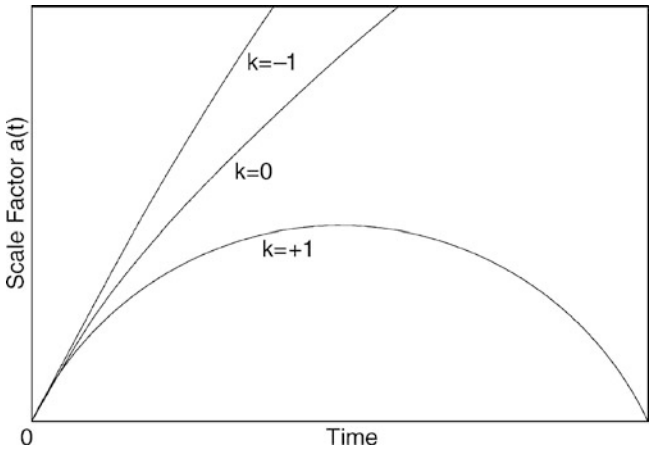


Fig. 4.8 The scale factor $a(t)$ as a function of cosmic time t for three models with a vanishing cosmological constant, $\Omega_\Lambda = 0$. Closed models ($K > 0$) attain a maximum expansion and then recollapse. In contrast, open models ($K \leq 0$) expand forever, and the Einstein–de Sitter model of $K = 0$ separates these two cases. In all models, the scale factor tends towards zero in the past; this time is called the Big Bang and defines the origin of the time axis

value $a = 0$ at some point in the past, at least formally. At this instant the 'size of the Universe' formally vanished. As $a \to 0$, both matter and radiation densities diverge so that the density in this state must have been singular. The epoch at which $a = 0$ and the evolution away from this state is called the *Big Bang*. It is useful to define this epoch ($a = 0$) as the origin of time, so that t is identified with the age of the Universe, the time since the Big Bang. As we will show, the predictions of the Big Bang model are in impressive agreement with observations.

The expansion history for the special case of a vanishing vacuum energy density is sketched in Fig. 4.8 for three values of the curvature.

To characterize whether the current expansion of the Universe is decelerated or accelerated, the *deceleration parameter*

$$q_0 := -\ddot{a}\, a/\dot{a}^2 \tag{4.34}$$

is defined where the right-hand side has to be evaluated at $t = t_0$. With (4.19) and (4.33) it follows that

$$q_0 = \Omega_\mathrm{m}/2 - \Omega_\Lambda \,. \tag{4.35}$$

If $\Omega_\Lambda = 0$ then $q_0 > 0$, $\ddot{a} < 0$, i.e., the expansion decelerates, as expected due to gravity. However, if Ω_Λ is sufficiently large the deceleration parameter may become negative, corresponding to an accelerated expansion of the universe. The reason for this behavior, which certainly contradicts intuition, is seen in the vacuum energy. Only a negative pressure can cause an accelerated expansion—more precisely, as seen from (4.22), $P < -\rho c^2/3$ is needed for $\ddot{a} > 0$. Indeed, we know today that our Universe is currently undergoing an accelerated expansion and thus that the cosmological constant differs significantly from zero.

Age of the universe. The age of the universe at a given scale factor a follows from $\mathrm{d}t = \mathrm{d}a(\mathrm{d}a/\mathrm{d}t)^{-1} = \mathrm{d}a/(aH)$. This relation can be integrated,

$$t(a) = \frac{1}{H_0}\int_0^a \mathrm{d}x\, \left[x^{-2}\Omega_\mathrm{r} + x^{-1}\Omega_\mathrm{m} + (1-\Omega_\mathrm{m}-\Omega_\Lambda) + x^2\Omega_\Lambda\right]^{-1/2} , \tag{4.36}$$

where the contribution from radiation for $a \gg a_\mathrm{eq}$ can be neglected because it is relevant only for very small a and thus only for a very small fraction of cosmic time. To obtain the current age t_0 of the universe, (4.36) is calculated for $a = 1$. For models of vanishing spatial curvature $K = 0$ and for those with $\Lambda = 0$, Fig. 4.9 displays t_0 as a function of Ω_m.

The qualitative behavior of the cosmological models is characterized by the density parameters Ω_m and Ω_Λ, whereas the Hubble constant H_0 determines 'only' the overall length- or time-scale. One can consider several special cases for the density parameters:

- Models without a cosmological constant, $\Lambda = 0$. The difficulties in deriving a 'sensible' value for Λ from particle physics has in the past often been used as an argument for neglecting the vacuum energy density. However, there are now very strong observational indications that in fact $\Lambda > 0$.
- Models with $\Omega_\mathrm{m} + \Omega_\Lambda = 1$, i.e., $K = 0$. Such flat models are preferred by the so-called inflationary models, which we will briefly discuss further below.
- A special case is the Einstein–de Sitter model, $\Omega_\mathrm{m} = 1$, $\Omega_\Lambda = 0$. For this model, $t_0 = 2/(3H_0) \approx 6.7\, h^{-1} \times 10^9$ yr.
- For many world models, t_0 is larger than the age of the oldest globular clusters, so they are compatible with this age determination. The Einstein–de Sitter model, however, is compatible with stellar ages only if H_0 is very small, considerably smaller than the value of H_0 derived from the HST Key Project discussed in Sect. 3.9. Hence, this model is ruled out by these observations.

The values of the cosmological parameters are now quite well known. We list them here for later reference without any further discussion. Their determination will be described in the course of this chapter and in Chap. 8. The values are approximately

$$\boxed{\Omega_\mathrm{m} \sim 0.3\,;\quad \Omega_\Lambda \sim 0.7\,;\quad h \sim 0.7}\,. \tag{4.37}$$

Early expansion. In the early phase of the universe, the curvature term and the vacuum energy density can be neglected in the expansion

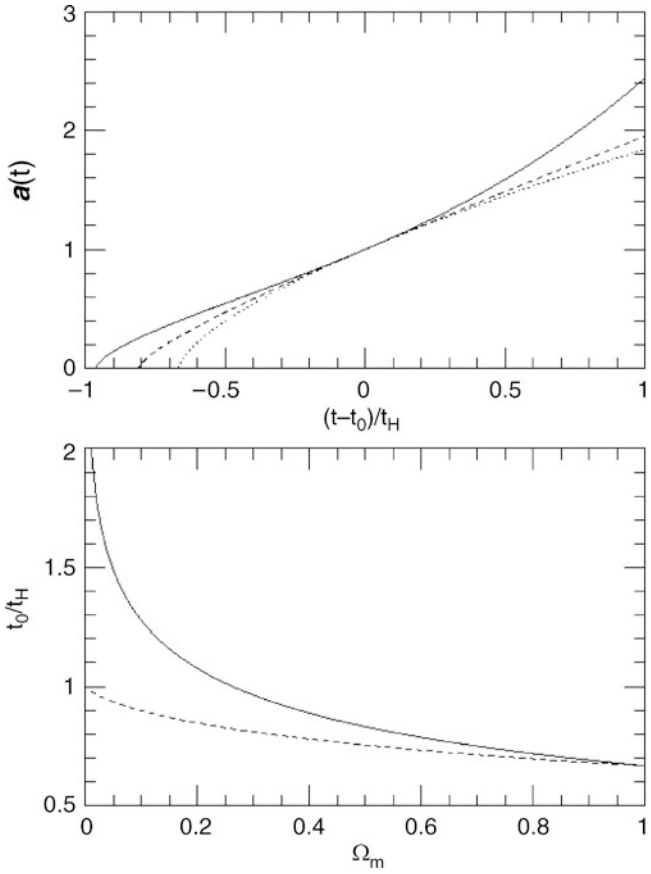


Fig. 4.9 *Top panel*: Scale factor $a(t)$ as a function of cosmic time, here scaled as $(t-t_0)H_0$, for an Einstein–de Sitter model ($\Omega_\mathrm{m} = 1$, $\Omega_\Lambda = 0$; *dotted curve*), an open universe ($\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0$; *dashed curve*), and a flat universe of low density ($\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0.7$; *solid curve*). At the current epoch, $t = t_0$ and $a = 1$. *Bottom panel*: Age of the universe in units of the Hubble time $t_\mathrm{H} = H_0^{-1}$ for flat world models with $K = 0$ ($\Omega_\mathrm{m} + \Omega_\Lambda = 1$; *solid curve*) and models with a vanishing cosmological constant (*dashed curve*). We see that for a flat universe with small Ω_m (thus large $\Omega_\Lambda = 1 - \Omega_\mathrm{m}$), t_0 may be considerably larger than H_0^{-1}. Credit: M. Bartelmann, MPA Garching

equation (4.33), which then simplifies to

$$H^2 = H_0^2 \left(\frac{\Omega_\mathrm{r}}{a^4} + \frac{\Omega_\mathrm{m}}{a^3} \right) = H_0^2 \, \Omega_\mathrm{m} \, a^{-3} \left(1 + \frac{a_\mathrm{eq}}{a} \right) ,$$

where we used (4.30). In this case, the relation (4.36) between time and scale factor can be integrated explicitly to yield

$$t = \frac{2}{3 H_0 \sqrt{\Omega_\mathrm{m}}} \left[a^{3/2} \left(1 - \frac{2 a_\mathrm{eq}}{a} \right) \sqrt{1 + \frac{a_\mathrm{eq}}{a}} + 2 a_\mathrm{eq}^{3/2} \right] . \qquad (4.38)$$

From this result we can infer that the scale factor behaves at $a \propto t^{1/2}$ for $a \ll a_\mathrm{eq}$ and that $a \propto t^{2/3}$ in the matter dominated era.

4.3.2 Redshift

The Hubble law describes a relation between the redshift, or the radial component of the relative velocity, and the distance of an object from us. Furthermore, (4.6) specifies that any observer is experiencing a local Hubble law with an expansion rate $H(t)$ which depends on the cosmic epoch. We will now derive a relation between the redshift of a source, which is directly observable, and the cosmic time t or the scale factor $a(t)$, respectively, at which the source emitted the light we receive today.

To do this, we consider a light ray that reaches us today. Along this light ray we imagine fictitious comoving observers. The light ray is parametrized by the cosmic time t, and is supposed to have been emitted by the source at epoch t_e. Two comoving observers along the light ray with separation $\mathrm{d}r$ from each other see their relative motion due to the cosmic expansion according to (4.6), $\mathrm{d}v = H(t)\,\mathrm{d}r$, and they measure it as a redshift of light, $\mathrm{d}\lambda/\lambda = \mathrm{d}z = \mathrm{d}v/c$. It takes a time $\mathrm{d}t = \mathrm{d}r/c$ for the light to travel from one observer to the other. Furthermore, from the definition of the Hubble parameter, $\dot{a} = \mathrm{d}a/\mathrm{d}t = H\,a$, we obtain the relation $\mathrm{d}t = \mathrm{d}a/(H\,a)$. Combining these relations, we find

$$\frac{\mathrm{d}\lambda}{\lambda} = \frac{\mathrm{d}v}{c} = \frac{H}{c}\,\mathrm{d}r = H\,\mathrm{d}t = \frac{\mathrm{d}a}{a} . \qquad (4.39)$$

The relation $\mathrm{d}\lambda/\lambda = \mathrm{d}a/a$ is now easily integrated since the equation $\mathrm{d}\lambda/\mathrm{d}a = \lambda/a$ obviously has the solution $\lambda = Ca$, where C is a constant. That constant is determined by the wavelength λ_obs of the light as observed today (i.e., at $a = 1$), so that

$$\lambda(a) = a\,\lambda_\mathrm{obs} . \qquad (4.40)$$

The wavelength at emission was therefore $\lambda_\mathrm{e} = a(t_\mathrm{e})\lambda_\mathrm{obs}$. On the other hand, the redshift z is defined as $(1 + z) = \lambda_\mathrm{obs}/\lambda_\mathrm{e}$. From this, we finally obtain the relation

$$\boxed{1 + z = \frac{1}{a}} \qquad (4.41)$$

between the observable z and the scale factor a which is linked via (4.36) to the cosmic time. The same relation can also be derived by considering light rays in GR.

The relation between redshift and the scale factor is of immense importance for cosmology because, for most sources, redshift is the only distance information that we are able to measure. If the scale factor is a monotonic function of time, i.e., if the right-hand side of (4.33) is different from zero for all $a \in [0, 1]$, then z is also a monotonic function of t. In this case, which corresponds to the Universe we happen to live in, a, t, and z are equally good measures of the distance of a source from us.

Local Hubble law. The Hubble law applies for nearby sources: with (4.8) and $v \approx zc$ it follows that

$$\boxed{z = \frac{H_0}{c}\,D \approx \frac{h\,D}{3000\,\mathrm{Mpc}} \quad \text{for } z \ll 1} , \qquad (4.42)$$

where D is the distance of a source with redshift z. This corresponds to a light travel time of $\Delta t = D/c$. On the other hand, due to the definition of the Hubble parameter, we have $\Delta a = (1-a) \approx H_0\,\Delta t$, where a is the scale factor at time $t_0 - \Delta t$, and we used $a(t_0) = 1$ and $H(t_0) = H_0$. This implies $D = (1-a)c/H_0$. Utilizing (4.42), we then find $z = 1-a$, or $a = 1-z$, which agrees with (4.41) in linear approximation since $(1+z)^{-1} = 1 - z + \mathcal{O}(z^2)$. Hence we conclude that the general relation (4.41) contains the local Hubble law as a special case.

Energy density in radiation. A further consequence of (4.41) is the dependence of the energy density of radiation on the scale parameter. As mentioned previously, the number density of photons is $\propto a^{-3}$ if we assume that photons are neither created nor destroyed. In other words, the number of photons in a comoving volume element is conserved. According to (4.41), the frequency ν of a photon changes due to cosmic expansion. Since the energy of a photon is $\propto \nu$, $E_\gamma = h_\mathrm{P}\nu \propto 1/a$, the energy density of photons decreases, $\rho_\mathrm{r} \propto n\,E_\gamma \propto a^{-4}$. Therefore (4.41) implies (4.24).

Cosmic microwave background. Assuming that, at some time t_1, the universe contained a blackbody radiation of temperature T_1, we can examine the evolution of this photon population in time by means of relation (4.41). We should recall that the Planck function B_ν (A.13) specifies the radiation energy of blackbody radiation that passes through a unit area per unit time, per unit frequency interval, and per unit solid angle. Using this definition, the number density $\mathrm{d}N_\nu$ of photons in the frequency interval between ν and $\nu + \mathrm{d}\nu$ is obtained as

$$\frac{\mathrm{d}N_\nu}{\mathrm{d}\nu} = \frac{4\pi\,B_\nu}{c\,h_\mathrm{P}\nu} = \frac{8\pi\nu^2}{c^3}\,\frac{1}{\exp\left(\frac{h_\mathrm{P}\nu}{k_\mathrm{B}T_1}\right)-1}\,. \tag{4.43}$$

At a later time $t_2 > t_1$, the universe has expanded by a factor $a(t_2)/a(t_1)$. An observer at t_2 therefore observes the photons redshifted by a factor $(1+z) = a(t_2)/a(t_1)$, i.e., a photon with frequency ν at t_1 will then be measured to have frequency $\nu' = \nu/(1+z)$. The original frequency interval is transformed accordingly as $\mathrm{d}\nu' = \mathrm{d}\nu/(1+z)$. The number density of photons decreases with the third power of the scale factor, so that $\mathrm{d}N'_{\nu'} = \mathrm{d}N_\nu/(1+z)^3$. Combining these relations, we obtain for the number density $\mathrm{d}N'_{\nu'}$ of photons in the frequency interval between ν' and $\nu' + \mathrm{d}\nu'$

$$\begin{aligned}\frac{\mathrm{d}N'_{\nu'}}{\mathrm{d}\nu'} &= \frac{\mathrm{d}N_\nu/(1+z)^3}{\mathrm{d}\nu/(1+z)} \\ &= \frac{1}{(1+z)^2}\,\frac{8\pi(1+z)^2\nu'^2}{c^3}\,\frac{1}{\exp\left(\frac{h_\mathrm{P}(1+z)\nu'}{k_\mathrm{B}T_1}\right)-1} \\ &= \frac{8\pi\nu'^2}{c^3}\,\frac{1}{\exp\left(\frac{h_\mathrm{P}\nu'}{k_\mathrm{B}T_2}\right)-1}\,,\end{aligned} \tag{4.44}$$

where we used $T_2 = T_1/(1+z)$ in the last step. The distribution (4.44) has the same form as (4.43) except that the temperature is reduced by a factor $(1+z)^{-1}$. If a Planck distribution of photons had been established at an earlier time, it will persist during cosmic expansion. As we have seen above, the CMB is such a blackbody radiation, with a current temperature of $T_0 = T_\mathrm{CMB} \approx 2.73\,\mathrm{K}$. We will show in Sect. 4.4 that this radiation originates in the early phase of the cosmos. Thus it is meaningful to consider the temperature of the CMB as the 'temperature of our Universe' which is a function of redshift,

$$\boxed{T(z) = T_0(1+z) = T_0\,a^{-1}}\,, \tag{4.45}$$

i.e., the Universe was hotter in the past than it is today. The energy density of the Planck spectrum is given by (4.26), i.e., proportional to T^4, so that ρ_r behaves like $(1+z)^4 = a^{-4}$ in accordance with (4.24).[7]

Finally, it should be stressed again that (4.41) allows all relations to be expressed as functions of a as well as of z. For example, the age of the Universe as a function of z is obtained by replacing the upper integration limit, $a \to (1+z)^{-1}$, in (4.36).

Interpretation of cosmological redshift. The redshift results from the fact that during the expansion of the universe, the energy of the photons decreases in proportion to $1/a$, which is the reason, together with the decreasing proper number density, that $\rho_\mathrm{r}(a) \propto a^{-4}$. Our considerations in this section have derived this $1/a$-dependence of the photon energy.

But maybe this is puzzling anyway—if photons lose energy during cosmic expansion, then, having in mind the concept of energy conservation, one might be tempted to ask: Where does this energy go to?

To answer this question, we start with pointing out that energy conservation in cosmology is expressed by the 'first law of thermodynamics' (4.17), which has as one of its consequences the $1/a$-behavior of photon energy. Thus, there is no reason to lose sleep about this issue.

But it may be useful to be more explicit here. We first point out that 'the energy' of a photon, or any other particle,

[7] Generally, it can be shown that the specific intensity I_ν changes due to redshift according to

$$\frac{I_\nu}{\nu^3} = \frac{I'_{\nu'}}{(\nu')^3}\,. \tag{4.46}$$

Here, I_ν is the specific intensity today at frequency ν and $I'_{\nu'}$ is the specific intensity at redshift z at frequency $\nu' = (1+z)\nu$.

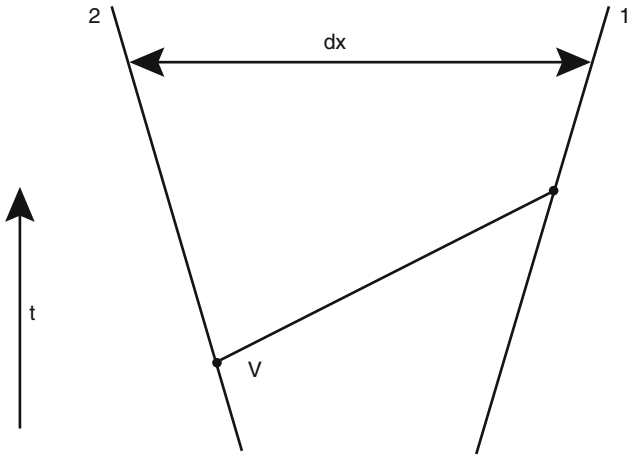


Fig. 4.10 Sketch of two comoving observers with comoving separation $\mathrm{d}x$, and a particle with velocity v, as measured by observer No. 2, moving from the worldline of observer No. 2 to that of observer No. 1

as such is not defined! To see this, consider two observers that measure the wavelength of photons; both are at the same location, but observer No. 2 moves in the direction of the light source, as seen from observer No. 1. Because of the Doppler effect, observer No. 2 measures a shorter wavelength than observer No. 1; thus, the two observers come to different conclusions about the energy of the photons. It's not like one of them is right, the other wrong—the energy of a photon is not an absolute quantity, but depends on the frame in which it is measured.

We have defined a reference frame in an expanding universe, that of comoving observers—they are the ones who see the universe being isotropic around them. When we state that 'the photon energy changes as $1/a$', we implicitly mean that it is the energy as measured by the comoving observer that changes. As these comoving observers move relative to each other, as expressed by the Hubble law, it should not be a surprise that they measure different frequencies of photons, as explicitly accounted for in (4.39). Thus, not the properties of the photons are changing in time, but the state of motion of the observers that measure the photon energy as they propagate through the universe.

In fact, one can show that in general the momentum (as measured by comoving observers!) of a freely moving particle changes as

$$p \propto \frac{1}{a} \,. \tag{4.47}$$

For photons, we have already shown this: since the momentum of a photon is given by its energy, divided by c, then $\nu \propto 1/a$ shows the validity of (4.47). The same is true for other relativistic particles as well. We can also derive this behavior easily for non-relativistic particles. Consider a particle of mass m which crosses the worldlines of the two neighboring observers 2 and 1 (see Fig. 4.10). Let $\mathrm{d}x$ be their comoving separation, then at epoch a their physical separation is $\mathrm{d}r = a\,\mathrm{d}x$. It takes the particle the time $\mathrm{d}t = \mathrm{d}r/v$ to travel between the two observers, where v is the velocity measured by observer No. 2. Observer No. 1 will measure a velocity $v - \mathrm{d}v$, since from his perspective, observer No. 2 is receding from him (this is simply a Galilean transformation), with a velocity $\mathrm{d}v = H(a)\,\mathrm{d}r$ given by the local Hubble law. Putting this together, there is a momentum change $\mathrm{d}p = -m\,\mathrm{d}v$ of the particle when measured by the two observers, where

$$\frac{\mathrm{d}p}{p} = -\frac{H(a)\,\mathrm{d}r}{v} = -H(a)\,\mathrm{d}t = -\frac{\mathrm{d}a/\mathrm{d}t}{a}\,\mathrm{d}t = -\frac{\mathrm{d}a}{a}\,, \tag{4.48}$$

where we made use of the definition of the Hubble function. The resulting equation $\mathrm{d}\ln p = -\mathrm{d}\ln a$ has the solution $pa = \text{const.}$, i.e., (4.47) holds. For semi-relativistic particles, the proof of (4.47) can be made with Special Relativity, but proceeds essentially in the same way.

The necessity of a Big Bang. We discussed in Sect. 4.3.1 that the scale factor must have attained the value $a = 0$ at some time in the past. One gap in our argument that inevitably led to the necessity of a Big Bang still remains, namely the possibility that at some time in the past $\dot{a} = 0$ occurred, i.e., that the universe underwent a transition from a contracting to an expanding state. This is possible only if $\Omega_\Lambda > 1$ and if the matter density parameter is sufficiently small (see Fig. 4.7). In this case, a attained a minimum value in the past. This minimum value depends on both Ω_m and Ω_Λ. For instance, for $\Omega_\mathrm{m} > 0.1$, the value is $a_\mathrm{min} \gtrsim 0.3$. But a minimum value for a implies a maximum redshift $z_\mathrm{max} = 1/a_\mathrm{min} - 1$. However, since we have observed quasars and galaxies with $z > 6$ and the density parameter is known to be $\Omega_\mathrm{m} > 0.1$, such a model without a Big Bang can be safely excluded.

4.3.3 Distances in cosmology

In the previous sections, different distance measures were discussed. Because of the monotonic behavior of the corresponding functions, each of a, t, and z provide the means to sort objects according to their distance. An object at higher redshift z_2 is more distant than one at $z_1 < z_2$ such that light from a source at z_2 may become absorbed by gas in an object at redshift z_1, but not vice versa. The object at redshift z_1 is located between us and the object at z_2. The more distant a source is from us, the longer the light takes to reach us, the earlier it was emitted, the smaller a was at emission, and the larger z is. Since z is the only observable of these parameters, distances in extragalactic astronomy are nearly always expressed in terms of redshift.

But how can a redshift be translated into a distance that has the dimension of a length? Or, phrasing this question differently, how many Megaparsecs away from us is a source with redshift $z = 2$? The corresponding answer is more complicated than the question suggests. For very small redshifts, the local Hubble relation (4.42) may be used, but this is valid only for $z \ll 1$.

In a static Euclidean space, the separation between two points is unambiguously defined, and several prescriptions exist for measuring a distance. We will give two examples here. A sphere of radius R situated at distance D subtends a solid angle of $\omega = \pi R^2/D^2$ on our sky. If the radius is known, D can be measured using this relation. As a second example, we consider a source of luminosity L at distance D which then has a measured flux $S = L/(4\pi D^2)$. Again, if the luminosity is known, the distance can be computed from the observed flux. If we use these two methods to determine, for example, the distance to the Sun, we would of course obtain identical results for the distance (within the range of accuracy), since these two prescriptions for distance measurements are defined to yield equal results.

In a non-Euclidean or expanding/contracting space-time like, for instance, our Universe this is no longer the case. The equivalence of different distance measures is only ensured in Euclidean space, and we have no reason to expect this equivalence to also hold in a curved spacetime. In cosmology, the same measuring prescriptions as in Euclidean space are used for defining distances, but the different definitions lead to different results. The two most important definitions of distance are:

- *Angular-diameter distance:* As above, we consider a source of radius R observed to cover a solid angle ω. The angular-diameter distance is defined as

$$\boxed{D_\mathrm{A}(z) = \sqrt{\frac{R^2\,\pi}{\omega}}}\,. \tag{4.49}$$

- *Luminosity distance:* We consider a source with bolometric luminosity L and flux S and define its luminosity distance as

$$\boxed{D_\mathrm{L}(z) = \sqrt{\frac{L}{4\pi\,S}}}\,. \tag{4.50}$$

These two distances agree locally, i.e., for $z \ll 1$; on small scales, the curvature of spacetime is not noticeable. In addition, they are *unique* functions of redshift. They can be computed explicitly. However, to do this some tools of GR are required. Since we have not discussed GR in this book, these tools are not available to us here. The distance-redshift relations depend on the cosmological parameters; Fig. 4.11 shows the angular-diameter distance for different models. For $\Lambda = 0$, the famous Mattig relation applies,

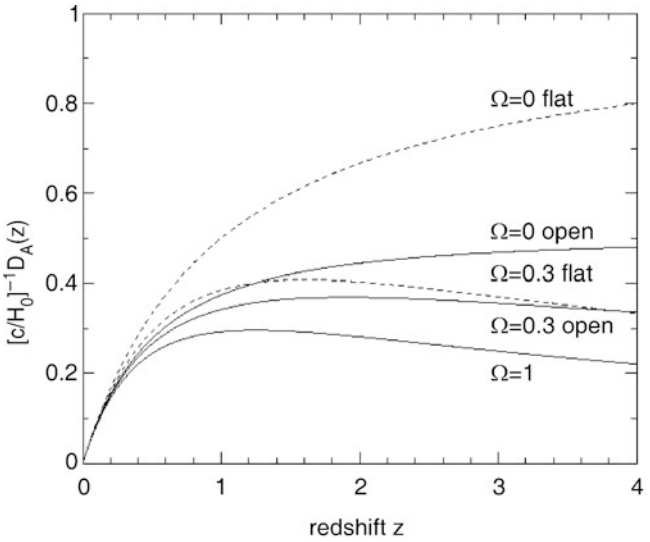


Fig. 4.11 Angular-diameter distance vs. redshift for different cosmological models. *Solid curves* display models with no vacuum energy; *dashed curves* show flat models with $\Omega_\mathrm{m} + \Omega_\Lambda = 1$. In both cases, results are plotted for Ω_m = 1, 0.3, and 0. Adopted from J.A. Peacock 1999, *Cosmological Physics*, Cambridge University Press

$$D_\mathrm{A}(z) = \frac{c}{H_0}\,\frac{2}{\Omega_\mathrm{m}^2\,(1+z)^2} \times \left[\Omega_\mathrm{m} z + (\Omega_\mathrm{m} - 2)\left(\sqrt{1+\Omega_\mathrm{m} z} - 1\right)\right]\,. \tag{4.51}$$

In particular, D_A is not necessarily a monotonic function of z. To better comprehend this, we consider the geometry on the surface of a sphere. Two great circles on Earth are supposed to intersect at the North Pole enclosing an angle $\varphi \ll 1$—they are therefore meridians. The separation L between these two great circles, i.e., the length of the connecting line perpendicular to both great circles, can be determined as a function of the distance D from the North Pole, which is measured as the distance along one of the two great circles. If θ is the geographical latitude ($\theta = \pi/2$ at the North Pole, $\theta = -\pi/2$ at the South Pole), $L = R\varphi\cos\theta$ is found, where R is the radius of the Earth. L vanishes at the North Pole, attains its maximum at the equator (where $\theta = 0$), and vanishes again at the South Pole; this is because both meridians intersect there again. Furthermore, $D = R(\pi/2-\theta)$, e.g., the distance to the equator $D = R\pi/2$ is a quarter of the Earth's circumference. Solving the last relation for θ, the distance is then given by $L = R\varphi\cos(\pi/2 - D/R) = R\varphi\sin(D/R)$. For the angular-diameter distance on the Earth's surface, we define $D_\mathrm{A}(D) = L/\varphi = R\sin(D/R)$, in analogy to the definition (4.49). For values of D that are considerably smaller than the curvature radius R of the sphere, we therefore obtain that $D_\mathrm{A} \approx D$, whereas for larger D, D_A deviates considerably from D. In particular, D_A is not a monotonic

function of D, rather it has a maximum at $D = \pi R/2$ and then decreases for larger D.

There exists a general relation between angular-diameter distance and luminosity distance,

$$\boxed{D_{\rm L}(z) = (1+z)^2\, D_{\rm A}(z)}\,. \tag{4.52}$$

The reader might now ask which of these distances is the *correct one?* Well, this question does not make sense since there is no unique definition of *the* distance in a curved spacetime like our Universe. Instead, the aforementioned measurement prescriptions must be used. The choice of a distance definition depends on the desired application of this distance. For example, if we want to compute the linear diameter of a source with observed angular diameter, the angular-diameter distance must be employed because it is defined just in this way. On the other hand, to derive the luminosity of a source from its redshift and observed flux, the luminosity distance needs to be applied. Due to the definition of the angular-diameter distance (length/angular diameter), those are the relevant distances that appear in the gravitational lens equation (3.63). A statement that a source is located "at a distance of 7 billion light years" away from us is meaningless unless it is mentioned which type of distance is meant. Again, in the low-redshift Universe ($z \ll 1$), the differences between different distance definitions are very small, and thus it *is* meaningful to state, for example, that the Coma cluster of galaxies lies at a distance of ~ 90 Mpc.

In Fig. 4.12 a Hubble diagram extending to high redshifts is shown, where the brightest galaxies in clusters of galaxies have been used as approximate standard candles. With an assumed constant intrinsic luminosity for these galaxies, the apparent magnitude is a measure of their distance, where the luminosity distance $D_{\rm L}(z)$ must be applied to compute the flux as a function of redshift.

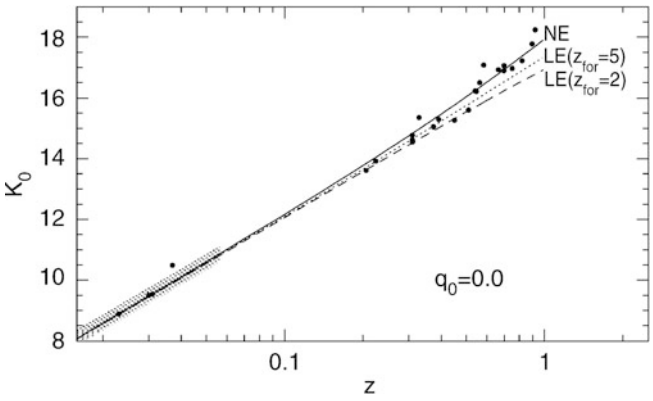


Fig. 4.12 A modern Hubble diagram: for several clusters of galaxies, the K-band magnitude of the brightest cluster galaxy is plotted versus the escape velocity, measured as redshift $z = \Delta\lambda/\lambda$ (*symbols*). If these galaxies all had the same luminosity, the apparent magnitude would be a measure of distance. For low redshifts, the curves follow the linear Hubble law (4.9), with $z \approx v/c$, whereas for higher redshifts modifications to this law are necessary. The *solid curve* corresponds to a constant galaxy luminosity at all redshifts, whereas the two other *curves* take evolutionary effects of the luminosity into account according to models of population synthesis (Sect. 3.5). Two different epochs of star formation were assumed for these galaxies. The diagram is based on a cosmological model with a deceleration parameter of $q_0 = 0$ [see (4.35)]. Source: A. Aragon-Salamanca et al. 1998, *The K-band Hubble diagram for the brightest cluster galaxies: a test of hierarchical galaxy formation models*, MNRAS 297, 427, p. 429, Fig. 1. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

We compile several expressions that are required to compute distances in general Friedmann–Lemaître models (see also problem 4.6). To do this, we need to define the function

$$f_K(x) = \begin{cases} 1/\sqrt{K}\,\sin(\sqrt{K}x) & K > 0 \\ x & K = 0 \\ 1/\sqrt{-K}\,\sinh(\sqrt{-K}x) & K < 0 \end{cases},$$

where K is the curvature scalar (4.32). The comoving radial distance x of a source at redshift z can be computed using $\mathrm{d}x = a^{-1}\,\mathrm{d}r = -a^{-1}\,c\,\mathrm{d}t = -c\,\mathrm{d}a/(a^2 H)$. Hence with (4.33)

$$x(z) = \int_{(1+z)^{-1}}^{1} \frac{\mathrm{d}a\ (c/H_0)}{\sqrt{a\Omega_{\rm m} + a^2(1-\Omega_{\rm m}-\Omega_\Lambda) + a^4\Omega_\Lambda}}\,. \tag{4.53}$$

The angular-diameter distance is then given as

$$D_{\rm A}(z) = \frac{1}{1+z}\, f_K\,[x(z)]\ , \tag{4.54}$$

and thus can be computed for all redshifts and cosmological parameters by (in general numerical) integration of (4.53). The luminosity distance then follows from (4.52). The angular-diameter distance of a source at redshift z_2, as measured by an observer at redshift $z_1 < z_2$, reads

$$D_{\rm A}(z_1, z_2) = \frac{1}{1+z_2}\, f_K\,[x(z_2) - x(z_1)]\ . \tag{4.55}$$

This is the distance that is required in equations of gravitational lens theory for $D_{\rm ds}$. In particular, $D_{\rm A}(z_1, z_2) \neq D_{\rm A}(z_2) - D_{\rm A}(z_1)$.

Sometimes, the look-back time is used as another quantity characterizing the 'distance' of a source. It is defined as the time the light traveled from a source at redshift z to us, and can be calculated in analogy to (4.36), with the lower and upper limit of integration being $a = (1+z)^{-1}$ and 1, respectively.

4.3.4 Special case: The Einstein–de Sitter model

As a final note in this section, we will briefly examine one particular cosmological model more closely, namely the model with $\Omega_\Lambda = 0$ and vanishing curvature, $K = 0$, and hence $\Omega_{\rm m} = 1$. We disregard the radiation component, which

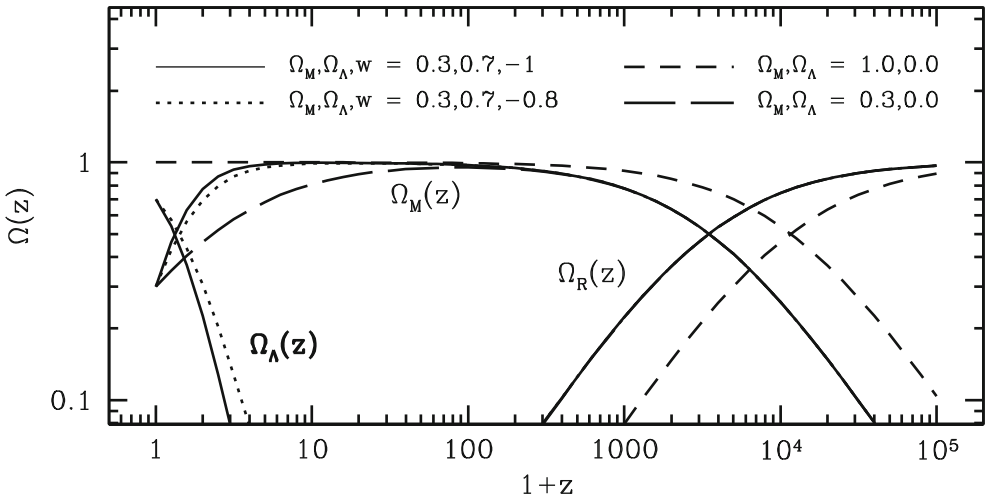


Fig. 4.13 The ratio of the density of the different components in the universe to the critical density $\rho_{\rm cr}(z) = 3H^2(z)/(8\pi G)$, as a function of redshift, for four different cosmological models: the *solid curves* correspond to the model which is presumably the one we live in, the *short-dashed curves* correspond to an Einstein–de Sitter model, and the *long-dashed curves* show a low density universe without dark energy. Finally, the *dotted curve* corresponds to a case with a different model for the dark energy. In all cases, radiation dominates the energy density of the universe at early times, i.e., at high redshifts, whereas for z below $\sim 10^4$ the universe becomes matter dominated. Only at redshift below ~ 2 does dark energy contribute significantly to the energy budget, but then quickly starts to dominate. Source: M. Voit 2004, *Tracing cosmic evolution with clusters of galaxies*, astro-ph/0410173, Fig. 2. Reproduced with permission of the author

contributes to the expansion only at very early times and thus for very small a. For a long time, this Einstein–de Sitter (EdS) model was the preferred model among cosmologists because inflation (see Sect. 4.5.3) predicts $K = 0$ and because a finite value for the cosmological constant was considered 'unnatural'. In fact, as late as the mid-1990s, this model was termed the 'standard model'. In the meantime we have learned that $\Lambda \neq 0$; thus we are not living in an EdS universe. But there is at least one good reason to examine this model a bit more, since the expansion equations become much simpler for these parameters and we can formulate simple explicit expressions for the quantities introduced above. These then yield estimates which for other model parameter values are only possible by means of numerical integration.

The resulting expansion equation $\dot{a} = H_0\, a^{-1/2}$ is easily solved by making the ansatz $a = (Ct)^\beta$ which, when inserted into the equation, yields the solution

$$a(t) = \left(\frac{3\,H_0\,t}{2}\right)^{2/3} = \left(\frac{t}{t_0}\right)^{2/3} . \qquad (4.56)$$

Setting $a = 1$, we obtain the age of the Universe, $t_0 = 2/(3H_0)$. The same result also follows immediately from (4.36) if the parameters of an EdS model are inserted there. Using $H_0 \approx 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ results in an age of about 10 Gyr, which is slightly too low to be compatible with the age of the oldest star clusters. The angular-diameter distance (4.49) in an EdS universe is obtained by considering the Mattig relation (4.41) for the case $\Omega_{\rm m} = 1$:

$$D_{\rm A}(z) = \frac{2c}{H_0}\,\frac{1}{(1+z)}\left(1 - \frac{1}{\sqrt{1+z}}\right) ,$$
$$D_{\rm L}(z) = \frac{2c}{H_0}\,(1+z)\left(1 - \frac{1}{\sqrt{1+z}}\right) , \qquad (4.57)$$

where we used (4.52) to obtain the second relation from the first.

4.3.5 Summary

We shall summarize the most important points of the two preceding lengthy sections:

- Observations are compatible with the fact that the Universe around us is isotropic and homogeneous on large scales. The cosmological principle postulates this homogeneity and isotropy of the Universe.
- General Relativity allows homogeneous and isotropic world models, the Friedmann–Lemaître models. In the language of GR, the cosmological principle reads as follows: "A family of solutions of Einstein's field equations exists such that a set of comoving observers see the same history of the universe; for each of them, the universe appears isotropic."
- The *shape* of these Friedmann–Lemaître world models is characterized by the density parameter $\Omega_{\rm m}$ and by the cosmological constant Ω_Λ, the *size* by the Hubble constant H_0. The cosmological parameters determine the expansion rate of the universe as a function of time.

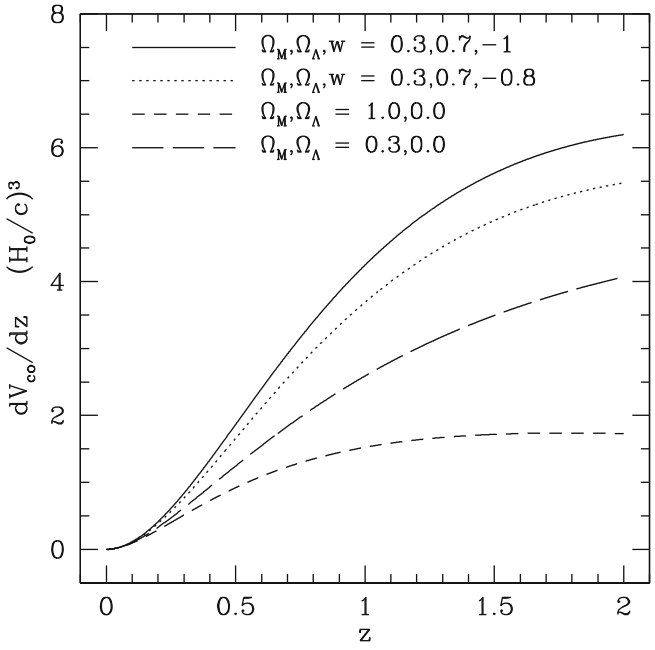


Fig. 4.14 The comoving volume of spherical shells per unit redshift interval, scaled by $(c/H_0)^3$, for the same cosmological models as shown in Fig. 4.13. Obviously, the transformation of a redshift interval into a volume element is a strong function of the density parameters; the volume is smallest for the Einstein–de Sitter model, and largest for the flat, low-density universe. Source: M. Voit 2004, *Tracing cosmic evolution with clusters of galaxies*, astro-ph/0410173, Fig. 3. Reproduced with permission of the author

- The scale factor $a(t)$ of the universe is a monotonically increasing function from the beginning of the universe until now; at earlier times the universe was smaller, denser, and hotter. There must have been an instant when $a \to 0$, which is called the Big Bang. The future of the expansion depends on Ω_m and Ω_Λ.
- The expansion of the universe causes a redshift of photons. The more distant a source is from us, the more its photons are redshifted.
- The relative contribution of radiation, matter and vacuum energy density changes over cosmic time, with radiation dominating in the first phase of the universe, changing to a matter-dominated universe, to become dark-energy dominated at late times, provided $\Omega_\Lambda > 0$ (see Fig. 4.13).
- 'Distance' in cosmology does not have a unique meaning. Depending on whether one relates fluxes to luminosity, or length scales with angular sizes, one needs to use different definitions of distances. The distance-redshift relations depend on the values of the cosmological parameters—they all scale with the Hubble length c/H_0, and depend on the density parameters. Accordingly, the volume of a spherical shell with given thickness in redshift also depends on the density parameters—which is important when source counts are used to infer number density of sources (see Fig. 4.14).

At http://www.astro.ucla.edu/~wright/CosmoCalc.html the reader can find an online calculator for distances, ages, lookback-times etc. as a function of redshift, for different cosmological parameters.

4.4 Thermal history of the Universe

Since $T \propto (1+z)$ our Universe was hotter at earlier times. For example, at a redshift of $z = 1100$ the temperature (of the CMB) was about $T \sim 3000\,\mathrm{K}$. And at an even higher redshift, $z = 10^9$, it was $T \sim 3 \times 10^9\,\mathrm{K}$, hotter than in a stellar interior. Thus we might expect energetic processes like nuclear fusion to have taken place in the early Universe.

In this section we shall describe the essential processes in the early universe. To do so we will assume that the laws of physics have not changed over time. This assumption is by no means trivial—we have no guarantee whatsoever that the cross sections in nuclear physics were the same 13 billion years ago as they are today. But if they have changed in the course of time the only chance of detecting this is through cosmology. Based on this assumption of time-invariant physical laws, we will study the consequences of the Big Bang model developed in the previous section and then compare them with observations. Only this comparison can serve as a test of the success of the model. A few comments should serve as preparation for the discussion in this section.

1. Temperature and energy may be converted into each other since $k_\mathrm{B}T$ has the dimension of energy. We use the electron volt (eV) to measure temperatures and energies, with the conversion $1\,\mathrm{eV} = 1.1605 \times 10^4\,k_\mathrm{B}\,\mathrm{K}$.
2. Elementary particle physics is very well understood for energies below $\sim 100\,\mathrm{GeV}$. For much higher energies our understanding of physics is a lot less certain. Therefore, we will begin the consideration of the thermal history of the cosmos at energies well below this scale.
3. Statistical physics and thermodynamics of elementary particles are described by quantum mechanics. A distinction has to be made between *bosons*, which are particles of integer spin (like the photon), and *fermions*, particles of half-integer spin (like, for instance, electrons, protons, neutrinos, and their anti-particles).
4. If particles are in thermodynamical and chemical equilibrium, their number density and their energy distribution are specified solely by the temperature—e.g., the Planck distribution (A.13), and thus the energy density of the radiation (4.26), is a function of T only.

The necessary condition for establishing chemical equilibrium is the possibility for particles to be created and destroyed, such as in e^+-e^--pair production and annihilation.

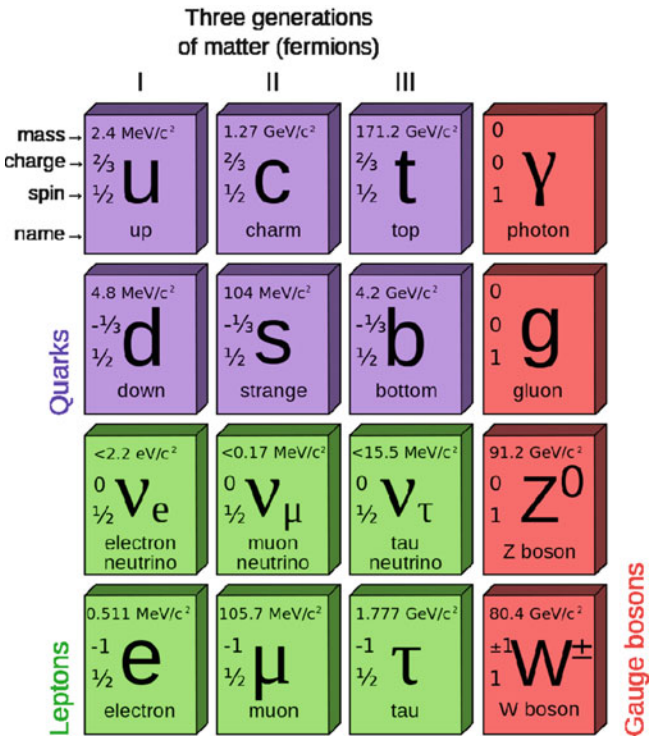


Fig. 4.15 The particles in the Standard Model: The six quarks (*violet*) and six leptons (*green*) are organized in three families; the four gauge bosons are shown in *red*. The *numbers* in each box indicate particle mass, charge (in units of the elementary charge) and spin. All particles have an antiparticle, except the photon and the Z-boson, which are their own antiparticles. Not shown in the diagram is the recently found Higgs particle. Source: Wikipedia

4.4.1 The Standard Model of particle physics

Before discussing the events in the early history of the Universe in more detail, it is useful to briefly summarize what we know about particle physics. Since about the 1970s, the Standard Model has been in place; it describes the elementary particles and their interactions—except for gravity.

Particle contents. According to this model, matter is composed of fermions, particles with half-integer spin, which obey the Pauli exclusion principle. The fermions are further divided into leptons, such as the electron and the electron neutrino, and quarks, such as the up and down quark. Figure 4.15 provides an overview on the particles of the Standard Model, together with information on their mass, charge and spin. All of the particles have anti-particles; for example, the positron is the anti-particle of the electron, its charge is minus the electron's charge; some of the particles are their own anti-particle, such as the photon. The quarks, particles with the strange property that their charge is not a multiple of the elementary charge, but thirds of it, form bound states, such as the proton and the neutron, being composed of three quarks each. In fact, according to the Standard Model, quarks do not occur isolated in nature, but are only found in compound systems of hadrons, such as the nucleons, or mesons, such as the pions. According to the model, the former consist of three quarks, the latter of a quark-antiquark system. Protons and neutrons are composed and up and down quarks, the lightest quarks. The electron, the electron neutrino, and the up and down quarks form the so-called first family of particles (see Fig. 4.15).

Except for the electron, there are two more charged leptons, the muon and the tau, which were discovered in 1936 and 1975, respectively. Both of these leptons are much heavier than the electron, and they are unstable: the muon decays into an electron and two neutrinos, whereas the tau can either decay into a muon or an electron, again accompanied with two neutrinos. However, these two charged leptons show properties very similar to those of the electron. For both of them, there is an associated neutrino, the μ-neutrino (ν_μ) and the τ-neutrino (ν_τ). Neutrinos are much more difficult to detect than the charged leptons, since they only interact weakly with matter. Therefore, they were directly detected only in 1956 (ν_e), 1962 (ν_μ) and 2000 (ν_τ)—note that the discovery of the tau-neutrino occurred well after the Standard Model had been well established and presents one of its successes.

Particle accelerators with ever increasing energies produced heavier particles. Among them were also particles which could not be described as composite particles consisting of up and down quarks only, but their understanding implied the presence of additional quarks. By 1970, the strange and charm quarks were thereby indirectly discovered. Together with the muon and its neutrino, these two new quarks form the second family of elementary particles. With the discovery of the tau, two more quarks were predicted, to form the third family—the bottom and top quark were indeed discovered in 1977 and 1995, again as bound states forming heavier compound particles.

Interactions; gauge bosons. According to the Standard Model, interactions between particles occur through the exchange of bosons. The carrier of the electromagnetic force is the photon. The exchange particles for the strong force between quarks is the gluon. Like the photon, it is electrically neutral, but unlike the photon, it carries a new property called color. As for the quarks, an isolated gluon has not been observed yet; nevertheless, its existence was indirectly verified in particle decays in 1979. The strong interaction between quarks transmitted by gluons is described by the theory of Quantum Chromodynamics (QCD), which is the strong interaction part of the Standard Model.

All leptons are subject to weak interactions; the Standard Model postulates the existence of two exchange bosons, the charged W-boson and the electrically neutral Z-boson. Their discovery had to await sufficiently powerful accelerators, but

both were finally found in 1983, a beautiful and impressive confirmation of the Standard Model. The Standard Model predicts that electromagnetic and weak interactions are unified to the electroweak interaction. However, at low energies these two interactions appear to be quite different. The explanation for this difference is that the W- and Z-boson are very massive, whereas the photon is massless. Thus, at energy scales below the Z-boson mass, weak interactions are considerably weaker than electromagnetic ones.

The Higgs mechanism. According to the Standard Model as outlined so far, all particles are intrinsically massless. Obviously, this is not the case; for example, the electron, the muon and the tau have a finite rest mass. One finds that the quarks have a finite rest mass as well. As an aside, we note that the mass of a nucleon is much higher than the sum of the masses of its three constituent quarks; most of the nucleon mass stems from the strong interactions transmitted by the gluons.

A mechanism for particles to obtain a finite rest mass was proposed in the 1960s, and this so-called Higgs-mechanism for symmetry breaking has been widely accepted and became part of the Standard Model. It is responsible for the large masses of the W and Z bosons, and thus for the different appearance of electroweak interactions as weak and electromagnetic ones at low energies. This Higgs mechanism implies the existence of an additional particle, called the Higgs particle. The search for this Higgs particle was one of the main drivers for building the most complex machine ever made by mankind—the Large Hadron Collider at CERN. It can generate sufficiently high energies for the Higgs particle to be generated and discovered. Indeed, in the summer of 2013, scientists from two different collaborations announced the discovery of a new elementary particle, which after a short period was verified as being the long-sought Higgs particle—a spectacular success! With that, the final missing piece of the Standard Model was found. The Nobel Prize in physics 2013 for F. Englert and P. Higgs for the theoretical development of this mechanism acknowledges also the importance of this discovery.

4.4.2 Expansion in the radiation-dominated phase

As mentioned above (4.30), the energy density of radiation dominates in the early universe, at redshifts $z \gg z_{\rm eq}$ where

$$\boxed{z_{\rm eq} = a_{\rm eq}^{-1} - 1 \approx 23\,900\,\Omega_{\rm m}\,h^2}\,. \qquad (4.58)$$

The radiation density behaves like $\rho_{\rm r} \propto T^4$, where the constant of proportionality depends on the number of species of relativistic particles (these are the ones for which $k_{\rm B}T \gg mc^2$). Since $T \propto 1/a$ and thus $\rho_{\rm r} \propto a^{-4}$, radiation then dominates in the expansion equation (4.18). The latter can be solved by a power law, $a(t) \propto t^{\beta}$, which after insertion into the expansion equation yields $\beta = 1/2$ and thus

$$\boxed{\begin{array}{l} a \propto t^{1/2}\,, \quad t = \sqrt{\dfrac{3}{32\pi G\rho}}\,, \\ t \propto T^{-2} \quad \text{in radiation-dominated phase} \end{array}}\,, \qquad (4.59)$$

where the constant of proportionality depends again on the number of relativistic particle species. Since the latter is known from particle physics, assuming thermodynamical equilibrium, the time dependence of the early expansion is uniquely specified by (4.59). This is reasonable because for early times neither the curvature term nor the cosmological constant contribute significantly to the expansion dynamics.

4.4.3 Decoupling of neutrinos

We start our consideration of the universe at a temperature of $T \approx 10^{12}\,\mathrm{K}$ which corresponds to $\sim 100\,\mathrm{MeV}$. This energy can be compared to the rest mass of various particles:

proton, $m_{\rm p} = 938.3\,\mathrm{MeV}/c^2$,
neutron, $m_{\rm n} = 939.6\,\mathrm{MeV}/c^2$,
electron, $m_{\rm e} = 511\,\mathrm{keV}/c^2$,
muon, $m_{\mu} = 140\,\mathrm{MeV}/c^2$.

Protons and neutrons (i.e., the baryons) are too heavy to be produced at the temperature considered. Thus all baryons that exist today must have been present already at this early time. Also, the production of muon pairs, according to the reaction $\gamma + \gamma \rightarrow \mu^+ + \mu^-$, is not efficient because the temperature, and thus the typical photon energy, is not sufficiently high. Hence, at the temperature considered the following relativistic particle species are present: electrons and positrons, photons and neutrinos. These species contribute to the radiation density $\rho_{\rm r}$. The mass of the neutrinos is not accurately known, though we recently learned that they have a small but finite rest mass. As will be explained in Sect. 8.7, cosmology allows us to obtain a very strict limit on the neutrino mass, which is currently below 1 eV. For the purpose of this discussion they may thus be considered as massless.

In addition to relativistic particles, non-relativistic particles also exist. These are the protons and neutrons, and probably also the constituents of dark matter. We assume that the latter consists of weakly interacting massive particles

(WIMPs), with rest mass larger than $\sim 100\,\mathrm{GeV}$ because up to these energies no WIMP candidates have been found in terrestrial particle accelerator laboratories. With this assumption, WIMPs are non-relativistic at the energies considered. Thus, like the baryons, they virtually do not contribute to the energy density in the early universe.

Apart from the WIMPs, all the aforementioned particle species are in equilibrium, e.g., by the following reactions:

$\mathrm{e}^{\pm} + \gamma \leftrightarrow \mathrm{e}^{\pm} + \gamma$: Compton scattering,
$\mathrm{e}^{+} + \mathrm{e}^{-} \leftrightarrow \gamma + \gamma$: pair-production and annihilation,
$\nu + \bar{\nu} \leftrightarrow \mathrm{e}^{+} + \mathrm{e}^{-}$: neutrino-antineutrino-scattering,
$\nu + \mathrm{e}^{\pm} \leftrightarrow \nu + \mathrm{e}^{\pm}$: neutrino-electron scattering,
$\mathrm{e}^{\pm} + \mathrm{p} \leftrightarrow \mathrm{e}^{\pm} + \mathrm{p} + \gamma$: Bremsstrahlung.

Reactions involving baryons will be discussed later. The energy density at this epoch is[8]

$$\rho = \rho_{\mathrm{r}} = 10.75\,\frac{\pi^2}{30}\,\frac{(k_{\mathrm{B}}\,T)^4}{(\hbar\,c)^3}\,, \tag{4.60}$$

which yields—see (4.59)—

$$t \approx 0.3\,\mathrm{s}\left(\frac{T}{1\,\mathrm{MeV}}\right)^{-2}\,. \tag{4.61}$$

Hence, about one second after the Big Bang the temperature of the Universe was about $10^{10}\,\mathrm{K}$. For the particles to maintain equilibrium, the reactions above have to occur at a sufficient rate. The equilibrium state, specified by the temperature, continuously changes due to the expansion of the Universe, so that the particle distribution needs to continually adjust to this changing equilibrium. This is possible only if the mean time between two reactions is much shorter than the time-scale on which equilibrium conditions change. The latter is given by the expansion. This means that the reaction rates (the number of reactions per particle per unit time) must be larger than the cosmic expansion rate $H(t)$ in order for the particles to maintain equilibrium.

The reaction rates Γ are proportional to the product of the number density n of the reaction partner particles and the cross section σ of the corresponding reaction. Both decrease with time: the number density decreases as $n \propto a^{-3} \propto t^{-3/2}$ because of the expansion. Furthermore, the cross sections for weak interaction, which is responsible for the reactions involving neutrinos, depend on energy, approximately as $\sigma \propto E^2 \propto T^2 \propto a^{-2}$. Together this yields $\Gamma \propto n\sigma \propto a^{-5} \propto t^{-5/2}$, whereas the expansion rate decreases only as $H \propto t^{-1}$. At sufficiently early times, the reaction rates were larger than the expansion rate, and thus particles remained in equilibrium. Later, however, the reactions no longer took place fast enough to maintain equilibrium. The time or temperature, respectively, of this transition can be calculated from the cross section of weak interaction,

$$\frac{\Gamma}{H} \approx \left(\frac{T}{1.6\times 10^{10}\,\mathrm{K}}\right)^3\,,$$

so that for $T \lesssim 10^{10}\,\mathrm{K}$ neutrinos are no longer in equilibrium with the other particles. This process of decoupling from the other particles is also called *freeze-out*; neutrinos freeze out at $T \sim 10^{10}\,\mathrm{K}$. At the time of freeze-out, they had a thermal distribution with the same temperature as the other particle species which stayed in mutual equilibrium. From this time on neutrinos propagate without further interactions, and so have kept their thermal distribution up to the present day, with a temperature decreasing as $T \propto 1/a$. This consideration predicts that these neutrinos, which decoupled from the rest of the matter about one second after the Big Bang, are still around in the Universe today. They have a number density of $113\,\mathrm{cm}^{-3}$ per neutrino family and are at a temperature of $1.9\,\mathrm{K}$ (this value will be explained in more detail below). However, these very low energy neutrinos are currently undetectable because of their extremely low cross section.

The expansion behavior is unaffected by the neutrino freeze-out and continues to proceed according to (4.61).

4.4.4 Pair annihilation

At temperatures smaller than $\sim 5\times 10^9\,\mathrm{K}$, or $k_{\mathrm{B}}T \sim 500\,\mathrm{keV}$, electron-positron pairs can no longer be produced efficiently since the number density of photons with energies above the pair production threshold of 511 keV is becoming too small. However, the annihilation $\mathrm{e}^{+} + \mathrm{e}^{-} \rightarrow \gamma + \gamma$ continues to proceed and, due to its large cross section, the density of $\mathrm{e}^{+}\mathrm{e}^{-}$-pairs decreases rapidly.

Pair annihilation injects additional energy into the photon gas, originally present as kinetic and rest mass energy of the $\mathrm{e}^{+}\mathrm{e}^{-}$ pairs. This changes the energy distribution of photons, which continues to be a Planck distribution but now with a modified temperature relative to that it would have had without annihilation. The neutrinos, already decoupled at this time, do not benefit from this additional energy. This means that after the annihilation the photon temperature exceeds that of the neutrinos. From the thermodynamics of this process, the change in photon temperature is computed as

[8] Compare this energy density with that of a blackbody photon distribution; they are the same except for the prefactor. This prefactor is determined by the number of bosonic and fermionic particle species which are relativistic at temperature T.

$$\boxed{\begin{aligned}&(a\,T)(\text{after annihilation})\\&=\left(\frac{11}{4}\right)^{1/3}(a\,T)(\text{before annihilation})\\&=\left(\frac{11}{4}\right)^{1/3}(a\,T_\nu)\end{aligned}}\,. \tag{4.62}$$

This temperature ratio is preserved afterwards, so that neutrinos have a temperature lower than that of the photons by $(11/4)^{1/3} \sim 1.4$—until the present epoch. This result has already been mentioned and taken into account in the estimate of $\rho_{\mathrm{r},0}$ in (4.28); we find $\rho_{\mathrm{r},0} = 1.68\rho_{\mathrm{CMB},0}$.

The factor 1.68 in the foregoing equation originates from the fact that the energy density of neutrinos is related to that of the photons through

$$\rho_\nu = N_{\mathrm{eff}}\,\frac{7}{8}\left(\frac{4}{11}\right)^{4/3}\rho_{\mathrm{CMB}}\,, \tag{4.63}$$

where N_{eff} is the number of neutrino families, the factor (7/8) is derived from quantum statistics and accounts for the fact that neutrinos are fermions, whereas photons are bosons, and the factor $(4/11)^{4/3}$ stem from the different temperatures of neutrinos and photons after pair annihilation. With three neutrino families, one has $N_{\mathrm{eff}} = 3$, according to the consideration above. However, since the temperature at which neutrino freeze-out happens, is very close to that of pair annihilation, the treatment of both processes as done above is slightly simplistic. We have assumed that the neutrinos are fully decoupled before pair annihilation sets in; an accurate treatment accounts for the fact that these processes are not fully decoupled. Such an accurate treatment confirms the relation (4.63), but with a slightly different value of $N_{\mathrm{eff}} = 3.046$.

After pair annihilation, the expansion law

$$t = 0.55\,\mathrm{s}\left(\frac{T}{1\,\mathrm{MeV}}\right)^{-2} \tag{4.64}$$

applies. This means that, as a result of annihilation, the constant in this relation changes compared to (4.61) because the number of relativistic particles species has decreased. Furthermore, the ratio η of the baryon-to-photon number density remains constant after pair annihilation.[9] The former is characterized by the density parameter $\Omega_{\mathrm{b}} = \rho_{\mathrm{b},0}/\rho_{\mathrm{cr}}$ in baryons (today), and the latter is determined by T_0:

$$\boxed{\eta := \left(\frac{n_{\mathrm{b}}}{n_\gamma}\right) = 2.74\times10^{-8}\,(\Omega_{\mathrm{b}}h^2)}\,. \tag{4.65}$$

As we will see in a moment, in our Universe $\Omega_{\mathrm{b}}h^2 \approx 0.02$, which means that for every baryon there are about two billion photons. Before pair annihilation there were about as many electrons and positrons as there were photons. After annihilation *nearly* all electrons were converted into photons—but not entirely because there was a very small excess of electrons over positrons to compensate for the positive electrical charge density of the protons. Therefore, the number density of electrons that survive the pair annihilation is exactly the same as the number density of protons, for the Universe to remain electrically neutral. Thus, the ratio of electrons to photons is also given by η, or more precisely by about 0.8η, since η includes both protons and neutrons.

[9] The total number of photons emitted during stellar evolution is negligible compared to the number of CMB photons.

4.4.5 Primordial nucleosynthesis

Protons and neutrons can fuse to form atomic nuclei if the temperature and density of the plasma are sufficiently high. In the interior of stars, these conditions for nuclear fusion are provided. The high temperatures in the early phases of the Universe suggest that atomic nuclei may also have formed then. As we will discuss below, in the first few minutes after the Big Bang some of the lightest atomic nuclei were formed. The quantitative discussion of this primordial nucleosynthesis (Big Bang nucleosynthesis, BBN) will explain observation (4) of Sect. 4.1.1.

Proton-to-neutron abundance ratio. As already discussed, the baryons (or nucleons) do not play any role in the expansion dynamics in the early universe because of their low density. The most important reactions through which they maintain chemical equilibrium with the rest of the particles are

$$\begin{aligned}\mathrm{p}+\mathrm{e}&\leftrightarrow\mathrm{n}+\nu\,,\\ \mathrm{p}+\bar\nu&\leftrightarrow\mathrm{n}+\mathrm{e}^+\,,\\ \mathrm{n}&\to\mathrm{p}+\mathrm{e}+\bar\nu\,.\end{aligned}$$

The latter is the decay of free neutrons, with a time-scale for the decay of $\tau_{\mathrm{n}} = 881\,\mathrm{s}$. The first two reactions maintain the equilibrium proton-to-neutron ratio as long as the corresponding reaction rates are large compared to the expansion rate. The equilibrium distribution is specified by the Boltzmann factor,

$$\frac{n_{\mathrm{n}}}{n_{\mathrm{p}}} = \exp\left(-\frac{\Delta m\,c^2}{k_{\mathrm{B}}T}\right)\,, \tag{4.66}$$

where $\Delta m = m_{\mathrm{n}} - m_{\mathrm{p}} = 1.293\,\mathrm{MeV}/c^2$ is the mass difference between neutrons and protons. Hence, neutrons are slightly heavier than protons; otherwise the neutron decay would not be possible. After neutrino freeze-out equilibrium reactions become rare because the above reactions are based on weak interactions, the same as those that kept the neutrinos in chemical equilibrium. At the time of neutrino decoupling, we have $n_{\mathrm{n}}/n_{\mathrm{p}} \approx 1/3$. After this, protons and neutrons are no longer in equilibrium, and their ratio is no longer described by (4.66). Instead, it changes only by the decay of free neutrons on the time-scale τ_{n}. To have neutrons

survive at all until the present day, they must quickly become bound in atomic nuclei.

Deuterium formation. The simplest compound nucleus is that of deuterium (D), consisting of a proton and a neutron and formed in the reaction

$$\mathrm{p} + \mathrm{n} \rightarrow \mathrm{D} + \gamma \ .$$

The binding energy of D is $E_\mathrm{b} = 2.225\,\mathrm{MeV}$. This energy is only slightly larger than $m_\mathrm{e}c^2$ and Δm—all these energies are comparable. The formation of deuterium is based on strong interactions and therefore occurs very efficiently. However, at the time of neutrino decoupling and pair annihilation, T is not much smaller than E_b. This has an important consequence: because photons are so much more abundant than baryons, a sufficient number of highly energetic photons, with $E_\gamma \geq E_\mathrm{b}$, exist in the Wien tail of the Planck distribution to instantly destroy newly formed D by photo-dissociation. Only when the temperature has decreased considerably, $k_\mathrm{B}T \ll E_\mathrm{b}$, can the deuterium abundance become appreciable. With the corresponding balance equations we can calculate that the formation rate exceeds the photo-dissociation rate of deuterium at about $T_\mathrm{D} \approx 8 \times 10^8\,\mathrm{K}$, corresponding to $t \sim 3\,\mathrm{min}$. Up to then, a fraction of the neutrons has thus decayed, yielding a neutron-proton ratio at T_D of $n_\mathrm{n}/n_\mathrm{p} \approx 1/7$.

After that time, everything happens very rapidly. Owing to the strong interaction, virtually all neutrons first become bound in D. Once the deuterium density has become appreciable, helium (He^4) forms, which is a nucleus with high binding energy ($\sim 28\,\mathrm{MeV}$) which can therefore not be destroyed by photo-dissociation. Except for a small (but, as we will later see, very important) remaining fraction, all deuterium is quickly transformed into He^4. For this reason, the dependence of helium formation on the small binding energy of D is known as the 'bottleneck of nucleosynthesis'.

Helium abundance. The number density of helium nuclei can now be calculated, since virtually all neutrons present are bound in He^4. First, $n_\mathrm{He} = n_\mathrm{n}/2$, since every helium nucleus contains two neutrons. Second, the number density of free protons after the formation of helium is $n_\mathrm{H} = n_\mathrm{p} - n_\mathrm{n}$, since He^4 contains an equal number of protons and neutrons. From this, the *mass fraction Y of* He^4 *of the baryon density* follows,

$$\boxed{Y = \frac{4n_\mathrm{He}}{4n_\mathrm{He} + n_\mathrm{H}} = \frac{2n_\mathrm{n}}{n_\mathrm{p} + n_\mathrm{n}} = \frac{2(n_\mathrm{n}/n_\mathrm{p})}{1 + (n_\mathrm{n}/n_\mathrm{p})} \approx 0.25} \ , \tag{4.67}$$

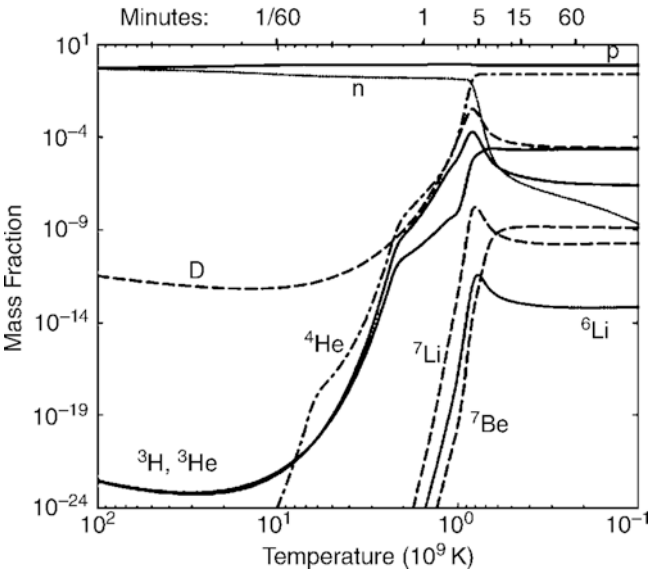


Fig. 4.16 The evolution of abundances of the light elements formed in BBN, as a function of temperature (lower axis) and cosmic time t (upper axis). The decrease in neutron abundance in the first $\sim 3\,\mathrm{min}$ is due to neutron decay. The density of deuterium increases steeply—linked to the steep decrease in neutron density—and reaches a maximum at $t \sim 3\,\mathrm{min}$ because then its density becomes sufficiently large for efficient formation of He^4 to set in. Only a few deuterium nuclei do not find a reaction partner and remain, with a mass fraction of $\sim 10^{-5}$. Only a few other light nuclei are formed in the Big Bang, mainly He^3 and Li^7. Source: D. Tytler, J.M. O'Meara, N. Suzuki & D. Lubin 2000, *Deuterium and the baryonic density of the universe*, Phys. Rep. 333, 409–432. Reprinted with permission from Elsevier

where in the last step we used the above ratio of $n_\mathrm{n}/n_\mathrm{p} \approx 1/7$ at T_D. This consideration thus leads to the following conclusion:

> About 1/4 of the baryonic mass in the Universe should be in the form of He^4. This is a robust prediction of Big Bang models, and it is in excellent agreement with observations.

The helium content in the Universe changes later by nuclear fusion in stars, which also forms heavier nuclei ('metals'), but as derived in problem 2.2, the total amount of helium produced in stars is expected to be smaller by about one order of magnitude compared to that in BBN. Observations of fairly unprocessed material (i.e., that which has a low metal content) reveal that in fact $Y \approx 0.25$. Figure 4.16 shows the result of a quantitative model of BBN where the mass fraction of several species is plotted as a function of time or temperature, respectively.

Dependence of the primordial abundances on the baryon density. At the end of the first 3 min, the composition of the baryonic component of our Universe is about as follows: 25 % of the baryonic mass is bound in helium nuclei, 75 % in hydrogen nuclei (i.e., protons), with traces of D, He^3 and Li^7. Heavier nuclei cannot form because no stable nucleus of mass number 5 or 8 exists and thus no new, stable nuclei can be formed in collisions of two helium nuclei or of a proton with a helium nucleus. Collisions between three nuclei are far too rare to contribute to nucleosynthesis. The density in He^4 and D depends on the baryon density in the Universe, as can be seen in Fig. 4.17 and through the following considerations:

- The larger the baryon density Ω_b, thus the larger the baryon-to-photon ratio η (4.65), the earlier D can form, i.e., the fewer neutrons have decayed, which then results in a larger n_n/n_p ratio. From this and (4.67) it follows that Y increases with increasing Ω_b.
- A similar argument is valid for the abundance of deuterium: the larger Ω_b is, the higher the baryon density during the conversion of D into He^4. Thus the conversion will be more efficient and more complete. This means that fewer deuterium nuclei remain without a reaction partner for helium formation. Thus fewer of them are left over in the end, so the fraction of D will be lower.

Baryon content of the Universe. From measurements of the primordial abundances of He^4 and D and their comparison with detailed models of nucleosynthesis in the early Universe, η or Ω_b, respectively, can be determined (see Fig. 4.17). The abundance of deuterium is a particularly sensitive measure for Ω_b. Measurements of the relative strength of the Lyα lines of H and D, which have slightly different transition frequencies due to the different masses of their nuclei, in QSO absorption lines (see Sect. 5.7) yields $D/H \approx 3.4 \times 10^{-5}$. Since the intergalactic gas producing these absorption lines is very metal-poor and thus presumably barely affected by nucleosynthesis in stars, its D/H-ratio should be close to the primordial value. Combining the quoted value of D/H with the model curves shown in Fig. 4.17 we find

$$\boxed{\Omega_b h^2 \approx 0.02} \,. \tag{4.68}$$

With a Hubble constant of $H_0 \sim 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, thus $h^2 \approx 1/2$, we have $\Omega_b \approx 0.04$. But since $\Omega_m > 0.1$, this result implies that baryons represent only a small fraction of the matter in the Universe. *The major fraction of matter is non-baryonic dark matter.*

To circumvent the conclusion of a dominant fraction of non-baryonic matter, inhomogeneous models of BBN have been investigated, but these also yield values for Ω_b which are too low and therefore do not provide a viable alternative.

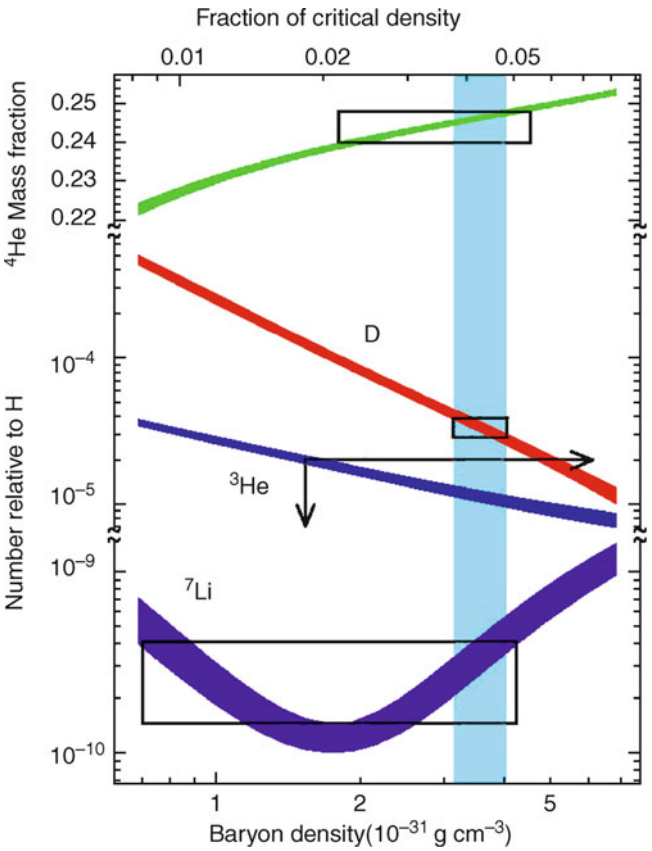


Fig. 4.17 BBN predictions of the primordial abundances of light elements as a function of today's baryon density ($\rho_{b,0}$, lower axis) and the corresponding density parameter Ω_b where $h = 0.65$ was assumed. The vertical extent of the rectangles marks the measured values of the abundances (*top*: He^4, *center*: D, *bottom*: Li^7). The horizontal extent results from the overlap of these intervals with curves computed from theoretical models. The ranges in Ω_b that are allowed by these three species do overlap, as is indicated by the vertical strip. The deuterium measurements yield the most stringent constraints for Ω_b. Source: D. Tytler, J.M. O'Meara, N. Suzuki & D. Lubin 2000, *Deuterium and the baryonic density of the universe*, Phys. Rep. 333, 409–432. Reprinted with permission from Elsevier

Dependence of BBN on the number of neutrino flavors. In the analysis of BBN we implicitly assumed that not more than three (relativistic, i.e., with $m_\nu < 1\,\mathrm{MeV}$) neutrino families exist. If $N_\nu > 3$, the quantitative predictions of BBN will change. In this case, the expansion would occur faster [see (4.59)] because $\rho(T)$ would be larger, leaving less time until the temperature has cooled down to T_D—thus, fewer neutrons would decay and the resulting helium abundance would be higher. Even before 1990, it was concluded from BBN (with relatively large uncertainties, however) that $N_\nu = 3$. In 1990, the value of $N_\nu = 3$ was then confirmed in the laboratory from Z-boson decay.

4.4.6 WIMPs as dark matter particles

There is a wide variety of evidence for the existence of dark matter, from scales of individual galaxies (rotation curves of spirals), clusters (velocity dispersion of galaxies, X-ray temperature, lensing), to cosmological scales, where the baryon density as inferred from BBN is lower than the lower bound on the total matter density. The MACHO experiments described in Sect. 2.5.3 rule out astronomical objects as the dominant contribution to dark matter, at least in the halo of the Milky Way; furthermore, all obvious candidates for astronomical dark matter objects would yield very strong conflicts with observations, for example concerning metallicity. In addition, the fact that the mass density in the Universe is ~ 6 times higher than the baryonic density precludes any 'normal' astronomical objects as the main constituent of dark matter. Therefore, the solution of the dark matter issue must likely come from particle physics.

Constraints on the dark matter particle. Since dark matter particles are 'dark' they must be electrically neutral in order not to interact electromagnetically. Furthermore, the particle must be stable, or at least have a lifetime much longer than the age of the Universe, so that they are still around today. The only known neutral particles in the Standard Model (see Sect. 4.4.1) are the neutrinos and the neutron. However, the neutron is baryonic, and its density in the early Universe is very well constrained by BBN (see Sect. 4.4.5); furthermore, the free neutron is unstable and thus clearly not a viable dark matter candidate. The neutrinos in principle could be good dark matter candidates if they have a finite mass, since we know they exist, and we actually know their abundance and their temperature, which were determined at their decoupling—see Sect. 4.4.3. But they would be hot dark matter, and as such lead to a large-scale structure in the Universe that would be very different from the one we observe, as will be explained in more detail in Sect. 7.4.1. In particular, they would be too 'hot' to cluster on scales of galaxies, with their thermal velocity exceeding that of the escape velocity from galaxy halos. We thus conclude that none of the known particles is a viable dark matter candidate.

Physics beyond the Standard Model. This Standard Model of particle physics has proved extremely successful in describing subatomic physics, as discussed in Sect. 4.4.1. Predictions of low-energy electromagnetic phenomena agree with laboratory measurements to better than one part in a billion, and for electro-weak interactions, the agreement is better than 10^{-3}. Because of its strengths and of non-linearities, strong interactions are far more difficult to describe quantitatively from first principles, and thus the strong interaction sector of the standard model—quantum chromodynamics (QCD)—is less accurately tested than the weak and electromagnetic part.

Despite its successes, the standard model is known to be incomplete, and in at least one aspect, it directly conflicts with observations: According to the standard model, neutrinos should be massless. However, the Solar neutrino problem and its solution has shown this to be not the case: The (electron) neutrinos generated through nuclear fusion in the center of the Sun can escape, due to their small interaction cross section. These Solar neutrinos can be detected in (big!) terrestrial detectors.[10] However, the measured rate of electron neutrinos from the Sun is only half as large as expected from Solar models. This *Solar neutrino problem* kept physicists and astrophysicists busy for decades. Its solution consists of a purely quantum-mechanical process, that of neutrino oscillations. It it possible that during its propagation, an electron neutrino transforms into a muon or tau neutrino, and back again. The existence of such oscillations was speculated for a long time, and provides a possible explanation for the missing rate of Solar electron neutrinos. Indeed, in 2001 the Sudbury Neutrino Observatory showed that neutrinos of all three flavors are received from the Sun, and that the sum of the rates of these neutrinos is compatible with the predicted rate from Solar models. In the meantime, these neutrino oscillation have also been observed in terrestrial experiments with neutrino beams.

Whereas neutrino oscillations are therefore well established today, they are in conflict with the Standard Model, according to which neutrinos have zero rest mass. From Special Relativity one can easily show that massless particles can not change their 'identity' while propagating through space. The existence of neutrino oscillations necessarily requires that neutrinos have a finite rest mass. Indeed, the oscillation experiments were able to constrain these masses, since they determine the length-scale over which the flavor of neutrinos changes—more precisely, it depends on the difference of their squared mass m_i^2. One finds that $m_2^2 - m_1^2 = (7.6 \pm 0.2) \times 10^{-5}\,\mathrm{eV}^2$, and $|m_3^2 - m_2^2| \approx |m_3^2 - m_1^2| = (2.4 \pm 0.2) \times 10^{-3}\,\mathrm{eV}^2$. These squared-mass differences thus do not provide us with an absolute mass scale of neutrinos, but their mass is non-zero. That means that neutrinos contribute to the cosmic matter density today, giving a contribution of

$$\Omega_\nu h^2 = \frac{\sum m_{\nu i}}{91.5\,\mathrm{eV}}\,, \tag{4.69}$$

which depends only on the sum of the three neutrino masses—since the number density of neutrinos is known

[10] For their research in the field of Solar neutrinos, Raymond Davis and Masatoshi Koshiba were awarded with one half of the Nobel Prize in Physics in 2002. The other half was awarded to Ricardo Giacconi for his pioneering work in the field of X-ray astronomy.

from the thermal history after the Big Bang. If the neutrino masses take the lowest values allowed by the squared-mass differences given above, this contribution is about 0.1 %. We will see in Chap. 8 that observations of the large-scale structure in the Universe show that neutrinos cannot contribute a substantial fraction to the matter density. Indeed, these observations yield a constraint of $\sum m_{\nu i} \lesssim 1\,\mathrm{eV}$, and thus the upper bound on neutrino masses from cosmology are much stricter than those obtained from laboratory experiments. For the electron neutrino, an upper limit on its mass was determined from decay experiments of tritium, yielding $m_{\nu_\mathrm{e}} \lesssim 2\,\mathrm{eV}$, which together with the results from neutrino oscillations implies a maximum density of $\Omega_\nu < 0.12$.

Physical motivation for WIMPs. Hence, the Standard Model needs an extension which allows the existence of a finite neutrino mass. In addition, there are other issues with the Standard Model—it is "technically unnatural" since the energy scale of the Higgs boson, $\sim$ 125 GeV, which is also comparable to the electro-weak mass scale, i.e., the masses of the W and Z bosons, is so much smaller than the Planck mass,[11] and it can also not explain why there are more baryons than anti-baryons in the current Universe. The former of these problems is called the gauge hierarchy problem; in order to solve it, one needs some new physics at an energy scale of $\sim$ 100 GeV. There are several models, extending the Standard Model, which introduce new physics at this scale. Arguably, the most promising of those is supersymmetry.

Then suppose that in the extended model, there exists a electrically neutral particle X which is stable, has a mass of order the energy scale of the model, i.e., somewhere between 100 GeV and 10 TeV, and interacts only weakly.[12] Such Weakly Interacting Massive Particles (WIMPs) are the most promising dark matter candidates—because, if such a particle exists, it would have the right cosmic density to account for the dark matter.

[11] From the fundamental constants G, c and h_P, one can form a unique combination with the dimension of a mass, $m_\mathrm{Pl} = \sqrt{h_\mathrm{P} c/G} \sim 10^{19}$ GeV, called the Planck mass. This is the mass scale where one expects that General Relativity ceases to be valid and that it has to be generalized to a quantum theory of gravity. Up to now, no plausible model for such a quantum gravity has been found.

[12] One might be surprised about the assumption that a particle of such high mass should be stable—given that there are only very few particles known to be stable, and they all have very small mass—the heaviest one being the proton. The Standard Model predicts that the baryon number is a conserved quantity, i.e., one cannot change the net number of baryons. For example, in creating a proton–antiproton pair, the net baryon number is changed by $+1 - 1 = 0$; in the decay of the free neutron, baryon number is conserved as well. Since the proton is the lightest baryon, there is no particle into which it can decay without violating baryon conservation—that's why we believe the proton is stable. If such a conserved quantity exists in the extended model of elementary particles (for example, in supersymmetry there is a so-called R-parity), then the lightest of the particles which 'carries' this quantity must be stable as well.

At first sight, it may be difficult to see how one can estimate the cosmic mass density of a particle whose existence and properties are as yet unknown, but indeed we can. For that, recall how we obtained the abundance of neutrinos in the Universe. Above a certain temperature, they were in thermodynamic equilibrium with the rest of the matter in the Universe, but when their interaction rate became too slow, they dropped out of equilibrium with the rest of the matter in the Universe, and kept their comoving number density from then on.

Now, let's assume the WIMP particle X exists. At sufficiently early times, it was in thermodynamic equilibrium. Since it is weakly interacting by assumption, we have a good idea about its cross section, and thus conclude that it stays in equilibrium during the phase when the temperature of the Universe drops below $T \sim m_\mathrm{X} c^2 = \mathcal{O}(1\,\mathrm{TeV})$, i.e., when the particles become non-relativistic. Once this happens, the equilibrium number density is determined by the Boltzmann factor,

$$n_{\mathrm{X,eq}} \propto (m_\mathrm{X}\, T)^{3/2}\, \mathrm{e}^{-m_\mathrm{X}/T}$$

and thus starts to decrease rapidly with decreasing T. At $T \approx 0.05 m_\mathrm{X}$, the interaction rate of X becomes too small to keep them in equilibrium with the other particles present, they freeze out, and from then on have constant comoving number density. Their mass density is then easily obtained as the product of their number density and m_X. Amazingly, if $m_\mathrm{X} \approx 300\,\mathrm{GeV}$, the resulting density of these WIMPs would yield $\Omega_\mathrm{X} \sim 0.2 \approx \Omega_\mathrm{dm}$, with an uncertainty of about a factor $\sim$ 3 (owing to the as yet unknown detailed properties and thus the precise value of the interaction cross section of X).

Hence, if a massive WIMP exists with properties expected from particle theory—weakly interacting, and with a mass near the weak interaction mass scale—the cosmological density of these particles is just the observed dark matter density! This indeed is an astonishing result, sometimes called the 'WIMP miracle', a miracle perhaps too good to be just a random coincidence. For that reason, such a WIMP is the favorite candidate for the dark matter particle. Fortunately, this model can be experimentally verified, and there are three ways how this can be achieved.

Direct detection. The first one is the direct detection. These WIMPs, if they constitute the dark matter in our Galaxy, should also be present in our neighborhood and pass through the Earth. Since they are weakly interacting only, their cross section with ordinary matter is very small, and they are difficult to detect. Nevertheless, experiments were built to search for such particles through their scattering with detector material, i.e., atomic nuclei. Due to scattering, the WIMP will transfer a momentum to the nucleus, and the resulting energy gain can be used for detection. One can estimate the associated recoil energy, given a plausible mass of $\sim$ 300 GeV and the characteristic velocity of $\sim$ 200 km/s,

corresponding to typical velocities in the Galaxy. Then, the kinematics of the scattering process implies that the recoil energy is small, $\sim$ 100 keV. This energy causes a tiny temperature increase of the detector, which must be probed.

Several different methods for the WIMP detection have been turned into experiments. Since WIMP events will be rare, one needs to place the detectors in a well-shielded environment, in laboratories deep underground, so that the background of cosmic ray particles is strongly suppressed. In order to test whether a measured signal is indeed due to particles from outside the Solar system, one checks for an annual variation: Due to the orbit of the Earth around the Sun, our velocity relative to the Galactic frame changes over the year, and the event rate should behave accordingly.

Existing experiments have imposed bounds on combinations of the WIMP mass and its cross section, and ruled out a significant fraction of plausible parameter space. Improvements in the experiments give rise to the expectation that WIMPs can be detected within the next few years. In fact, some experiments have claimed a detection, and also saw an annual modulation. However, the corresponding estimates of the mass and cross section are ruled out by other experiments, so that the interpretation of these results is controversial at present.

Particle colliders. The Large Hadron Collider (LHC) at CERN started operation in 2009; its two major science drivers were the search for the Higgs particle, which has been achieved in the meantime, and the search for phenomena beyond the Standard Model of particle physics. As we argued before, there are good reasons to assume that new physics will appear beyond $\sim$ 100 GeV, an energy range probed by the LHC. However, although the LHC will probably be able to produce WIMPs—if they exist—it will not lead to a direct WIMP detection, due to the low interaction cross sections with matter. Therefore, indirect methods must be used. For example, if supersymmetry is the correct extension of the Standard Model, supersymmetric particles will be produced and decay in the detector, thereby producing the lightest supersymmetric particle—presumably the WIMP. From adding up the charges, momenta and energy of all particles in the reaction, one could then conclude that a neutral particle has left the detector, and get an estimate on its mass. This particle must have a lifetime of $\sim 10^{-7}$ s in order not to decay inside the detector. This lower bound on the lifetime is far away from the requested lifetime $\gg 10^{10}$ yr of the WIMP. Therefore, even though the LHC may point towards the correct physical nature of the WIMP candidate, only its direct detection can prove that it is indeed the dark matter particle. However, from the measured cross sections of other supersymmetric particles, one can determine the free parameters of the model (at least in its simplest version), and from that get an estimate of the WIMP annihilation cross section. Since this, in combination with the WIMP mass, determines its cosmological density, as explained above, Ω_X can be estimated in the laboratory! If this value agrees with $\Omega_\mathrm{dm} = (\Omega_\mathrm{m} - \Omega_\mathrm{b})$, then this neutral particle will be indeed an excellent candidate for the dark matter.

Indirect astrophysical detections. In its simplest form, we expect from supersymmetry that the WIMP is its own antiparticle, and thus two WIMPs can annihilate. That happened in the early Universe before the freeze-out of WIMPs, but since then became very rare. Nevertheless, in regions of high dark matter density, some annihilation may occur. The resulting signal depends on the kind of particles into which they annihilate, but in general one would expect that high-energy photons are generated in the decay chain, which may be visible in hard γ-radiation. The number density of annihilation events is proportional to the square of the WIMP density, and therefore the most promising places to look for these γ-rays are probably the centers of dark matter halos—in particular, the center of the Galaxy and that of nearby dwarf galaxies. Of course, the problem of distinguishing the annihilation signal from other γ-ray sources needs to be overcome.

Another indirect method is based on the fact that some WIMPs which cross the Earth or the Sun get scattered by atomic nuclei, thereby change their velocity, which may become lower than the escape velocity from these objects, and thus they are gravitationally captured. After that, they will orbit within the Sun (or the Earth), and due to the high density there, they will scatter again, and finally sink toward the center of the body. Therefore, the density of WIMPs can be strongly enhanced there, and correspondingly the rate of annihilations. The annihilation products will decay, or be stopped, in the body, except for neutrinos which escape. The signature of the annihilation are thus neutrinos, with an energy much higher that produced in nuclear fusion processes. Hence, such high-energy neutrino signals from the center of the Earth or the Sun would be a unique signature of WIMP annihilation. Existing neutrino detectors, such as IceCube in Antarctica, are beginning to probe interesting regions in the WIMP parameter space of mass and cross section.

4.4.7 Recombination

About 3 min after the Big Bang, BBN comes to an end. At this time, the Universe has a temperature of roughly $T \sim 8 \times 10^8$ K and consists of photons, protons, helium nuclei, traces of other light elements, and electrons. In addition, there are neutrinos that dominate, together with photons, the energy density and thus also the expansion rate, and

there are (probably) WIMPs. Except for the neutrinos and the WIMPs, all particle species have the same temperature, which is established by interactions of charged particles with the photons, which resemble some kind of heat bath.

At $z = z_{\rm eq} \approx 23\,900\,\Omega_{\rm m} h^2$, pressureless matter (i.e., the so-called dust) begins to dominate the cosmic energy density and thus the expansion rate. The second term in (4.33) then becomes largest, i.e., $H^2 \approx H_0^2\,\Omega_{\rm m}/a^3$. If a power-law ansatz for the scale factor, $a \propto t^\beta$, is inserted into the expansion equation, we find that $\beta = 2/3$, and hence

$$\boxed{a(t) = \left(\frac{3}{2}\sqrt{\Omega_{\rm m}}\,H_0\,t\right)^{2/3} \quad \text{for } a_{\rm eq} \ll a \ll 1}\,. \tag{4.70}$$

This describes the expansion behavior until either the curvature term or, if this is zero or very small, the Λ-term starts to dominate.

After further cooling, the free electrons can combine with the nuclei and form neutral atoms. This process is called *recombination*, although this expression is misleading: since the Universe was fully ionized until then, it is not a *re*combination but rather the (first) transition to a neutral state—however the expression 'recombination' has now long been established. The recombination of electrons and nuclei is in competition with the ionization of neutral atoms by energetic photons (photoionization), whereas collisional ionization can be disregarded completely since η—see (4.65)—is so small. Because photons are so much more numerous than electrons, cooling has to proceed to well below the ionization temperature, corresponding to the binding energy of an electron in hydrogen, before neutral atoms become abundant. This happens for the same reasons as apply in the context of deuterium formation: there are plenty of ionizing photons in the Wien tail of the Planck distribution, even if the temperature is well below the ionization temperature. The ionization energy of hydrogen is $\chi = 13.6\,\text{eV}$, corresponding to a temperature of $T > 10^5\,\text{K}$, but T has to first decrease to $\sim 3000\,\text{K}$ before the ionization fraction

$$x = \frac{\text{number density of free electrons}}{\text{total number density of existing protons}} \tag{4.71}$$

falls considerably below 1, for the reason mentioned above. At temperatures $T > 10^4\,\text{K}$ we have $x \approx 1$, i.e., virtually all electrons are free. Only below $z \sim 1300$ does x deviate significantly from unity.

The onset of recombination can be described by an equilibrium consideration which leads to the so-called Saha equation,

$$\frac{1-x}{x^2} \approx 3.84\,\eta\left(\frac{k_{\rm B}T}{m_{\rm e}c^2}\right)^{3/2} \exp\left(\frac{\chi}{k_{\rm B}T}\right)\,,$$

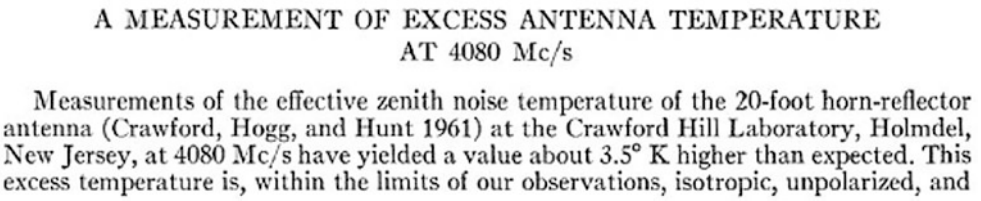
which describes the ionization fraction x as a function of temperature. However, once recombination occurs, the assumption of thermodynamical equilibrium is no longer justified. This can be seen as follows:

Any recombination directly to the ground state leads to the emission of a photon with energy $E_\gamma > \chi$. However, these photons can ionize other, already recombined (thus neutral), atoms. Because of the large cross section for photoionization, this happens very efficiently. Thus for each recombination to the ground state, one neutral atom will become ionized, yielding a vanishing net effect. But recombination can also happen in steps, first into an excited state and then evolving into the ground state by radiative transitions. Each of these recombinations will yield a Lyman-series photon in the transition from an excited state into the ground state. This Lyman photon will then immediately excite another atom from the ground state into an excited state, which has an ionization energy of $\leq \chi/4$. This yields no net production of atoms in the ground state. Since the density of photons with $E_\gamma > \chi/4$ is very much larger than of those of $E_\gamma > \chi$, the excited atoms are more easily ionized, and this indeed happens. Stepwise recombination thus also provides no route towards a lower ionization fraction.

The processes described above cause a small distortion of the Planck spectrum due to recombination radiation (in the range $\chi \gg k_{\rm B}T$) which affects recombination. One cannot get rid of these energetic photons—in contrast to gas nebulae like HII regions, in which the Lyα photons may escape due to the finite geometry.

Ultimately, recombination takes place by means of a very rare process, the two-photon decay of the first excited level. This process is less probable than the direct Lyα transition by a factor of $\sim 10^8$. However, it leads to the emission of two photons, neither of which is sufficiently energetic to excite an atom from the ground state. This 2γ-transition is therefore a net sink for energetic photons.[13] Taking into account all relevant processes and using a rate equation, which describes the evolution of the distribution of particles and photons even in the absence of thermodynamic equilibrium, gives for the ionization fraction in the relevant redshift range $800 \lesssim z \lesssim 1200$

[13] The recombination of hydrogen—and also that of helium which occurred at slightly higher redshifts—perturbed the exact Planck shape of the photon distribution, adding to it the Lyman-alpha photons and the photon pairs from the two-photon transition. This slight perturbation in the CMB spectrum should in principle still be present today. Unfortunately, it lies in a wavelength range ($\sim 200\,\mu\text{m}$) where the dust emission from the Galaxy is very strong; in addition, the wavelength range coincides with the peak of the far-infrared background radiation (see Sect. 9.5.1). Therefore, the detection of this spectral distortion will be extremely difficult.



A MEASUREMENT OF EXCESS ANTENNA TEMPERATURE AT 4080 Mc/s

Measurements of the effective zenith noise temperature of the 20-foot horn-reflector antenna (Crawford, Hogg, and Hunt 1961) at the Crawford Hill Laboratory, Holmdel, New Jersey, at 4080 Mc/s have yielded a value about 3.5° K higher than expected. This excess temperature is, within the limits of our observations, isotropic, unpolarized, and

Fig. 4.18 The first lines of the article by Penzias and Wilson (1965), ApJ 142, 419

$$\boxed{x(z) = 2.4 \times 10^{-3} \, \frac{\sqrt{\Omega_\mathrm{m} \, h^2}}{\Omega_\mathrm{b} \, h^2} \left(\frac{z}{1000} \right)^{12.75}} \, . \qquad (4.72)$$

The ionization fraction is thus a very strong function of redshift since x changes from 1 (complete ionization) to $x \sim 10^{-4}$ (where essentially all atoms are neutral) within a relatively small redshift range. The recombination process is not complete, however. A small ionization fraction of $x \sim 10^{-4}$ remains since the recombination rate for small x becomes smaller than the expansion rate—some nuclei do not find an electron fast enough before the density of the Universe becomes too low. From (4.72), the optical depth for Thomson scattering (scattering of photons by free electrons) can be computed (see problem 4.12),

$$\boxed{\tau(z) = 0.37 \left(\frac{z}{1000} \right)^{14.25}} \, , \qquad (4.73)$$

which is virtually independent of cosmological parameters. Equation (4.73) implies that photons can propagate from $z \sim 1000$ (the *'last-scattering surface'*) until the present day essentially without any interaction with matter—provided the wavelength is larger than 1216 Å. For photons of smaller wavelength, the absorption cross section of neutral atoms is large. Disregarding these highly energetic photons here—their energies are $\gtrsim 10$ eV, compared to $T_\mathrm{rec} \sim 0.3$ eV, so they are far out in the Wien tail of the Planck distribution—we conclude that the photons present after recombination have been able to propagate without further interactions until the present epoch. Before recombination they followed a Planck spectrum. As was discussed in Sect. 4.3.2, the distribution will remain a Planck spectrum with only its temperature changing. Thus these photons from the early Universe should still be observable today, redshifted into the microwave regime of the electromagnetic spectrum.

Our consideration of the early Universe predicts thermal radiation from the Big Bang, as was first realized by George Gamow in 1946—the cosmic microwave background. The CMB is therefore a visible relic of the Big Bang.

The CMB was detected in 1965 by Arno Penzias & Robert Wilson (see Fig. 4.18), who were awarded the 1978 Nobel prize in physics for this very important discovery. At the beginning of the 1990s, the COBE satellite measured the spectrum of the CMB with a very high precision—it is the most perfect blackbody ever measured (see Fig. 4.3). From upper bounds of deviations from the Planck spectrum, very tight limits for possible later energy injections into the photon gas, and thus on energetic processes in the Universe, can be obtained.[14]

We have only discussed the recombination of hydrogen. Since helium has a higher ionization energy it recombines earlier than hydrogen. Although recombination defines a rather sharp transition, (4.73) tells us that we receive photons from a recombination layer of finite thickness ($\Delta z \sim 60$). This aspect will be of importance later.

The gas in the intergalactic medium at lower redshift is highly ionized. If this were not the case we would not be able to observe any UV photons from sources at high redshift ('Gunn-Peterson-test', see Sect. 8.5.1). Sources with redshifts $z > 6$ have been observed, and we also observe photons with wavelengths shorter than the Lyα line of these objects. Thus at least at the epoch corresponding to redshift $z \sim 6$, the Universe must have been nearly fully ionized or else these photons would have been absorbed by photoionization of neutral hydrogen. This means that at some time between $z \sim 1000$ and $z \sim 6$, a reionization of the intergalactic medium must have occurred, presumably by a first generation of stars or by the first AGNs. The results from the new CMB satellites WMAP and Planck suggest a reionization at redshift $z \sim 10$; this will be discussed more thoroughly in Sect. 8.7.

[14] For instance, there exists a cosmic X-ray background (CXB; see Sect. 9.5) which is radiation that appeared isotropic in early measurements. For a long time, a possible explanation for this was suggested to be a hot intergalactic medium with temperature of $k_\mathrm{B} T \sim 40$ keV emitting bremsstrahlung radiation. But such a hot intergalactic gas would modify the spectrum of the CMB via the scattering of CMB photons to higher frequencies by energetic electrons (inverse Compton scattering). This explanation for the source of the CXB was excluded by the COBE measurements. From observations by the X-ray satellites ROSAT, Chandra, and XMM-Newton, with their high angular resolution, we know today that the CXB is a superposition of radiation from discrete sources, mostly AGNs.

4.4.8 Summary

We will summarize this somewhat long section as follows:

- Our Universe originated from a very dense, very hot state, the so-called *Big Bang*. Shortly afterwards, it consisted of a mix of various elementary particles, all interacting with each other.
- We are able to examine the history of the Universe in detail, starting at an early epoch where it cooled down by expansion such as to leave only those particle species known to us (electrons, protons, neutrons, neutrinos, and photons), and probably a dark matter particle.
- Because of their weak interaction and the decreasing density, the neutrinos experience only little interaction at temperatures below $\sim 10^{10}\,\mathrm{K}$, their decoupling temperature.
- At $T \sim 5 \times 10^9\,\mathrm{K}$, electrons and positrons annihilate into photons. At this low temperature, pair production ceases to take place.
- Protons and neutrons interact and form deuterium nuclei. As soon as $T \sim 10^9\,\mathrm{K}$, deuterium is no longer efficiently destroyed by energetic photons. Further nuclear reactions produce mainly helium nuclei. About 25 % of the mass in nucleons is transformed into helium, and traces of lithium are produced, but no heavier elements.
- At about $T \sim 3000\,\mathrm{K}$, some 400 000 years after the Big Bang, the protons and helium nuclei combine with the electrons, and the Universe becomes essentially neutral (we say that it 'recombines'). From then on, photons can travel without further interactions. At recombination, the photons follow a blackbody distribution (i.e., a thermal spectrum, or a Planck distribution). By the ongoing cosmic expansion, the temperature of the spectral distribution decreases, $T \propto (1+z)$, though its Planck property remains.
- After recombination, the matter in the Universe is almost completely neutral. However, we know from the observation of sources at very high redshift that the intergalactic medium is essentially fully ionized at $z \lesssim 6$. Before $z > 6$, our Universe must therefore have experienced a phase of reionization. This effect cannot be explained in the context of the *strictly homogeneous* world models; rather it must be examined in the context of structure formation in the Universe and the formation of the first stars and AGNs. These aspects will be discussed in Sect. 10.3.

4.5 Achievements and problems of the standard model

To conclude this chapter, we will evaluate the cosmological model which has been presented. We will review its achievements and successes, but also apparent problems, and point out the route by which those might be understood. As is always the case in natural sciences, problems with an otherwise very successful model are often the key to a new and deeper understanding.

4.5.1 Achievements

The standard model of the Friedmann–Lemaître universe described above has been extremely successful in numerous ways:

- It predicts that gas which has not been subject to much chemical processing (i.e., metal-poor gas) should have a helium content of $\sim 25\,\%$. This is in extraordinarily good agreement with observations.
- It predicts that sources of lower redshift are closer to us than sources of higher redshift.[15] Therefore, modulo any peculiar velocities, the absorption of radiation from sources at high redshift must happen at smaller redshifts. Not a single counter-example has been found yet.
- It predicts the existence of a microwave background, which indeed was found.
- It predicts the correct number of neutrino families, which was confirmed in laboratory experiments of the Z-boson decay.

Further achievements will be discussed in the context of structure evolution in the Universe.

A good physical model is one that can also be falsified. In this respect, the Friedmann–Lemaître universe is also an excellent model: a single observation could either cause a lot of trouble for this model or even disprove it. To wit, it would be incompatible with the model

1. if the helium content of a gas cloud or of a low-metallicity star was significantly below 25 %;
2. if it was found that one of the neutrinos has a rest mass $\gtrsim 100\,\mathrm{eV}$;
3. if the Wien-part of the CMB had a smaller amplitude compared to the Planck spectrum;
4. if a source with emission lines at z_e was found to show absorption lines at $z_\mathrm{a} \gg z_\mathrm{e}$;
5. if the cosmological parameters were such that $t_0 \lesssim 10\,\mathrm{Gyr}$.

On (1): While the helium content may increase by stellar evolution due to fusion of hydrogen into helium, only a small fraction of helium is burned in stars. In this process, heavier elements are of course produced. A gas cloud or a star with low metallicity therefore cannot consist of material in which helium has been destroyed; it must contain at least the helium abundance from BBN. On (2): Such a

[15] We ignore peculiar motions here which may cause an additional (Doppler-)redshift. These are typically $\lesssim 1000\,\mathrm{km/s}$ and are thus small compared to cosmological redshifts.

neutrino would lead to $\Omega_\mathrm{m} > 2$, which is in strict contradiction to the derived model parameters. On (3): Though it is possible to generate additional photons by energetic processes in the past, thereby increasing the Wien-part of the coadded spectrum compared to that of a Planck function, it is thermodynamically impossible to extract photons from the Wien-part. On (4): Such an observation would question the role of redshift as a monotonic measure of relative distances and thus remove one of the pillars of the model. On (5): Our knowledge of stellar evolution allows us to determine the age of the oldest stars with a precision of better than $\sim$ 20 %. An age of the Universe below $\sim$ 10 Gyr would be incompatible with the age of the globular clusters—naturally, these have to be younger than the age of our Universe, i.e., the time after the Big Bang.

Although these predictions have been known for more than 40 years, no observation has yet been made which disproves the standard model. Indeed, at any given time there have been astronomers who like to disagree with the standard model. These astronomers have tried to make a discovery, like the examples above, which would pose great difficulties for the model. So far without success; this does not mean that such results cannot be found in the literature, but rather such results did not withstand closer examination. The simple opportunities to falsify the model and the lack of any corresponding observation, together with the achievements listed above, have made the Friedmann–Lemaître model *the* standard model of cosmology. Alternative models have either been excluded by observation (such as steady-state cosmology) or have been unable to make any predictions. Currently, there is no serious alternative to the standard model.

4.5.2 Problems of the standard model

Despite these achievements, there are some aspects of the model which require further consideration. Here we will describe two conceptual problems with the standard model more thoroughly—the horizon problem and the flatness problem.

Horizons. The finite speed of light implies that we are only able to observe a finite part of the Universe, namely those regions from which light can reach us within a time t_0. Since $t_0 \approx 13.8$ Gyr, our visible Universe has—roughly speaking—a radius of 13.8 billion light years. More distant parts of the Universe are at the present time unobservable for us. This means that there exists a horizon beyond which we cannot see. Such horizons do not only exist for us today: at an earlier time t, the size of the horizon was about ct, hence smaller than today. We will now describe this aspect quantitatively.

In a time interval $\mathrm{d}t$, light travels a distance $c\,\mathrm{d}t$, which corresponds to a comoving distance interval $\mathrm{d}x = c\,\mathrm{d}t/a$ at scale factor a. From the Big Bang to a time t (or redshift z) the light traverses a comoving distance of

$$r_\mathrm{H,com}(z) = \int_0^t \frac{c\,\mathrm{d}t}{a(t)}\,.$$

From $\dot{a} = \mathrm{d}a/\mathrm{d}t$ we get $\mathrm{d}t = \mathrm{d}a/\dot{a} = \mathrm{d}a/(aH)$, so that

$$\boxed{r_\mathrm{H,com}(z) = \int_0^{(1+z)^{-1}} \frac{c\,\mathrm{d}a}{a^2\,H(a)}}\,. \tag{4.74}$$

If $z_\mathrm{eq} \gg z \gg 0$, the main contribution to the integral comes from times (or values of a) in which pressureless matter dominates the expansion rate H. Then with (4.33) we find $H(a) \approx H_0\sqrt{\Omega_\mathrm{m}}a^{-3/2}$, and (4.74) yields

$$\boxed{r_\mathrm{H,com}(z) \approx 2\frac{c}{H_0}\,\frac{1}{\sqrt{(1+z)\Omega_\mathrm{m}}} \quad \text{for} \quad z_\mathrm{eq} \gg z \gg 0}\,. \tag{4.75}$$

In earlier phases, $z \gg z_\mathrm{eq}$, H is radiation-dominated, $H(a) \approx H_0\sqrt{\Omega_\mathrm{r}}/a^2$, and (4.74) becomes

$$\boxed{r_\mathrm{H,com}(z) \approx \frac{c}{H_0\sqrt{\Omega_\mathrm{r}}}\,\frac{1}{(1+z)} \quad \text{for} \quad z \gg z_\mathrm{eq}}\,. \tag{4.76}$$

The earlier the cosmic epoch, the smaller the comoving horizon length, as was to be expected. In particular, we will now consider the recombination epoch, $z_\mathrm{rec} \sim 1000$, for which (4.75) applies (see Fig. 4.19). The comoving length $r_\mathrm{H,com}$ corresponds to a physical proper length $r_\mathrm{H,prop} = a\,r_\mathrm{H,com}$, and thus

$$r_\mathrm{H,prop}(z_\mathrm{rec}) = 2\frac{c}{H_0}\,\Omega_\mathrm{m}^{-1/2}\,(1+z_\mathrm{rec})^{-3/2} \tag{4.77}$$

is the horizon length at recombination. We can then calculate the angular size on the sky that this length corresponds to,

$$\theta_\mathrm{H,rec} = \frac{r_\mathrm{H,prop}(z_\mathrm{rec})}{D_\mathrm{A}(z_\mathrm{rec})}\,,$$

where D_A is the angular-diameter distance (4.49) to the last scattering surface of the CMB. Using (4.51), we find that in the case of $\Omega_\Lambda = 0$

$$D_\mathrm{A}(z) \approx \frac{c}{H_0}\,\frac{2}{\Omega_\mathrm{m} z} \quad \text{for} \quad z \gg 1\,,$$

and hence

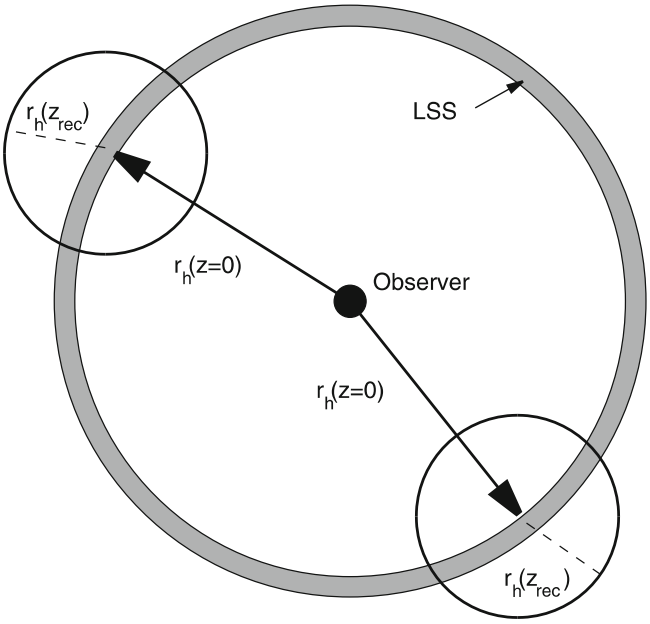


Fig. 4.19 The horizon problem: the region of space which was in causal contact before recombination has a much smaller radius than the spatial separation between two regions from which we receive the CMB photons. Thus the question arises how these two regions may 'know' of each other's temperature. Adapted from: Alan Guth 1998, The inflationary Universe, Basic Books

$$\boxed{\theta_{\mathrm{H,rec}} \approx \sqrt{\frac{\Omega_\mathrm{m}}{z_\mathrm{rec}}} \sim \frac{\sqrt{\Omega_\mathrm{m}}}{30} \sim \sqrt{\Omega_\mathrm{m}}\, 2^\circ \quad \text{for } \Omega_\Lambda = 0} \,. \tag{4.78}$$

This means that the horizon length at recombination subtends an angle of about 1° on the sky.

The horizon problem: Since no signal can travel faster than light, (4.78) means that CMB radiation from two directions separated by more than about one degree originates in regions that were not in causal contact before recombination, i.e., the time when the CMB photons interacted with matter the last time. Therefore, these two regions have never been able to exchange information, for example about their temperature. Nevertheless their temperature is the same, as seen from the high degree of isotropy of the CMB, which shows relative fluctuations of only $\Delta T/T \sim 10^{-5}$!

Redshift-dependent density parameter. We have defined the density parameters Ω_m and Ω_Λ as the current density divided by the critical mass density ρ_cr today. These definitions can be generalized. If we existed at a different time, the densities and the Hubble constant would have had different values and consequently we would obtain different values for the density parameters. Thus we define the total density parameter for an arbitrary redshift

$$\boxed{\Omega_0(z) = \frac{\rho_\mathrm{m}(z) + \rho_\mathrm{r}(z) + \rho_\mathrm{v}}{\rho_\mathrm{cr}(z)}} \,, \tag{4.79}$$

where the critical density ρ_cr is also a function of redshift,

$$\boxed{\rho_\mathrm{cr}(z) = \frac{3H^2(z)}{8\pi G}} \,. \tag{4.80}$$

Then by inserting (4.24) into (4.79), we find

$$\Omega_0(z) = \left(\frac{H_0}{H}\right)^2 \left(\frac{\Omega_\mathrm{m}}{a^3} + \frac{\Omega_\mathrm{r}}{a^4} + \Omega_\Lambda\right) \,.$$

Using (4.33), this yields

$$\boxed{1 - \Omega_0(z) = F\ [1 - \Omega_0(0)]} \,, \tag{4.81}$$

where $\Omega_0(0)$ is the total density parameter today, and

$$F = \left(\frac{H_0}{a\, H(a)}\right)^2 \,. \tag{4.82}$$

From (4.81) we can now draw two important conclusions. Since $F > 0$ for all a, the sign of $(\Omega_0 - 1)$ is preserved and thus is the same at all times as today. Since the sign of $(\Omega_0 - 1)$ is the same as that of the curvature—see (4.32)—the sign of the curvature is preserved in cosmic evolution: a flat Universe will be flat at all times, a closed Universe with $K > 0$ will always have a positive curvature.

The second conclusion follows from the analysis of the function F at early cosmic epochs, e.g., at $z \gg z_\mathrm{eq}$, thus in the radiation-dominated Universe. Back then, with (4.33), we have

$$F = \frac{1}{\Omega_\mathrm{r}(1+z)^2} \,,$$

so that for very early times, F becomes very small. For instance, at $z \sim 10^{10}$, the epoch of neutrino freeze-out, $F \sim 10^{-15}$. Today, Ω_0 is of order unity; from observations, we know that certainly $0.1 \lesssim \Omega_0(0) \lesssim 2$, where this is a *very* generous estimate,[16] so that $|1 - \Omega_0(0)| \lesssim 1$. Since F is so small at large redshifts, this means that $\Omega_0(z)$ must have been

[16] From the most recent CMB measurements (see Sect. 8.7) we are able to constrain this interval to better than [0.99,1.01].

very, very close to 1; for example at $z \sim 10^{10}$ it is required that $|\Omega_0 - 1| \lesssim 10^{-15}$.

Flatness problem: For the total density parameter to be of order unity today, it must have been extremely close to 1 at earlier times, which means that a very precise 'fine tuning' of this parameter was necessary.

This aspect can be illustrated very well by another physical example. If we throw an object up into the air, it takes several seconds until it falls back to the ground. The higher the initial velocity, the longer it takes to hit the ground. To increase the time of flight we need to increase the initial velocity, for instance by using a cannon. In this way, the time of flight may be extended to up to about a minute. Assume that we want the object to be back only after one day; in this case we must use a rocket. But we know that if the initial velocity of a rocket exceeds the escape velocity $v_{\mathrm{esc}} \sim 11.2\,\mathrm{km/s}$, it will leave the gravitational field of the Earth and never fall back. On the other hand, if the initial velocity is too much below v_{esc}, the object will be back in significantly less than a day. So the initial velocity must be *very* well chosen for the object to return after being up for at least a day. The flatness problem is completely analogous to this.

Let us consider the consequences of the case where Ω_0 had not been so extremely close to 1 at $z \sim 10^{10}$; then, the universe would have recollapsed long ago, or it would have expanded significantly faster than the universe we live in. In either case, the consequences for the evolution of life in the universe would have been catastrophic. In the first case, the total lifetime of the universe would have been much shorter than is needed for the formation of the first stars and the first planetary systems, so that in such a world no life could be formed. In the second case, extreme expansion would have prevented the formation of structure in the universe. In such a universe no life could have evolved either.

This consideration can be interpreted as follows: we live in a universe which had, at a very early time, a very precisely tuned density parameter, because only in such a universe can life evolve and astronomers exist to examine the flatness of the universe. In all other conceivable universes this would not be possible. This approach is meaningful only if a large number of universes existed—in this case we should not be too surprised about living in one of those where this initial fine-tuning took place—in the other ones, we, and the question about the cosmological parameters, would just not exist. This approach is called the *anthropic principle*. It may either be seen as an 'explanation' for the flatness of *our* Universe, or as a capitulation—where we give up to explore a physical reason for the origin of the flatness of our Universe.

The example of the rocket given above is helpful in understanding another aspect of cosmic expansion. If the rocket is supposed to have a long time of flight but not escape the gravitational field of the Earth, its initial velocity must be very, very close to, but a tiny little bit smaller than v_{esc}. In other words, the absolute value of the sum of kinetic and potential energy has to be very much smaller than either of these two components. This is also true for a large part of the initial trajectory. Independent of the exact value of the time of flight, the initial trajectory can be approximated by the limiting case $v_0 = v_{\mathrm{esc}}$ at which the total energy is exactly zero. Transferred to the Hubble expansion, this reads as follows: independent of the exact values of the cosmological parameters, the curvature term can be disregarded in the early phases of expansion (as we have already seen above). This is because our Universe can reach its current age only if at early times the modulus of potential and kinetic energy were nearly exactly equal, i.e., the curvature term in (4.14) must have been a lot smaller than the other two terms.

4.5.3 Extension of the standard model: inflation

We will consider the horizon and flatness problems from a different, more technical point of view. Einstein's field equations of GR, one solution of which has been described as our world model, are a system of coupled partial differential equations. As is always the case for differential equations, their solutions are determined by (1) the system of equations itself and (2) the initial conditions. If the initial conditions at, e.g., $t = 1\,\mathrm{s}$ were as they have been described, the two aforementioned problems would not exist. But why are the conditions at $t = 1\,\mathrm{s}$ such that they lead to a homogeneous, isotropic, (nearly) flat model? The set of homogeneous and isotropic solutions to the Einstein equation is of measure zero (i.e., nearly all solutions of the Einstein equation are not homogeneous and isotropic); thus these particular solutions are *very* special. Taking the line of reasoning that the initial conditions 'just happened to be so' is not satisfying because it does not explain anything. Besides the anthropic principle, the answer to this question can only be that processes must have taken place even earlier, due to known or as yet unknown physics, which have produced these 'initial conditions' at $t = 1\,\mathrm{s}$. The initial conditions of the normal Friedmann–Lemaître expansion thus have a physical origin. Cosmologists believe they have found such a physical reason: the inflationary model.

Inflation. In the early 1980s, a model was developed which was able to solve the flatness and horizon problems (and some others as well). As a motivation for this model, we first recall that the physical laws and properties of elementary particles are well known up to energies of $\sim 100\,\mathrm{GeV}$ because they were experimentally tested in particle accelerators. For

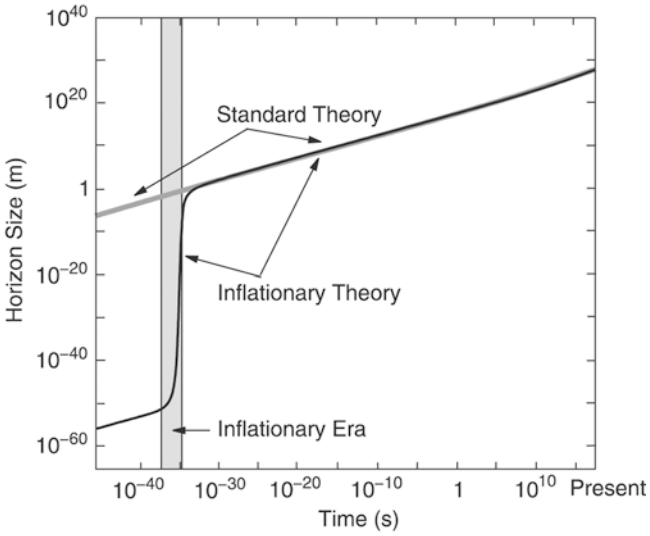


Fig. 4.20 During an inflationary phase, indicated here by the *gray bar*, the universe expands exponentially; see (4.83). This phase comes to an end when a phase transition transforms the vacuum energy into matter and radiation, after which the universe follows the normal Friedmann expansion. Adapted from: Alan Guth 1998, The inflationary Universe, Basic Books

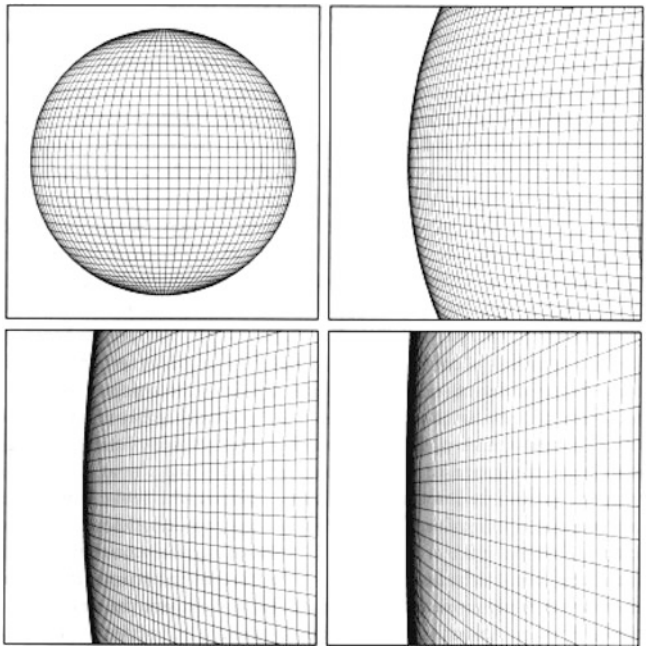


Fig. 4.21 Due to tremendous expansion during inflation, even a universe with initial curvature will appear to be a flat universe by the end of the inflationary phase. Source: A.H. Guth 1998, *The Inflationary Universe*, Basic Books

higher energies, particles and their interactions are unknown. This means that the history of the Universe, as sketched above, can be considered secure only up to energies of 100 GeV. The extrapolation to earlier times, up to the Big Bang, is considerably less certain. From particle physics we expect new phenomena to occur at an energy scale of the Grand Unified Theories (GUTs), at about 10^{14} GeV, corresponding to $t \sim 10^{-34}$ s.

In the inflationary scenario it is presumed that at very early times the vacuum energy density was much higher than today, so that it dominated the Hubble expansion. Then from (4.18) we find that $\dot{a}/a \approx \sqrt{\Lambda/3}$. This implies an exponential expansion of the Universe,

$$a(t) = C \, \exp\left(\sqrt{\frac{\Lambda}{3}}\, t\right) \, . \tag{4.83}$$

Obviously, this exponential expansion (or inflationary phase) cannot last forever. We assume that a phase transition took place in which the vacuum energy density is transformed into normal matter and radiation (a process called reheating), which ends the exponential expansion and after which the normal Friedmann evolution of the Universe begins. Figure 4.20 sketches the expansion history of the universe in an inflationary model.

Inflation solves the horizon problem. During inflation, $H(a) = \sqrt{\Lambda/3}$ is constant so that the integral (4.74) for the comoving horizon length formally diverges. This implies that the horizon may become arbitrarily large in the inflationary phase, depending on the duration of the exponential expansion. For illustration we consider a very small region in space of size $L < ct_\mathrm{i}$ at a time $t_\mathrm{i} \sim 10^{-34}$ s prior to inflation which is in causal contact. Through inflation, it expands tremendously, e.g., by a factor $\sim 10^{40}$; the original $L \sim 10^{-24}$ cm inflate to about 10^{16} cm by the end of the inflationary phase, at $t_\mathrm{f} \sim 10^{-32}$ s. By today, this spatial region will have expanded by another factor of $\sim 10^{25}$ by following (for $t > t_\mathrm{f}$) the normal cosmic expansion, to $\sim 10^{41}$ cm. This scale is considerably larger than the size of the currently visible Universe, c/H_0. According to this scenario, the whole Universe visible today was in causal contact prior to inflation, so that the homogeneity of the physical conditions at recombination, and with it the nearly perfect isotropy of the CMB, is provided by causal processes.

Inflation solves the flatness problem as well. Due to the tremendous expansion, any initial curvature is straightened out (see Fig. 4.21). Formally this can be seen as follows: during the inflationary phase we have

$$\Omega_\Lambda = \frac{\Lambda}{3H^2} = 1 \, ,$$

and since it is assumed that the inflationary phase lasts long enough for the vacuum energy to be completely dominant, when it ends we then have $\Omega_0 = 1$. Hence the universe is flat to an extremely good approximation.

The inflationary model of the very early universe predicts that today $\Omega_0 = 1$ is valid to very high precision; any other value of Ω_0 would require another fine-tuning. Thus our Universe is expected to be flat.

The physical details of the inflationary scenario are not very well known. In particular it is not yet understood how the phase transition at the end of the inflationary phase took place and why it did not occur earlier. But the two achievements presented above (and some others) make an inflationary phase appear a very plausible scenario. As we will see below (Chap. 8), the prediction of a flat universe was recently accurately tested and it was indeed confirmed. Furthermore, the inflationary model provides a natural, and in fact the only plausible, explanation for the origin of density fluctuations in the Universe which must have been present at very early epochs as the seeds of structure formation. We will discuss these aspects further in Chap. 7.

4.6 Problems

4.1. Big Bang Nucleosynthesis.

1. Calculate the baryon density at the epoch of nucleosynthesis. How does it compare with the density in the central regions of stars where nuclear burning takes place?
2. It takes the Sun some ten billion years to convert $\sim 10\,\%$ of its hydrogen into helium, whereas in BBN, all helium is formed on a time-scale of a minute. Can you speculate about the reasons for this difference?
3. During BBN, energy is released from the fusion process. Obtain an estimate of this fusion energy generated per unit volume, and compare it to the energy density of the photons at the epoch of BBN. Does BBN cause a substantial heating of the Universe?

4.2. Deceleration parameter. Assume that the energy density of the universe is composed of N different species. Each of these species is characterized by a density parameter Ω_i and an equation-of-state parameter w_i which describes the relation between pressure and density, $P_i = w_i \rho_i c^2$. Calculate the deceleration parameter q_0 for this cosmological model. By specializing to the three energy components discussed in this chapter, can you rederive (4.35)?

4.3. The qualitative behavior of the cosmic expansion. The general discussion of the qualitative behavior of the solutions of the Friedmann equation (4.33) is tedious, but some special results can be derived quite easily. In the following, neglect the (very small) contribution from Ω_r.

1. For $\Omega_\Lambda = 0$, show that the universe has been expanding for all $0 < a \leq 1$, that it will continue to expand forever in the future if $\Omega_\mathrm{m} \leq 1$, and that it reaches a maximum expansion at t_max, corresponding to the maximum scale factor a_max. Calculate a_max as a function of Ω_m.
2. Show that the universe expands forever in the future if $\Omega_\Lambda \geq 0$ and $\Omega_\mathrm{m} \leq 1$, and that it has been expanding in the past if $0 \leq \Omega_\Lambda < 1$, irrespective of Ω_m.
3. For the case of a closed universe which has a maximum expansion factor, or that of a bouncing model where a minimum scale factor occurs, show that $a(t_\mathrm{ex} - t) = a(t_\mathrm{ex} + t)$, where t_ex is the time where the extremum of the scale factor is attained.

4.4. Expansion law in a flat universe. You will now solve the Friedmann equation (4.33) for the case of vanishing curvature and vanishing radiation density. This describes the model of the Universe we live in, for scale factors $a \gg a_\mathrm{eq}$.

1. As a first step, write $a = v^\beta$, and choose β such that the Friedmann equation can be brought to the form $\dot{v}^2 = A + Bv^2$.
2. Then make the ansatz $v(t) = v_0 \sinh(t/t_\mathrm{a})$, and determine v_0 and t_a such that the foregoing equation is solved. Note that $\sinh'(x) = \cosh(x)$ and $\cosh^2(x) = 1 + \sinh^2(x)$.
3. Combining these two steps, write the full solution $a(t)$ explicitly. What is the behavior of the solution for $t \ll t_\mathrm{a}$ and for $t \gg t_\mathrm{a}$—does it agree with your expectations? Does this solution describe the transition from a decelerating expansion to an accelerated one?

4.5. The onset of inflation. Solve the Friedmann equation for a flat universe with vanishing matter density—the situation perhaps approximating the situation in our Universe before the end of inflation. Use the same steps as in the previous problem to obtain the solution. Again, there is a characteristic time scale t_a occurring in the solution; what is the behavior of the scale factor for $t \ll t_\mathrm{a}$ and for $t \gg t_\mathrm{a}$? Does this correspond to what is written in the main text for radiation dominance and vacuum energy dominance, respectively?

4.6. Distances in cosmology. In Sect. 4.3.3, we quoted the expressions for the angular-diameter distance as a function of redshift; these expressions shall be derived here.

1. Consider a radial light ray reaching us today, i.e., at scale factor $a = 1$. From the relation (4.39), derive the relation between a small interval $\mathrm{d}a$ and a comoving radial distance interval $\mathrm{d}x$ along this ray.
2. Using this result, show that the comoving distance as a function of redshift is given by (4.53).
3. For a flat universe with $K = 0$, show the angular-diameter distance is given by (4.54).

4.7. Expansion law in an $\Omega_\Lambda = 0$ universe. For a model with vanishing vacuum density, the expansion law can be obtained analytically.

1. As preparation for the solution, we consider a differential equation of the form

$$\left(\frac{\mathrm{d}f}{\mathrm{d}t}\right)^2 = \frac{C}{f} - K \,, \tag{4.84}$$

where $C > 0$ and K are constants. Solutions of (4.84) are given in parametric form. Show by insertion that the solution with $f(t_1) = 0$ reads

$$f(\theta) = \frac{C}{2K}(1-\cos\theta)\,, \quad t(\theta) = t_1 + \frac{C}{2K^{3/2}}(\theta - \sin\theta) \tag{4.85}$$

for $K > 0$ and $0 \le \theta \le 2\pi$, and

$$f(\theta) = \frac{C}{2|K|}(\cosh\theta - 1)\,,$$
$$t(\theta) = t_1 + \frac{C}{2|K|^{3/2}}(\sinh\theta - \theta) \tag{4.86}$$

for $K < 0$ and $\theta \ge 0$. Note that $\mathrm{d}f/\mathrm{d}t = (\mathrm{d}f/\mathrm{d}\theta)(\mathrm{d}t/\mathrm{d}\theta)^{-1}$. In the special case of $K = 0$, show that the solution is

$$f(t) = \left(\frac{9C}{4}\right)^{1/3}(t-t_1)^{2/3}\,. \tag{4.87}$$

For the case of $K > 0$, show that f reaches a maximum value $f_{\max} = C/K$ at time $t_{\max} = t_1 + \pi C K^{-3/2}/2$, and that $f_{\mathrm{coll}} = 0$ at time $t_{\mathrm{coll}} = t_1 + \pi C K^{-3/2}$.

2. Show that the Friedmann equation (4.33) in the case of $\Omega_\Lambda = 0 = \Omega_\mathrm{r}$ is of the form (4.84), and derive the expansion law in parametric form. Does the maximum scale factor $a_{\max}$ that occurs for $\Omega_\mathrm{m} > 1$ agree with what you found in problem 4.3? At what cosmic time does this maximum scale factor occur? When does such a universe recollapse?
3. As in problem 1.4, consider a sphere of mass M and initial radius r_0 at time $t = 0$, collapsing due to gravity. Show that the equation of motion for $r(t)$ can be written in the form (4.84), and determine the physical meaning of C and K. Show that the solution derived in problem 1.4 corresponds to the case $K = 0$. If the sphere at $t = 0$ was at rest, $\dot r(0) = 0$, show that the sphere collapses to point within the *free-fall time*

$$t_{\mathrm{ff}} = \sqrt{\frac{3\pi}{32G\bar\rho}}\,, \tag{4.88}$$

where $\bar\rho$ is the initial mean density of the sphere.

4.8. The time of return for a upward-moving object. In the text, an analogon of a nearly flat universe has been given, namely that of an object shot vertically upwards from the surface of the Earth. The time at which it returns to the surface is either 'short', or the velocity has to be very well fine-tuned. Using the parametric solution of the equation of motion derived in the previous exercise, we can now consider this situation quantitatively.

1. Show that the equation of motion for the object can be written in the form $\dot r^2 = 2GM_\mathrm{E}/r - K$, where M_E is the mass of the Earth. Relate the integration constant K to the initial velocity v_0 of the object, and assume in the following that $v_0 < v_{\mathrm{esc}}$, where $v_{\mathrm{esc}} = \sqrt{2GM_\mathrm{E}/r_\mathrm{E}} \approx 11.2\,\mathrm{km/s}$, with $r_\mathrm{E} \approx 6380\,\mathrm{km}$ being the Earth's radius.
2. From the parametric solution of the last problem, calculate the time t_{ret} at which the object returns to the Earth surface. For this, you can assume that the time-of-flight is 'long', i.e., much longer than r_E/v_0. Then find the relation between K and the time t_{ret}.
3. Combining the last two steps, obtain the relation between the initial velocity v_0 and the time of return t_{ret}. What fraction of the escape velocity does the object have to have initially if it should return after 1 day (1 year)?

4.9. Baryon cooling in the Universe. Suppose that at some epoch after recombination, the baryons are fully decoupled from the photons, so that there is no energy transfer from one species to the other. Use (4.47) to derive the expected redshift dependence of the baryon temperature during this epoch.

4.10. Thermal velocity of the cosmic neutrino background. Using (4.47), calculate the current characteristic velocity of neutrinos that decoupled in the early phase of the Big Bang.

4.11. Some properties of the Einstein–de Sitter model. Consider the Einstein–de Sitter model.

1. Calculate the look-back time $\tau(z)$. At what redshift was the age of the Universe half of its current age?
2. What is the volume of the spherical shell between redshifts z and $z + \Delta z$?
3. Assume the comoving density n_{com} of a class of cosmic objects is constant; how many of these are contained in sphere around us with maximum redshift z? Check that your result agrees with the expected one for $z \ll 1$.

4.12. The dependence of BBN on $\Omega_\mathrm{b}h^2$. The expansion law (4.61) yields the cosmic time vs. temperature. Why does this relation not depend on the Hubble constant? Why does the helium yield Y depend on the combination $\Omega_\mathrm{b}h^2$—see (4.68)—and not just on Ω_b?

4.13. Recombination optical depth. Using (4.72), show that the optical depth to Thomson scattering is almost independent of the cosmological parameters, as given in (4.73).

Active galactic nuclei

5

The light of normal galaxies in the optical and near infrared part of the spectrum is dominated by stars, with small contributions by gas and dust. This is thermal radiation since the emitting plasma in stellar atmospheres is basically in thermodynamical equilibrium. To a first approximation, the spectral properties of a star can be described by a Planck spectrum whose temperature depends on the stellar mass and the evolutionary state of the star. As we have seen in Sect. 3.5, the spectrum of galaxies can be described quite well as a superposition of stellar spectra. The temperature of stars varies over a relatively narrow range: Only few stars are found with $T \gtrsim 40\,000\,\mathrm{K}$, and those with $T \lesssim 3000\,\mathrm{K}$ hardly contribute to the spectrum of a galaxy, due to their low luminosity. Therefore, as a rough approximation, the light distribution of a galaxy can be described by a superposition of Planck spectra from a temperature range that covers about one decade. Since the Planck spectrum has a very narrow energy distribution around its maximum at $h_\mathrm{P}\nu \sim 3k_\mathrm{B}T$, the spectrum of a galaxy is basically confined to a range between $\sim$ 4000 and $\sim$ 20 000 Å. If the galaxy is actively forming stars, young hot stars extend this frequency range to somewhat higher frequency, and the thermal radiation from dust, heated by these new-born stars, extends the emission to the far-infrared.

However, there are galaxies which have a much broader energy distribution. Some of these show significant emission in the full range from radio wavelengths to the X-ray and even gamma range (see Fig. 3.4). This emission originates mainly from a very small central region of such an *active galaxy* which is called the *active galactic nucleus (AGN)*. This small emission region is structured and consists of multiple components with different physical properties, as we will see below. Active galaxies form a family of many different types of AGNs which differ in their spectral properties, including a wide range of ratios of radio-to-optical emission strength, their total luminosities and their ratio of nuclear luminosity to that of the stellar light. The optical spectra of three AGNs are presented in Fig. 5.1.

Some classes of AGNs, in particular the quasars, belong to the most luminous sources in the Universe, and they have been observed out to the highest measured redshifts ($z \sim 7$). The luminosity of quasars can exceed the luminosity of normal galaxies by a factor of a thousand. This luminosity originates from a tiny region in space, $r \lesssim 1\,\mathrm{pc}$. The optical/UV spectra of quasars are dominated by numerous strong and very broad emission lines, some of them emitted by highly ionized atoms (see Figs. 5.2 and 5.3). The processes in AGNs are among the most energetic ones in astrophysics. The enormous bandwidth of AGN spectra suggests that the radiation is non-thermal, i.e., not a superposition of (approximately) thermally radiating sources. As we will discuss later, processes in AGNs can produce highly energetic particles which are the origin of the non-thermal radiation.

After an introduction in which we will briefly present the history of the discovery of AGNs and their basic properties, in Sect. 5.2 we will describe the most important subgroups of the AGN family. In Sect. 5.3, we will discuss several arguments which lead to the conclusion that the energy source of an AGN originates in accretion of matter onto a supermassive black hole (SMBH). In particular, we will learn about the phenomenon of superluminal motion, where *apparent* velocities of source components are larger than the speed of light. We will then consider the different components of an AGN where radiation in different wavelength regions is produced.

Of particular importance for understanding the phenomenon of active galaxies are the unified models of AGNs that will be discussed next. We will see that the seemingly quite different appearances of AGNs can all be explained by geometric or projection effects. Finally, we will consider AGNs as cosmological probes. Due to their enormous luminosity they are observable up to very high redshifts.

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_5,

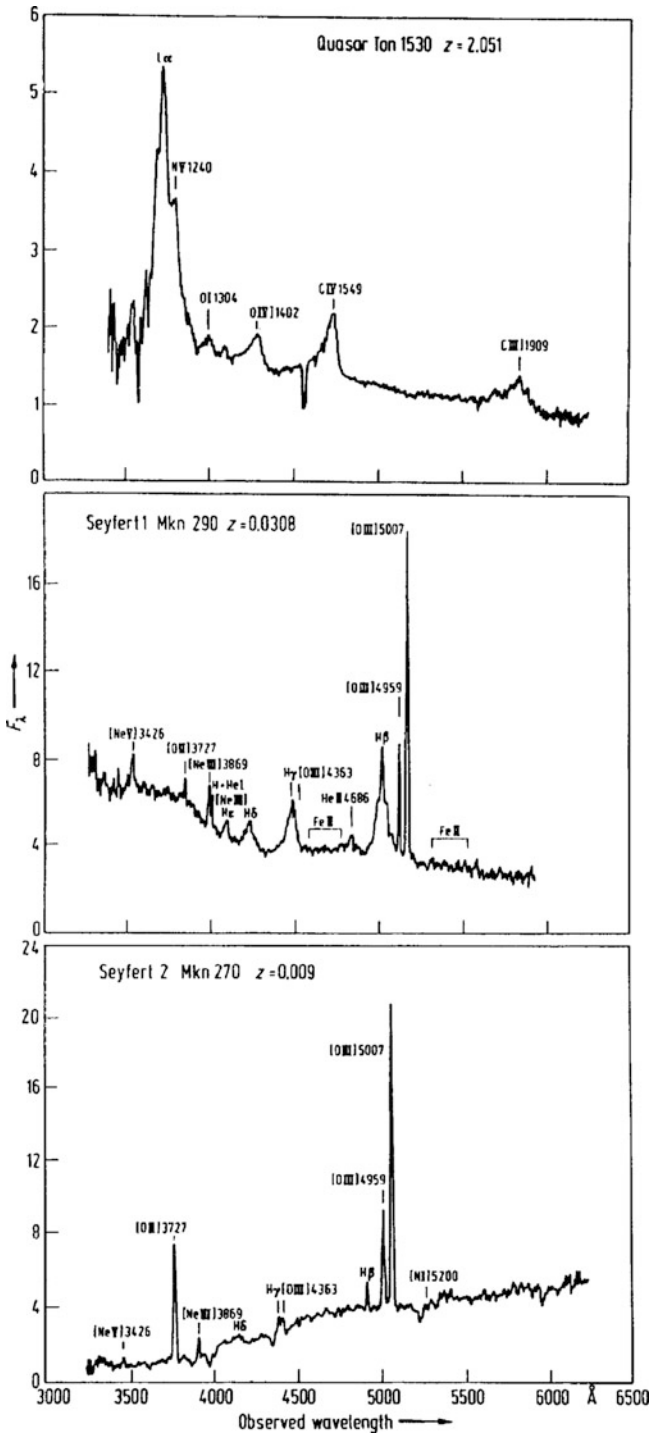


Fig. 5.1 Optical spectra of three AGNs. The *top panel* displays the spectrum of a quasar at redshift $z \sim 2$, which shows the characteristic broad emission lines. The strongest are Lyα of hydrogen, and the CIV-line and CIII]-line of triple and double ionized carbon, respectively (where the *squared bracket* means that this is a semi-forbidden transition, as will be explained in Sect. 5.4.2). The *middle panel* shows the spectrum of a nearby Seyfert galaxy of Type 1. Here both very broad emission lines and narrow lines, in particular of double ionized oxygen, are visible. In contrast, the spectrum in the *bottom panel*, of a Seyfert galaxy of Type 2, shows only relatively narrow emission lines. Source: H. Netzer 1990, in "Active Galactic Nuclei", eds. R.D. Blandford, H. Hetzer, L. Woltjer, Springer-Verlag 1990

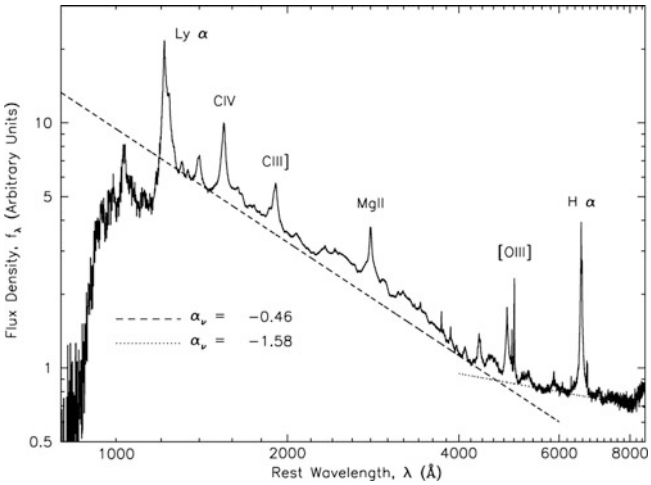


Fig. 5.2 The combined optical spectrum of more than 2200 QSOs from the SDSS. This 'mean' spectrum has a considerably better signal-to-noise ratio and a larger wavelength coverage than individual spectra. It was combined from the individual quasar spectra by transforming their wavelengths into the sources' rest-frames. The most prominent lines are marked. The *dashed* and *dotted lines* show power-law fits of the estimated underlying continuum emission of the QSOs. Source: D.E. Vanden Berk et al. 2001, *Composite Quasar Spectra from the Sloan Digital Sky Survey*, AJ 122, 549, p. 553, Fig. 3. ©AAS. Reproduced with permission

These observations allow us to draw conclusions about the properties of the early Universe.

5.1 Introduction

5.1.1 Brief history of AGNs

As long ago as 1908, strong emission lines were discovered in the galaxy NGC 1068 (see Fig. 5.4), which showed a line width of up to $\sim$ 3000 km/s. However, only the systematic analysis by Carl Seyfert in 1943 drew the focus of astronomers to this new class of galaxies. The cores of these *Seyfert galaxies* have an extremely high surface brightness, as demonstrated in Fig. 5.5, and the spectrum of their central region is dominated by emission lines of very high excitation. Some of these lines are extremely broad (see Fig. 5.1). The line width, when interpreted as Doppler broadening, $\Delta\lambda/\lambda = \Delta v/c$, yields values of up to $\Delta v \sim$ 8500 km/s for the full line width. The high excitation energy of some of the line-emitting atoms shows that they must have been excited by photons that are more energetic than photons from young stars that are responsible for the ionization of HII-regions. The hydrogen lines are often broader than other spectral lines. Most of the Seyfert galaxies are spirals, but one cD galaxy is also found in his original catalog.

In 1959, Lodewijk Woltjer argued that the extent of the cores of Seyfert galaxies cannot be larger than $r \lesssim 100$ pc because they appear point-like on optical images, i.e., they

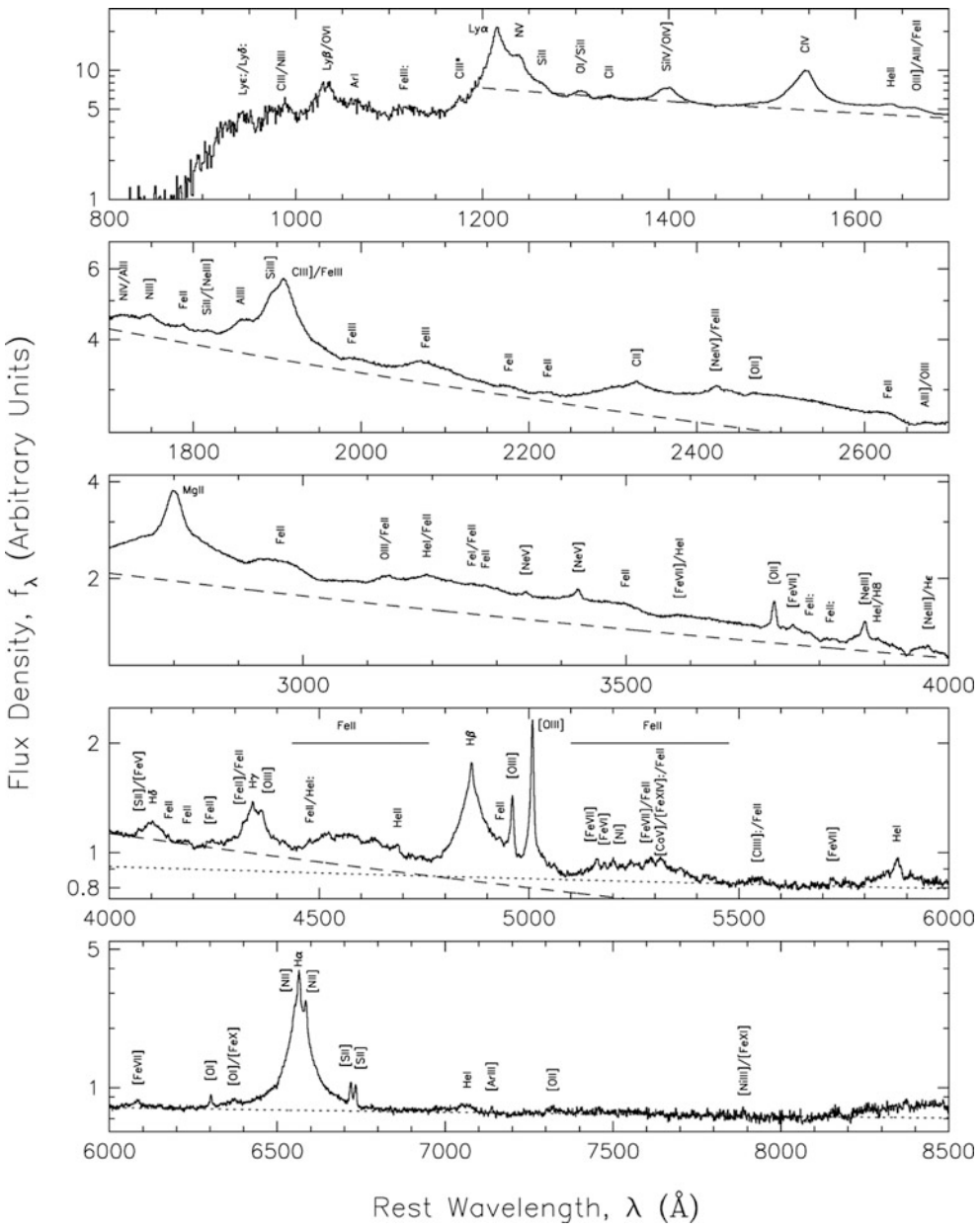


Fig. 5.3 An enlargement of the composite QSO spectrum shown in Fig. 5.2. Here, weaker spectral features are also visible. Also clearly seen is the break in the spectral flux bluewards of the Lyα line which is caused by the Lyα forest (Sect. 5.7), absorption by intergalactic hydrogen along the line-of-sight. The *dashed* and *dotted lines* indicate the average continuum. The substantial deviation of the spectrum from its estimated underlying continuum between 1600 and 3800 Å, even in spectral regions without obvious strong emission lines, is due to such a large number of overlapping iron lines that they blend into a quasi-continuum, and Balmer continuum (i.e., free-bound) radiation. Source: D.E. Vanden Berk et al. 2001, *Composite Quasar Spectra from the Sloan Digital Sky Survey*, AJ 122, 549, p. 555, Fig. 6. ©AAS. Reproduced with permission

are spatially not resolved. If the line-emitting gas is gravitationally bound, the relation

$$\frac{GM}{r} \simeq v^2$$

between the central mass $M(< r)$, the separation r of the gas from the center, and the typical velocity v must be satisfied. The latter is obtained from the line width: typically $v \sim 1000\,\mathrm{km/s}$. Therefore, with $r \lesssim 100\,\mathrm{pc}$, a mass estimate is immediately obtained,

$$M \gtrsim 10^{10} \left(\frac{r}{100\,\mathrm{pc}}\right) M_\odot \,. \tag{5.1}$$

Thus, either $r \sim 100\,\mathrm{pc}$, which implies an enormous mass concentration in the center of these galaxies, or r is much smaller than the estimated upper limit, which then implies an enormous energy density inside AGNs.

An important milestone in the history of AGNs was made with the 3C and 3CR radio catalogs which were completed around 1960. These are surveys of the northern ($\delta > -22°$) sky at 158 and 178 MHz, with a flux limit of $S_{\min} = 9\,\mathrm{Jy}$ (a Jansky is the flux unit used by radio astronomers, where $1\,\mathrm{Jy} = 10^{-23}\,\mathrm{erg\,s^{-1}\,cm^{-2}\,Hz^{-1}}$). Many of these 3C sources could be identified with relatively nearby galaxies, but the low angular resolution of radio telescopes at these low frequencies and the resulting large positional uncertainty of the respective sources rendered the identification with

Fig. 5.4 Optical image of the Seyfert galaxy NGC 1068, obtained with the Hubble Space Telescope. This spiral galaxy, located at a distance of ~ 15 Mpc from us, is the prototype of the Type 2 Seyfert galaxies. Its active nucleus is seen as the intense, high surface brightness center; it is powered by accretion onto a $\sim 15 \times 10^6\, M_\odot$ central supermassive black hole. Credit: NASA, ESA & A. van der Hoeven

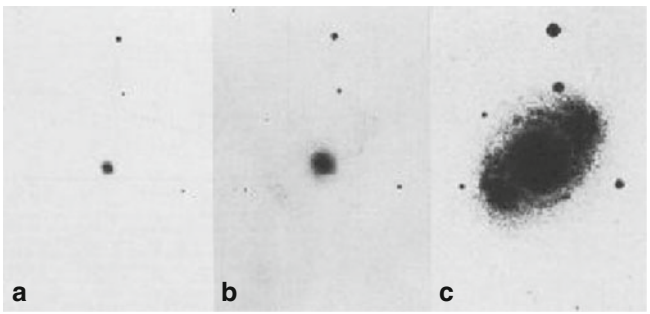


Fig. 5.5 Three images of the Seyfert galaxy NGC 4151, with the exposure time increasing to the right. In short exposures, the source appears point-like, with longer exposures displaying the galaxy. Source: W.W. Morgan 1968, *A Comparison of the Optical Forms of Certain Seyfert Galaxies with the N-Type Radio Galaxies*, ApJ 153, 27, PLATE 2, Fig. 2. ©AAS. Reproduced with permission

optical counterparts very difficult. If no striking nearby galaxy was found on optical photoplates within the positional uncertainty, the source was at first marked as unidentified.[1]

[1]The complete optical identification of the 3CR catalog, which was made possible by the enormously increased angular resolution of interferometric radio observations and thus by a considerably improved positional accuracy, was finalized only in the 1990s—some of these luminous radio sources are very faint optically.

In 1963, Thomas Matthews and Allan Sandage showed that 3C48 is a point-like ('stellar-like') source of $m = 16$ mag. It has a complex optical spectrum consisting of a blue continuum and strong, broad emission lines which could not be assigned to any atomic transition, and thus could not be identified. In the same year, Maarten Schmidt succeeded in identifying the radio source 3C273 with a point-like optical source which also showed strong and broad emission lines at unusual wavelengths. This was achieved by a lunar eclipse: the Moon passed in front of the radio source and eclipsed it. From the exact measurement of the time when the radio emission was blocked and became visible again, the position of the radio source was pinned down accurately. Schmidt could identify the emission lines of the source with those of the Balmer series of hydrogen, but at, for that time, an extremely high redshift of $z = 0.158$. Presuming the validity of the Hubble law and interpreting the redshift as cosmological redshift, 3C273 is located at the large distance of $D \sim 500h^{-1}$ Mpc. This huge distance of the source then implies an absolute magnitude of $M_B = -25.3 + 5 \log h$, i.e., it is about ~ 100 times brighter than normal (spiral) galaxies. Since the optical source had not been resolved but appeared point-like, this enormous luminosity must originate from a small spatial region. With the improving determination of radio source positions, many such *quasars* (quasi-stellar

radio sources = quasars) were identified in quick succession, the redshifts of some being significantly higher than that of 3C273.

5.1.2 Fundamental properties of quasars

In the following, we will review some of the most important properties of quasars. Although quasars are not the only class of AGNs, we will at first concentrate on them because they incorporate most of the properties of the other types of AGNs.

As already mentioned, quasars were discovered by identifying radio sources with point-like optical sources. Quasars emit at all wavelengths, from the radio to the X-ray, or even gamma-ray domain of the spectrum; see Fig. 3.4 for a sketch of the broad-band energy distribution of the quasar 3C 273. The radiation in the different frequency bands comes from various source components, as will be explained in the course of this chapter.

Interestingly, the flux of the source varies at nearly all frequencies, where the variability time-scale differs among the objects and also depends on the wavelength. As a rule, it is found that the variability time-scale of the observed radiation is smaller, and its amplitude larger, at higher frequencies. The optical spectrum is very blue; most quasars at redshifts $z \lesssim 2$ have $U - B < -0.3$ (for comparison: only hot white dwarfs have a similarly blue color index). Besides this blue continuum, very broad emission lines are characteristic of the optical and UV spectrum. Some of them correspond to transitions of very high ionization energy (see Fig. 5.3).

The continuum spectrum of a quasar can often be described, over a broad frequency range, by a power law of the form

$$S_\nu \propto \nu^{-\alpha} \,, \tag{5.2}$$

where α is the spectral index. $\alpha = 0$ corresponds to a flat spectrum, whereas $\alpha = 1$ describes a spectrum in which the same energy is emitted in every logarithmic frequency interval. Incidentally, the energy distribution of 3C 273 in Fig. 3.4 corresponds approximately to the latter case, over more than ten orders of magnitude in frequency, although over smaller frequency ranges, the spectral shape differs markedly from $\alpha = 1$.

5.1.3 AGNs as radio sources: synchrotron radiation

The morphology of quasars and other AGNs in the radio regime depends on the observed frequency and can often be very complex, consisting of several extended source components and one compact central one. In most cases, the extended component is observed as a double source in the form of two radio lobes situated more or less symmetrically around the optical position of the quasar. These lobes are frequently connected to the central core by jets, which are thin emission structures probably related to the energy transport from the core into the lobes (see Fig. 5.6 for an example). The observed length-scales are often impressive, in that the total extent of the radio source can reach values of up to 1 Mpc. The position of the optical quasar coincides with the compact radio source, which has an angular extent of $\ll 1''$ and is in some cases not resolvable even with VLBI methods. Thus the extent of these sources is $\lesssim 1$ mas, corresponding to $r \lesssim 1$ pc. This dynamical range in the extent of quasars is thus extremely large.

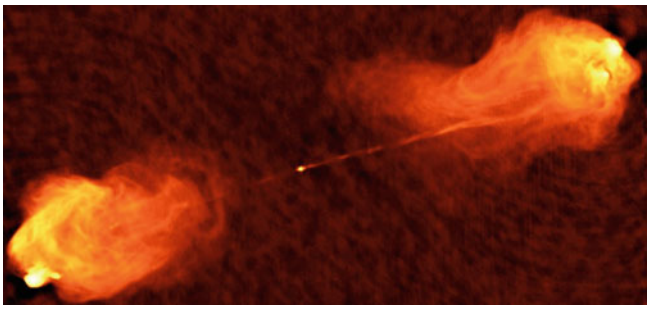


Fig. 5.6 The radio galaxy Cygnus A. Discovered by Grote Reber in 1939, it was the first very luminous active galaxy known. Cgynus A is one of the brightest radio sources in the sky, partly due to its low redshift ($z = 0.056$). This 6 cm map, covering a field of $2'.3 \times 1'.3$, impressively shows the narrow jet and counter-jet, and the two radio lobes on either side of the nucleus, in which the hot spots are embedded. Note how thin, straight, and highly collimated the jets are. Credit: Image courtesy of NRAO/AUI; Investigators: R. Perley, C. Carilli & J. Dreher

Classification of radio sources. Extended radio sources are often divided into two classes. *Fanaroff–Riley Type I* (FR I) are brightest close to the core, and the surface brightness decreases outwards. They typically have a luminosity of $L_\nu(1.4\,\mathrm{GHz}) \lesssim 10^{32}\,\mathrm{erg\,s^{-1}\,Hz^{-1}}$. In contrast, the surface brightness of *Fanaroff–Riley Type II* sources (FR II) increases outwards, and their luminosity is in general higher than that of FR I sources, $L_\nu(1.4\,\mathrm{GHz}) \gtrsim 10^{32}\,\mathrm{erg\,s^{-1}\,Hz^{-1}}$. One example for each of the two classes is shown in Fig. 5.7. FR II radio sources often have *jets*; they are extended linear structures that connect the compact core with a radio lobe. Jets often show internal structure such as knots and kinks. Their appearance indicates that they transport energy from the core out into the radio lobe. One of the most impressive examples of this is displayed in Fig. 5.8.

The jets are not symmetric. Often only one jet is observed, and in most sources where two jets are found one of them (the 'counter-jet') is much weaker than the other. The relative intensity of core, jet, and extended components varies with frequency, for sources as a whole and also within a source, because the components have different spectral indices. For

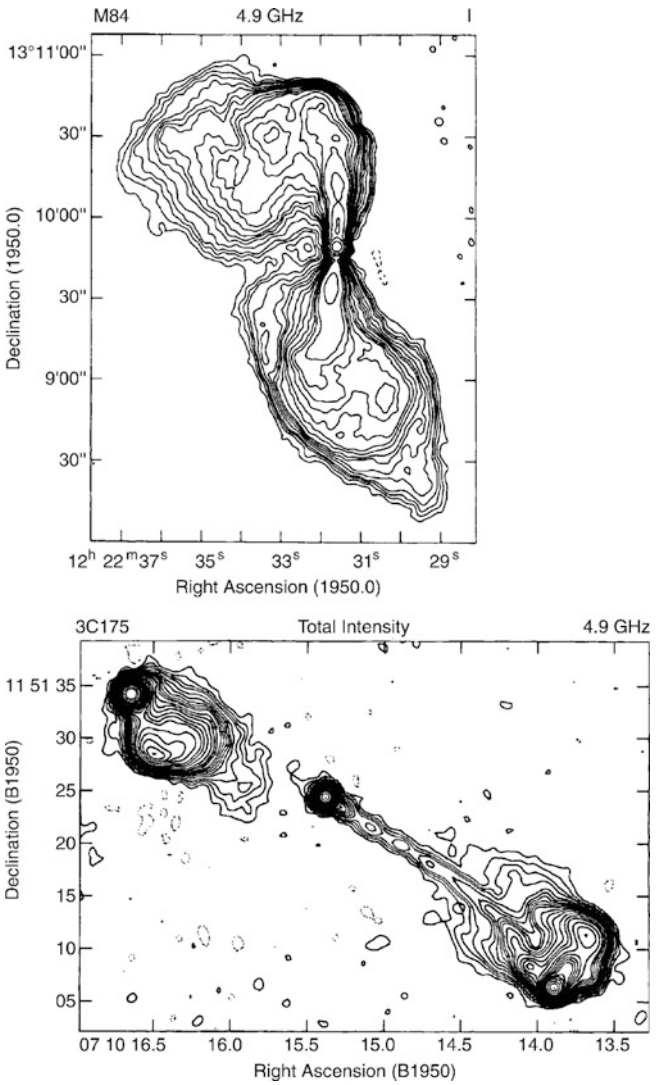


Fig. 5.7 Radio maps at $\lambda = 6$ cm of two radio galaxies: the *top* one is M84, an FR I-radio source, the *bottom* one is 3C175, an FR II-source. The radiation from M84 in the radio is strongest near the center and decreases outwards, whereas in 3C175 the most prominent components are the two radio lobes. The radio lobe on the right is connected to the compact core by a long and very thin jet, whereas on the opposite side no jet (counter-jet) is visible. Source: M84: R.A. Laing & A.H. Bridle 1987, *Rotation measure variation across M84*, MNRAS 228, 557, p. 559, Fig. 1. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society. 3C175: A.H. Bridle et al. 1994, *Deep VLA imaging of twelve extended 3CR quasars*, AJ 108, 766, p. 775, Fig. 7. ©AAS. Reproduced with permission

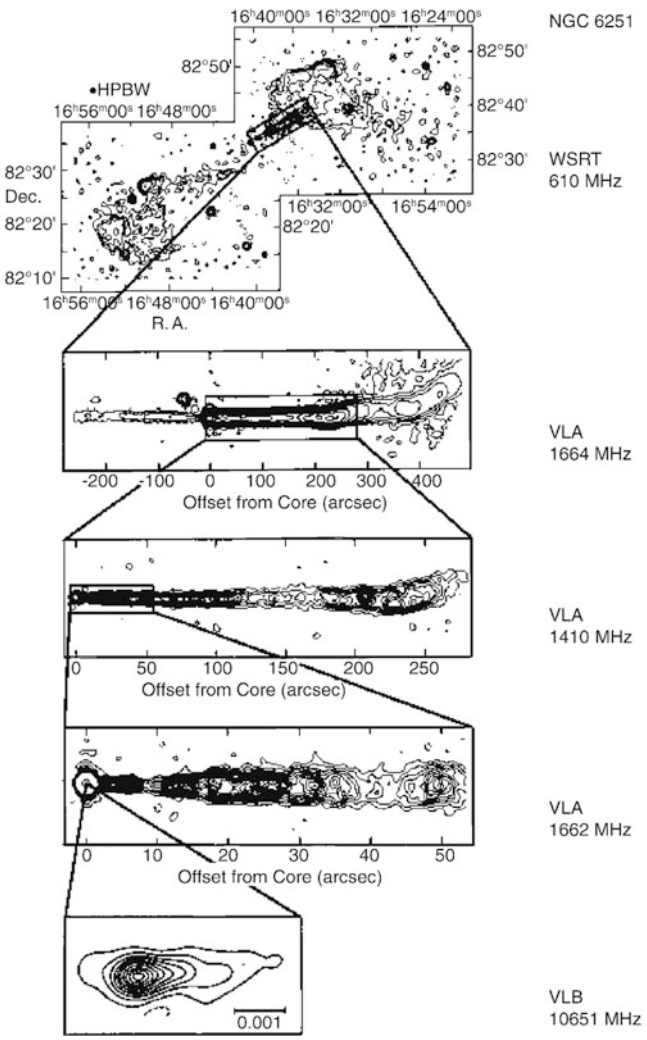


Fig. 5.8 The radio galaxy NGC 6251, with angular resolution increasing towards the bottom. On large scales (and at low frequencies), the two radio lobes dominate, while the core and the jets are clearly prominent at higher frequencies. NGC 6251 has a counter-jet, but with significantly lower luminosity than the main jet. Even at the highest resolution obtained by VLBI, structure can still be seen. The radio jets have a very small opening angle and are therefore strongly collimated. Source: A.H. Bridle & R. Perley 1984, *Extragalactic Radio Jets*, ARA&A 22, 319. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 22 ©1984 by Annual Reviews www.annualreviews.org

this reason, radio catalogs of AGNs suffer from strong selection effects. Catalogs that are sampled at low frequencies will predominantly select sources that have a steep spectrum, i.e., in which the extended structures dominate, whereas high-frequency samples will preferentially contain core-dominated sources with a flat spectrum.[2]

Synchrotron radiation. Over a broad range in wavelengths, the radio spectrum of AGNs follows a power law of the form (5.2), with $\alpha \sim 0.7$ for the extended components and $\alpha \sim 0$ for the compact core components. Radiation in the radio is often linearly polarized, where the extended radio source may reach a degree of polarization up to 30 % or even more. The spectral form and the high degree of polarization are interpreted such that the radio emission is

[2]For this reason, radio surveys for gravitational lens systems, which were mentioned in Sect. 3.11.3, concentrate on sources with a flat spectral index because these are dominated by the compact nucleus. Multiple image systems are thus more easily recognized as such.

produced by *synchrotron radiation* of relativistic electrons. Electrons in a magnetic field propagate along a helical, i.e., corkscrew-shaped path, so that they are continually accelerated by the Lorentz force. Since accelerated charges emit electromagnetic radiation, this motion of the electrons leads to the emission of synchrotron radiation. Because of its importance for our understanding of the radio emission of AGNs, we will review some aspects of synchrotron radiation next.

The radiation can be characterized as follows. If an electron has energy $E = \gamma\, m_\mathrm{e}\, c^2$, the characteristic frequency of the emission is

$$\nu_\mathrm{c} = \frac{3\gamma^2 eB}{4\pi\, m_\mathrm{e} c} \sim 4.2 \times 10^6\, \gamma^2 \left(\frac{B}{1\mathrm{G}}\right) \mathrm{Hz}\,, \qquad (5.3)$$

where B denotes the magnetic field strength, e the electron charge, and $m_\mathrm{e} = 511\,\mathrm{keV}/c^2$ the mass of the electron. The *Lorentz factor* γ, and thus the energy of an electron, is related to its velocity v via

$$\gamma := \frac{1}{\sqrt{1-(v/c)^2}}\,. \qquad (5.4)$$

For frequencies considerably lower than ν_c, the spectrum of a single electron is $\propto \nu^{1/3}$, whereas it decreases exponentially for larger frequencies. To a first approximation, the spectrum of a single electron can be considered as quasi-monochromatic, i.e., the width of the spectral distribution is small compared to the characteristic emission frequency ν_c. The synchrotron radiation of a single electron is linearly polarized, where the observed polarization direction depends on the orientation of the magnetic field projected onto the sky. The degree of polarization of the radiation from an ensemble of electrons depends on the complexity of the magnetic field. If the magnetic field is uniform in the spatial region from which the radiation is measured, the observed polarization may reach values of up to 75 %. However, if the spatial region that lies within the telescope beam contains a complex magnetic field, with the direction changing strongly within this region, the polarizations partially cancel each other out and the observed degree of linear polarization is significantly reduced.

To produce radiation at cm wavelengths ($\nu \sim 10\,\mathrm{GHz}$) in a magnetic field of strength $B \sim 10^{-4}\,\mathrm{G}$, $\gamma \sim 10^5$ is required, i.e., the *electrons need to be highly relativistic!* To obtain particles at such high energies, very efficient processes of particle acceleration must occur in the inner regions of quasars. It should be mentioned in this context that cosmic ray particles of considerably higher energies are observed (see Sect. 2.3.4). The majority of cosmic rays are presumably produced in the shock fronts of supernova remnants. Thus, it is supposed that the energetic electrons in quasars (and other AGNs) are also produced by 'diffusive shock acceleration', where here the shock fronts are not caused by supernova explosions but rather by other hydrodynamical phenomena. As we will see later, we find clear indications in AGNs for outflow velocities that are considerably higher than the speed of sound in the plasma, so that the conditions for the formation of shock fronts are satisfied.

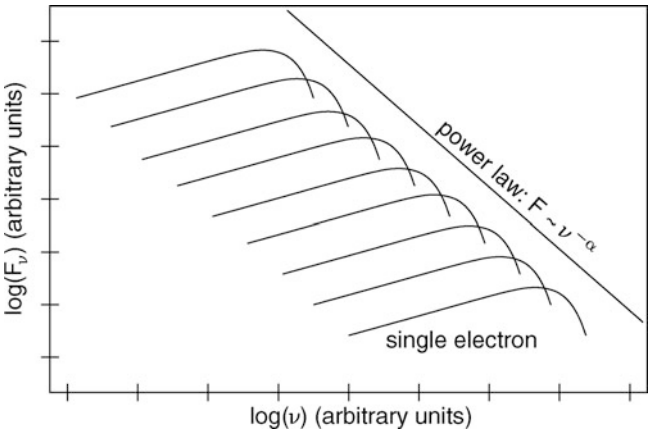


Fig. 5.9 Electrons at a given energy emit a synchrotron spectrum which is indicated by the *individual curves*; the maximum of the radiation is at ν_c (5.3), which depends on the electron energy. The superposition of many such spectra, corresponding to an energy distribution of the electrons, results in a power-law spectrum provided the energy distribution of the electrons follows a power law. Adapted from: B.W. Carroll & D.A. Ostlie 1996, *An introduction to Modern Astrophysics*, Reading

Spectral shape. Synchrotron radiation will follow a power law if the energy distribution of relativistic electrons also behaves like a power law (see Fig. 5.9). If $N(E)\,\mathrm{d}E \propto E^{-s}\,\mathrm{d}E$ represents the number density of electrons with energies between E and $E + \mathrm{d}E$, the power-law index of the resulting radiation will be $\alpha = (s-1)/2$, i.e., the slope in the power law of the electron distribution in energy defines the spectral shape of the resulting synchrotron emission (see Problem 5.1). In particular, an index of $\alpha = 0.7$ results for $s = 2.4$. An electron distribution with $N(E) \propto E^{-2.4}$ is very similar to the energy distribution of the cosmic rays in our Galaxy, which may be another indicator for the same or at least a similar mechanism being responsible for the generation of this energy spectrum.

The synchrotron spectrum is self-absorbed at low frequencies, i.e., the optical depth for absorption due to the synchrotron process is close to or larger than unity. In this case, the spectrum becomes flatter and, for small ν, it may even rise. In the limiting case of a high optical depth for self-absorption, we obtain $S_\nu \propto \nu^{2.5}$ for $\nu \to 0$. The extended radio components are optically thin at cm wavelength, so that $\alpha \sim 0.7$, whereas the compact core component is often

optically thick and thus self-absorbed, which yields $\alpha \sim 0$, or even inverted so that $\alpha < 0$.

Energy loss. Through emission, the electrons lose energy. Thus, the electrons cool and for only a limited time can they radiate at the frequency described by (5.3). The power emitted by an electron of Lorentz factor γ, integrated over all frequencies, is

$$P = -\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{4}{9}\frac{e^4 B^2 \gamma^2}{m_\mathrm{e}^2 c^3} \,. \tag{5.5}$$

The characteristic time in which an electron loses its energy is then obtained from its energy $E = \gamma m_\mathrm{e} c^2$ and its energy loss rate $\dot{E} = -P$ as

$$t_\mathrm{cool} = \frac{E}{P} = 2.4 \times 10^5 \left(\frac{\gamma}{10^4}\right)^{-1} \left(\frac{B}{10^{-4}\,\mathrm{G}}\right)^{-2} \mathrm{yr} \,. \tag{5.6}$$

For relatively low-frequency radio emission, this lifetime is longer than or comparable to the age of radio sources. But as we will see later, high-frequency synchrotron emission is also observed for which t_cool is considerably shorter than the age of a source component. The corresponding relativistic electrons can then only be generated locally. This means that the processes of particle acceleration are not confined to the inner core of an AGN, but also occur in the extended source components.

Since the characteristic frequency (5.3) of synchrotron radiation depends on a combination of the Lorentz factor γ and the magnetic field B, we cannot measure these two quantities independently. Therefore, it is difficult to estimate the magnetic field of a synchrotron source. In most cases, the (plausible) assumption of an equipartition of the energy density in the magnetic field and the relativistic particles is made, i.e., one assumes that the energy density $B^2/(8\pi)$ of the magnetic field roughly agrees with the energy density

$$\int \mathrm{d}\gamma \; n_\mathrm{e}(\gamma)\, \gamma\, m_\mathrm{e} c^2$$

of the relativistic electrons. Such approximate equipartition holds for the cosmic rays in our Galaxy and its magnetic field. Another approach is to estimate the magnetic field such that the total energy of relativistic electrons and magnetic field is minimized for a given source luminosity. The resulting value for B basically agrees with that derived from the assumption of equipartition.

5.1.4 Broad emission lines

The UV and optical spectra of quasars feature strong and very broad emission lines. Typically, lines of the Balmer series and Lyα of hydrogen, and metal lines of ions like MgII, CIII, CIV are observed[3]—these are found in virtually all quasar spectra. In addition, a large number of other emission lines occur which are not seen in every spectrum (Fig. 5.2).

To characterize the *strength* of an emission line, we define the *equivalent width* of a line W_λ as

$$W_\lambda = \int \mathrm{d}\lambda \, \frac{S_\mathrm{l}(\lambda) - S_\mathrm{c}(\lambda)}{S_\mathrm{c}(\lambda)} \approx \frac{F_\mathrm{line}}{S_\mathrm{c}(\lambda_0)} \,, \tag{5.7}$$

where $S_\mathrm{l}(\lambda)$ is the total spectral flux, and $S_\mathrm{c}(\lambda)$ is the spectral flux of the continuum radiation interpolated across the wavelength range of the line. F_line is the total flux in the line and λ_0 its wavelength. Hence, W_λ is the width of the wavelength interval over which the continuum needs to be integrated to obtain the same flux as measured in the line. Therefore, the equivalent width is a measure of the strength of a line relative to the continuum intensity.

The *width* of a line is characterized as follows: after subtracting the continuum, interpolated across the wavelength range of the line, the width is measured at half of the maximum line intensity. This width $\Delta\lambda$ is called the FWHM (full width at half maximum); it may be specified either in Å, or in km/s if the line width is interpreted as Doppler broadening, with $\Delta\lambda/\lambda_0 = \Delta v/c$. It should be noted that the width $\Delta\lambda$ and the equivalent width W_λ are very different quantities. For example, a strong narrow line can have a large W_λ, but a small $\Delta\lambda$. Conversely, a weak broad emission line can have $\Delta\lambda \gg W_\lambda$.

Broad emission lines in quasars often have a FWHM of $\sim$ 10 000 km/s, while narrower emission lines still have widths of several 100 km/s. Thus the 'narrow' emission lines are still broad compared to the typical velocities in normal galaxies.

5.1.5 Quasar demographics

Quasar surveys are always flux limited, i.e., one tries to find all quasars in a certain sky region with a flux above a predefined threshold. Only with such a selection criterion are the samples obtained of any statistical value. In addition, the selection of sources may include further criteria such as color, variability, radio or X-ray flux. For instance, radio surveys are defined by $S_\nu > S_\mathrm{lim}$ at a specific frequency. The optical identification of such radio sources reveals that quasars have a very broad redshift distribution. For decades, quasars have been the only sources known at $z > 3$. Below we will discuss different kinds of AGN surveys.

In the 1993 issue of the quasar catalog by Hewitt & Burbidge, 7236 sources are listed. This catalog contains a

[3] The ionization stages of an element are distinguished by Roman numbers. A neutral atom is denoted by 'I', a singly ionized atom by 'II', and so on. So, CIV is three times ionized carbon.

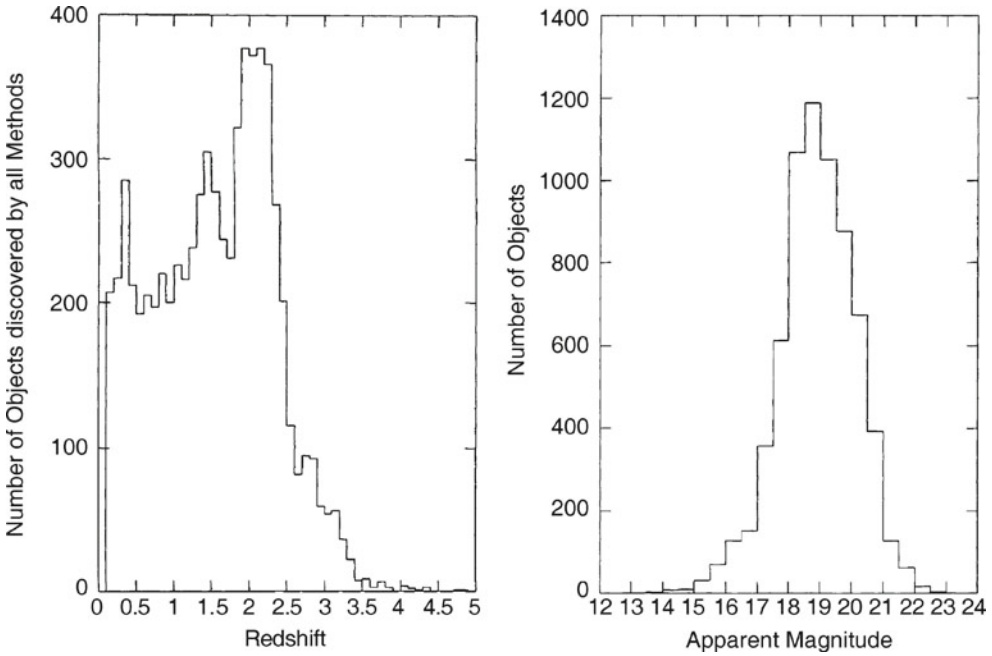


Fig. 5.10 The redshift (*left*) and brightness distribution (*right*) of QSOs in the 1993 Hewitt & Burbidge catalog. These distributions provide no proper statistical information, but they clearly show the width of the distributions. The decrease in abundances for $z \geq 2.3$ is a selection effect: many early QSO surveys started with an optical color selection, typically $U - B < -0.3$. If $z \geq 2.3$, the strong Lyα emission line moves into the B-filter and hence the quasar becomes redder in this color index and drops out of the color selection. Source: A. Hewitt & G. Burbidge 1993, *A revised and updated catalog of quasi-stellar objects*, ApJS 87, 451. ©AAS. Reproduced with permission

broad variety of different AGNs. Although it is statistically not well-defined, this catalog provides a good first indication of the width of the redshift and brightness distribution of AGNs (see Fig. 5.10).

The luminosity function of quasars extends over a very large range in luminosity, nearly three orders of magnitude in L (and over an even broader range if lower-luminosity AGNs are accounted for as well). It is steep at its bright end and has a significantly flatter slope at lower luminosities (see Sect. 5.6.2). We can compare this to the luminosity function of galaxies which is described by a Schechter function (see Sect. 3.10). While the faint end of the distribution is also described here by a relatively shallow power law, the Schechter function decreases exponentially for large L, whereas that of quasars decreases as a power law. For this reason, one finds quasars whose luminosity is much larger than the value of L^* where the break in the luminosity function occurs.

5.2 AGN zoology

Quasars are the most luminous members of the class of AGNs. Seyfert galaxies are another type of AGNs and were mentioned previously. In fact, a wide range of objects are subsumed under the name AGN, all of which have in common strong non-thermal emission in the core of a galaxy (*host galaxy*). We will mention the most important types of AGNs in this section. It is important to keep in mind that the frequency range in which sources are studied affects the source classification. We shall return to this point at the end of this section.

The classification of AGNs described below is very confusing at first glance. Different classes refer to different appearances of AGNs but do not necessarily correspond to the physical nature of these sources. A sample of optical spectra for different types of AGNs is displayed in Fig. 5.11 which illustrates a large variety of spectral properties. Similarly, the properties of the emission of AGNs in different wavebands (such as radio or gamma-rays) can differ most strongly. However, as we will discuss in Sect. 5.5, the large variety of appearances of AGNs can be understood, at least to a first approximation, by geometric considerations. The emission of an AGN is not isotropic; we will see that the flow of material which causes the energy release near the central black hole occurs in the form of a disk (the so-called accretion disk—see Sect. 5.3.2), which defines a pair of preferred directions, i.e., those perpendicular to the plane in which the disk lies. In the context of unified models, the way an AGN appears to us depends strongly on the angle between this disk axis and the line-of-sight to the source.

Outline of the unified model. In Fig. 5.12, this geometric picture of an AGN is sketched. The motivation for this model

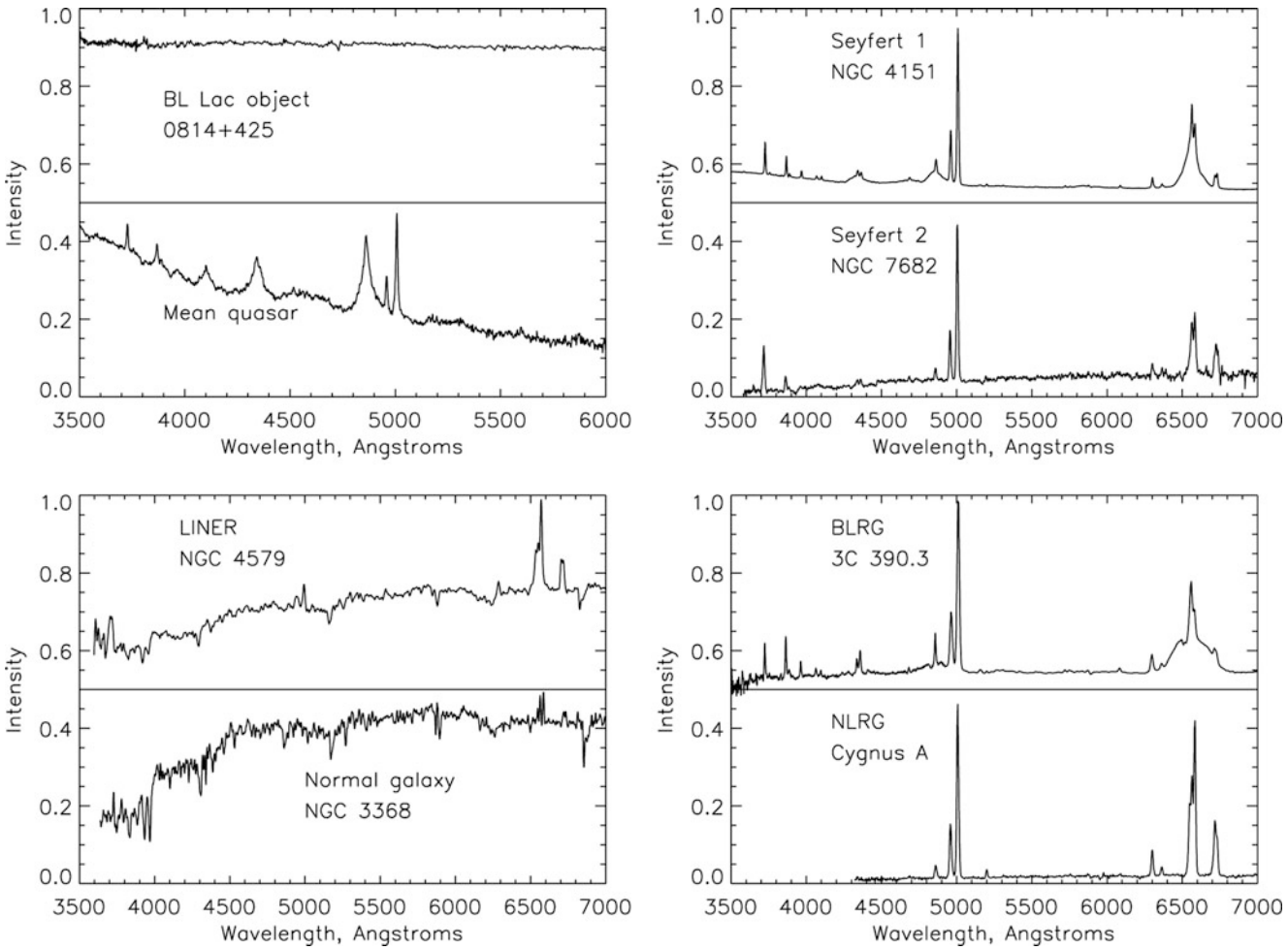


Fig. 5.11 Comparison of the optical spectra of various types of active galaxies with that of a 'normal', inactive galaxy (*bottom left*). *From top to bottom*, *left column*: a BL Lac object, the mean QSO spectrum, a LINER; in the *right column*: Seyfert 1 and 2, a Broad and a Narrow Line Radio Galaxy. Source: Bill Keel's WWW Gallery http://www.astr.ua.edu/keel/agn/spectra.html. Reproduced with permission of Bill Keel

will be explained in the course of this chapter, but we briefly summarize it here to provide a guide for the subsequent description of the various AGN classes. Surrounding the central supermassive black hole is an accretion disk which emits the bulk part of the optical and UV continuum emission. The central region around the accretion disk is the source of most of the X-ray radiation. Gas clouds above and below the accretion disk are responsible for the broad emission lines. In the plane of the disk, a distribution of gas and dust is present, which can absorb radiation from the inner region of the AGN; this obscuring material is sometimes depicted as a torus, though its geometry is probably more complicated. Nevertheless, the appearance of the AGN depends on whether the observer is located near the plane of the disk—where radiation is partly absorbed by the material in the torus—or placed in a direction closer to the axis of the disk. This concerns in particular the broad line emission, which may be fully obscured for an observer in the plane of the disk. In contrast, the gas responsible for the narrow emission lines is located at much larger distances from the black hole, so that it cannot be fully hidden by the obscuring torus.

The radio jets discussed before are launched very close to the central black hole along the direction of the disk axis. The emission from these jets is highly anisotropic, because the velocity in the inner part of the jets is close to the speed of light; then, according to the theory of Special Relativity, the jet emission is strongly beamed in the direction of jet motion. This implies that the appearance of the jet depends on how close the line-of-sight to an observer is to the jet axis. If the jet points almost directly at the observer, the jet emission can outshine all the other radiation from the AGN.

In Fig. 5.12, the different green arrows indicate different lines-of-sight to observers, and they are labeled with the characteristic AGN class the corresponding observer will see. In the upper half of the figure, it is assumed that the AGN produces strong jets, whereas in the lower part, weaker jets (or none at all) are assumed. With this picture in mind, we shall now describe the various types of AGNs.

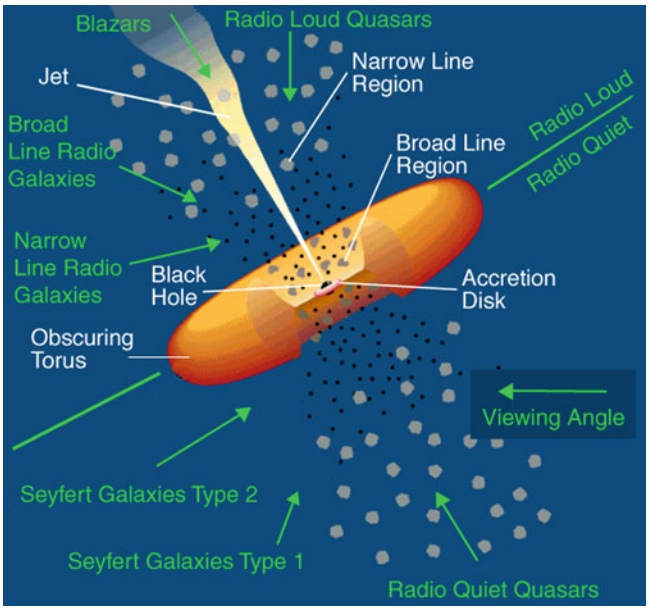


Fig. 5.12 Sketch of our current understanding of the unification of AGN types. The accretion disk is surrounded by a thick 'torus' containing dust which thus obscures the view to the center of the AGN. When looking from a direction near the plane of the disk, a direct view of the continuum source and the BLR is blocked, whereas it is directly visible from directions closer to the symmetry axis of the disk. The difference between Seyfert 1 (and BLRG) and Seyfert 2 (and NLRG) is therefore merely a matter of orientation relative to the line-of-sight. If an AGN is seen exactly along the jet axis, it appears as a blazar. Credit: NASA

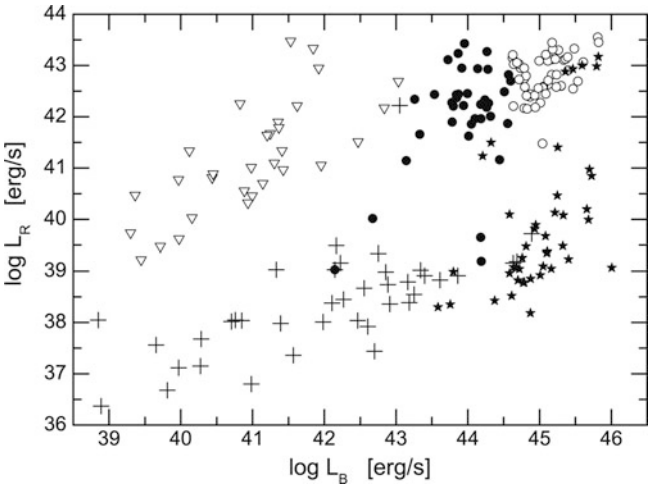


Fig. 5.13 Radio vs. optical luminosity of AGN, as measured at 5 GHz and in the B-band. Different types of AGNs are shown with *different symbols*: FR I radio galaxies (*open triangles*), Broad-Line Radio Galaxies (*filled circles*), radio-loud QSOs (*open circles*), Seyfert galaxies and LINERs (*crosses*), and a sample a $U - B$ color-selected bright QSOs, the Palomar-Green sample (*filled stars*). Apparently, the AGN population is divided into two populations, characterized by their radio-to-optical flux ratio R. Diagrams like this one suggest that there is a bimodal distribution in R, according to radio-loud and radio-quiet AGNs. Source: M. Sikora et al. 2007, *Radio Loudness of Active Galactic Nuclei: Observational Facts and Theoretical Implications*, ApJ 658, 815, p. 823, Fig. 1. ©AAS. Reproduced with permission

5.2.1 QSOs

The unusually blue color of quasars suggested the possibility of searching for them not only with radio observations but also at optical wavelengths, namely to look for point-like sources with a very blue $U - B$ color index. These photometric surveys were very successful. In fact, many more such sources were found than expected from radio counts. Most of these sources are (nearly) invisible in the radio domain of the spectrum; such sources are called radio-quiet. Their optical properties are virtually indistinguishable from those of quasars. In particular, they have a blue optical energy distribution (of course, since this was the search criterion!), strong and broad emission lines, and in general a high redshift.

Apart from their radio properties, these sources appear to be like quasars. Therefore they were called *radio-quiet quasars*, or quasi-stellar objects, QSOs. Today this terminology is no longer very common because the clear separation between sources with and without radio emission is not considered valid any more. Radio-quiet quasars also show radio emission if they are observed at sufficiently high sensitivity. In modern terminology, the expression QSO encompasses both the quasars and the radio-quiet QSOs. About 10 times more radio-quiet QSOs than quasars are thought to exist.

In fact, there is as yet not a clear consensus in whether QSOs show a bimodal distribution in their ratio of radio-to-optical luminosity. Figure 5.13 shows several different samples of AGN; in particular, optically-selected QSOs from the Palomar-Green survey (filled stars) and radio-loud QSOs (open circles). It seems that the ratio between radio and optical luminosity falls into two broad ranges, with a clear gap in between. Therefore, diagrams like that argue in favor of a bimodal distribution. However, this apparent division into two classes can at least partly be attributed to selection effects: the distribution of the radio-to-optical flux ratio depends on the selection of the QSO sample. Obviously, selecting them by their radio emission will favor those with a large $L_{\rm radio}/L_{\rm opt}$-ratio. Furthermore, the fraction of QSOs for which this ratio is large (i.e., which would be termed as radio-loud QSOs) depends on optical luminosity and on redshift: One finds a significantly higher radio-loud fraction amongst more luminous, and lower-redshift QSOs.

The QSOs are the most luminous AGNs. Their core luminosity can be as high as a thousand times that of an L^*-galaxy. Therefore they can outshine their host galaxy and appear point-like on optical images. For QSOs of lower L, their host galaxies were identified and spatially resolved with the HST (see Fig. 1.14). According to our current understanding, AGNs are the active cores of galaxies. These galaxies are supposed to be fairly normal galaxies, except for

their intense nuclear activity, and we will discuss possible reasons for the onset of this activity further below.

5.2.2 Seyfert galaxies

Seyfert galaxies are the AGNs that were detected first. Their nuclear luminosity is considerably lower than that of QSOs. On optical images they are identified as spiral galaxies which have an extraordinarily bright core (Fig. 5.5) whose spectrum shows strong emission lines which are broader than typical velocities in galaxies.

We distinguish between Seyfert galaxies of Type 1 and Type 2: Seyfert 1 galaxies have both very broad and also narrower emission lines, where 'narrow' still means several hundred km/s and thus a significantly larger width than characteristic velocities (like rotational velocities) found in normal galaxies. Seyfert 2 galaxies show only the narrower lines. Later, it was discovered that intermediate variants exist—one now speaks of Seyfert 1.5 and Seyfert 1.8 galaxies, for instance—in which very broad lines exist but with a smaller ratio of broad-to-narrow line flux than in Seyfert 1 galaxies. The classical Seyfert 1 galaxy is NGC 4151 (see Fig. 5.5), while NGC 1068 (Fig. 5.4) is a typical Seyfert 2 galaxy.

The optical spectrum of the nucleus of Seyfert 1 galaxies is very similar to that of QSOs. A smooth transition exists between (radio-quiet) QSOs and Seyfert 1 galaxies. Formally, these two classes of AGNs are separated at an absolute magnitude of $M_B = -21.5+5\log h$. The separation of Seyfert 1 galaxies and QSOs is historical since these two categories were introduced only because of the different methods of discovering them. However, except for the different core luminosity, no fundamental physical difference seems to exist. Often both classes are combined under the name Type 1 AGNs.

5.2.3 LINERs

The least luminous, and by far most common type of AGNs are the LINERs, low-ionization nuclear emission-line regions. In fact, at least one third of all nearby galaxies contain a LINER in their core, characterized by emission lines from neutral atoms or ions with rather low ionization energies. In contrast, emission from lines of strongly ionized ions is either weak or absent. Furthermore, the width of emission lines in LINERs is typically smaller than the narrow emission lines in Seyfert galaxies, and not much larger than the rotational velocity of the galaxy. However, in some LINERs one can find low-luminosity broad emission wings of the Balmer lines; these are sometimes called Type-1 LINERS.

Interestingly, spectra similar to LINERs are frequently found from low-density warm ionized gas in early-type galaxies, with the emission region being spatially extended. For these sources, it can be ruled out that the energy source of the line emission is due to a central AGN. Given the spectral similarity with the LINER emission from the center of spirals, it is sometimes questioned whether the latter phenomenon is indeed a signature of an AGN, or whether LINERs can be powered by star-formation activity, namely by post-AGB stars. In addition, if LINERs are AGNs, then there is no general consensus whether they form a distinct subclass, or whether they are the low-luminosity end of the distribution function of Seyfert galaxies.

5.2.4 Radio galaxies

Radio galaxies are elliptical galaxies with an active nucleus. They were the first sources that were identified with optical counterparts in the early radio surveys. Characteristic radio galaxies are Cygnus A (Fig. 5.6) and Centaurus A (see Fig. 5.48 below).

Similarly to Seyfert galaxies, for radio galaxies we also distinguish between those with and without broad emission lines: broad-line radio galaxies (BLRG) and narrow-line radio galaxies (NLRG), respectively. In principle, the two types of radio galaxies can be considered as radio-loud Seyfert 1 and Seyfert 2 galaxies. A smooth transition between BLRG and quasars also seems to exist, again separated by optical luminosity of the nucleus as for Seyfert galaxies.

Besides the classification of radio galaxies into BLRG and NLRG with respect to the optical spectrum, they are distinguished according to their radio morphology. As was discussed in Sect. 5.1.2, radio sources are divided into FR I and FR II sources.

5.2.5 OVVs

One subclass of QSOs is characterized by the very strong and rapid variability of its optical radiation. The flux of these sources, which are known as Optically Violently Variables (OVVs), can vary by a significant fraction on time-scales of days (see Fig. 5.14). Besides this strong variability, OVVs also stand out because of their relatively high polarization of optical light, typically a few percent, whereas the polarization of normal QSOs is below $\sim 1\,\%$. OVVs are usually strong radio emitters. Their radiation also varies in other wavelength regions besides the optical, with shorter time-scales and larger amplitudes at higher frequencies.

5.2.6 BL Lac objects

The class of AGNs called BL Lac objects (or short: BL Lacs) is named after its prototypical source BL Lacertae. They are

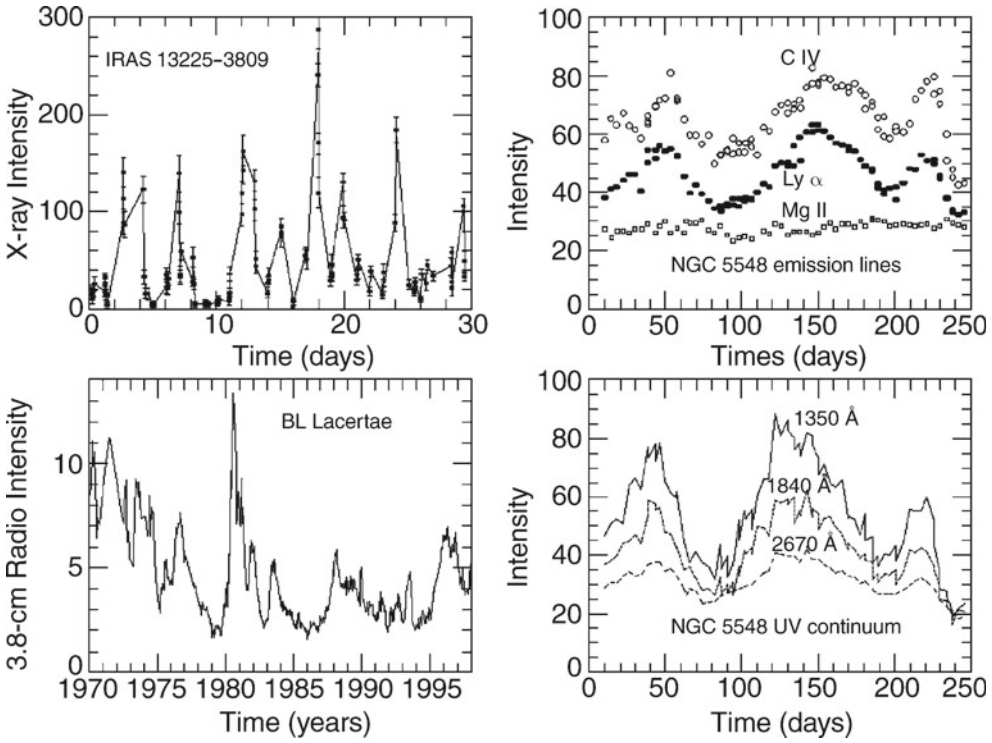


Fig. 5.14 Quasars, BL Lac objects, and Seyfert galaxies all show clear variability at many different wavelengths. In the *upper left panel*, the X-ray light curve of the Seyfert 1 galaxy IRAS 13225−3809 is plotted (observed by ROSAT); on time-scales of days, the source frequently varies by more than a factor of 20. The radio light curve of BL Lacertae at $\lambda = 3.8$ cm covering a period of 28 years is shown in the *lower left panel*. Short-term variations of such blazars are observed in a number of bursts, some overlapping (see, e.g., the burst in 1981). The UV variability of NGC 5548, a Seyfert 1 galaxy, observed by the IUE satellite is plotted for three wavelengths in the *lower right panel*. Variations at these frequencies appear to be in phase, but the amplitude becomes larger towards smaller wavelengths. Simultaneously, the line strengths of three broad emission lines of this Seyfert 1 galaxy have been measured and are plotted in the *upper right panel*. It is found that lines of high ionization potentials, like CIV, have higher variability amplitudes than those of low ionization potentials, like MgII. From the relative temporal shift in the line variability and the continuum flux, the size of the broad line region can be estimated—see Sect. 5.4.2. Credit: Webpage William C. Keel, University of Alabama

AGNs with very strongly varying radiation, like the OVVs, but without strong emission and absorption lines. As for OVVs, the optical radiation of BL Lacs is highly polarized. Since no emission lines are observed in the spectra of BL Lacs, the determination of their redshift is often difficult and sometimes impossible. In some cases, absorption lines are detected in the spectrum which are presumed to derive from the host galaxy of the AGN and are then identified with the redshift of the BL Lac.

The optical luminosity of some BL Lacs varies by several magnitudes if observed over a sufficiently long time period. Particularly remarkable is the fact that in epochs of low luminosity, emission lines are sometimes observed and then a BL Lac appears like an OVV. For this reason, OVVs and BL Lacs are collectively called *blazars*. All known blazars are radio sources. Besides the violent variability, blazars also show highly energetic and strongly variable γ-radiation (Fig. 5.15). Table 5.1 summarizes the fundamental properties of the different classes of AGNs.

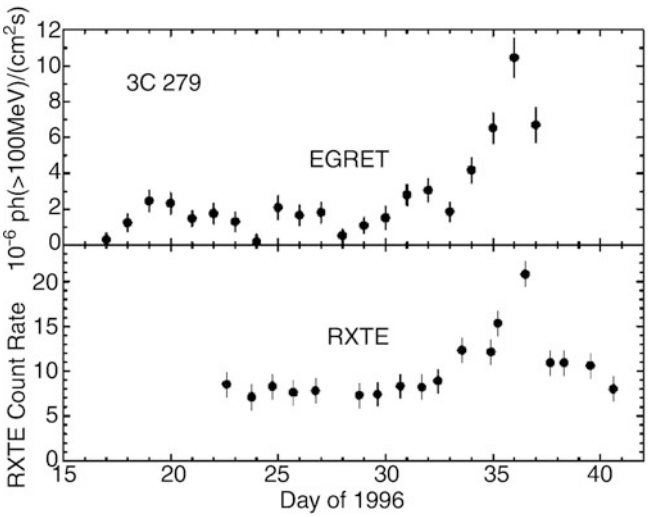


Fig. 5.15 Variability of the blazar 3C279 in X-ray (*bottom*) and in γ-radiation at photon energies above 100 MeV (*top*). On time-scales of a few days, the luminosity varies by a factor ~ 10. Credit: EGRET Team; NASA; research article: R.C. Hartman et al. 2001, ApJ 558, 583

Table 5.1 Overview of the classification of active galactic nuclei

	Normal galaxy	Radio galaxy	Seyfert galaxy	Quasar	Blazar
Example	Milky Way	M87, Cygnus A	NGC 4151	3C273	BL Lac, 3C279
Galaxy type	Spiral	Elliptical, Irregular	Spiral	Irregular	Elliptical?
$L_{\rm AGN}/L_\odot$	$< 10^4$	10^6–10^8	10^8–10^{11}	10^{11}–10^{14}	10^{11}–10^{14}
$M_{\rm BH}/M_\odot$	4×10^6	3×10^9	10^6–10^9	10^6–10^9	10^6–10^9
Radio emission	Weak	Core, jets, lobes	Only $\approx$5 % radio-loud	Only $\approx$5 % radio-loud	Strong, Short-time variable
X-ray emission	Weak	Strong	Strong	Strong	Strong
Gamma emission	Weak	Weak	Medium	Strong	Strong

5.3 The central engine: a black hole

We have previously mentioned that the energy production in AGNs must be related to a supermassive black hole (SMBH) in its center. We will present arguments for this conclusion in this section. To do this, we will first summarize some of the relevant observational facts for AGNs.

- The extent of some radio sources in AGNs may reach $\gtrsim$1 Mpc. From this length-scale a minimum lifetime for the activity in the nucleus of these objects can be derived, since even if the radio source expands outwards from the core with the speed of light, the age of such a source would be $\tau \gtrsim 10^7$ yr.
- Luminous QSOs have a luminosity of up to $L_{\rm bol} \sim 10^{47}$ erg/s. Assuming that the luminosity does not change substantially over the lifetime of the source, a total energy can be estimated from the luminosity and the minimum age,

$$E \gtrsim 10^{47}\,\mathrm{erg/s} \times 10^7\,\mathrm{yr} \sim 3 \times 10^{61}\,\mathrm{erg}\;; \qquad (5.8)$$

  however, the assumption of an essentially constant luminosity is not necessarily justified.
- The luminosity of some AGNs varies by more than 50 % on time-scales of a day. From this variability time-scale, an upper limit for the spatial extent of the source can be determined, because the source luminosity can change substantially only on such time-scales where the source as a whole, or at least a major part of the emitting region, is in causal contact. Otherwise 'one end' of the source does not know that the 'other end' is about to vary. This yields a characteristic extent of the central source of $R \lesssim$ 1 lightday $\sim 3 \times 10^{15}$ cm.

5.3.1 Why a black hole?

We will now combine the aforementioned observations and derive from them that the basic energy production in AGNs has to be of a gravitational nature. To do this, we note that the most efficient 'classical' method of energy production is nuclear fusion, as is taking place in stars. We will therefore make the provisional assumption (which will soon lead to a contradiction) that the energy production in AGNs is based on thermonuclear processes.

By burning hydrogen into iron—the nucleus with the highest binding energy per nucleon—8 MeV/nucleon are released, or $0.008\,m_{\rm p}c^2$ per nucleon. The maximumefficiency of nuclear fusion is therefore $\epsilon \lesssim 0.8\,\%$, where ϵ is defined as the mass fraction of 'fuel' that is converted into energy, according to

$$\boxed{E = \epsilon\, mc^2}\,. \qquad (5.9)$$

To generate the energy of $E = 3 \times 10^{61}$ erg by nuclear fusion, a total mass m of fuel would be needed, where m is given by

$$m = \frac{E}{\epsilon c^2} \sim 4 \times 10^{42}\,\mathrm{g} \sim 2 \times 10^9 M_\odot\,, \qquad (5.10)$$

where we used the energy estimate from (5.8). If the energy of an AGN was produced by nuclear fusion, burnt-out matter of mass m [more precisely, $(1-\epsilon)m$] must be present in the core of the AGN.

However, the Schwarzschild radius of this mass is (see Sect. 3.8.1)

$$r_{\rm S} = \frac{2Gm}{c^2} = \frac{2GM_\odot}{c^2}\frac{m}{M_\odot}$$
$$= 3 \times 10^5\,\mathrm{cm}\,\frac{m}{M_\odot} \sim 6 \times 10^{14}\,\mathrm{cm}\,,$$

i.e., the Schwarzschild radius of the 'nuclear cinder' is of the same order of magnitude as the above estimate of the extent of the central source. This argument demonstrates that the gravitational binding energy of the cinder is far higher than the energy released from nuclear burning. Hence, gravitational effects *must* play a crucial role—the assumption

of thermonuclear energy as prime energy source has been disproven because its efficiency ϵ is too low. The only known mechanism yielding larger ϵ is gravitational energy production.

Through the infall of matter onto a central black hole, potential energy is converted into kinetic energy. If it is possible to convert part of this inward-directed kinetic energy into internal energy (heat) and subsequently emit this in the form of radiation, ϵ can be larger than that of thermonuclear processes. From the theory of accretion onto black holes, a maximum efficiency of $\epsilon \sim 6\,\%$ for accretion onto a non-rotating black hole (also called a Schwarzschild hole) is derived. A black hole with the maximum allowed angular momentum can have an efficiency of $\epsilon \sim 29\,\%$.

5.3.2 Accretion

Due to its broad astrophysical relevance beyond the context of AGNs, we will consider the accretion process in somewhat more detail.

The principle of accretion. Gas falling onto a compact object loses its potential energy, which is first converted into kinetic energy. If the infall is not prevented, the gas will fall into the black hole without being able to radiate this energy. In general one can expect that the gas has finite angular momentum. Thus it cannot fall straight onto the compact object, since this is prevented by the angular momentum barrier. Through friction with other gas particles and by the resulting momentum transfer, the gas will assemble in a disk oriented perpendicular to the direction of the angular momentum vector. The frictional forces in the gas are expected to be much smaller than the gravitational force. Hence the disk will locally rotate with approximately the Kepler velocity. Since a Kepler disk rotates differentially, in the sense that the angular velocity depends on radius, the gas in the disk will be heated by internal friction. In addition, the same friction causes a slight deceleration of the rotational velocity, whereby the gas will slowly move inwards. The energy source for heating the gas in the disk is provided by this inward motion—namely the conversion of potential energy into kinetic energy, which is then converted into internal energy (heat) by friction.

According to the virial theorem, half of the potential energy released is converted into kinetic energy; in the situation considered here, this is the rotational energy of the disk. The other half of the potential energy can be converted into internal energy. We now present an approximately quantitative description of this process, specifically for accretion onto a black hole.

Temperature profile of a geometrically thin, optically thick accretion disk. When a mass m falls from radius $r + \Delta r$ to r, the energy

$$\Delta E = \frac{GM_\bullet m}{r} - \frac{GM_\bullet m}{r+\Delta r} \approx \frac{GM_\bullet m}{r}\,\frac{\Delta r}{r}$$

is released. Here $M_\bullet$ denotes the mass of the SMBH, assumed to dominate the gravitational potential, so that self-gravity of the disk can be neglected. Half of this energy is converted into heat, $E_{\rm heat} = \Delta E/2$. If we assume that this energy is emitted locally, the corresponding luminosity is

$$\Delta L = \frac{GM_\bullet \dot{m}}{2r^2}\Delta r\ , \tag{5.11}$$

where $\dot{m}$ denotes the accretion rate, which is the mass that falls into the black hole per unit time. In the stationary case, $\dot{m}$ is independent of radius since otherwise matter would accumulate at some radii. Hence the same amount of matter per unit time flows through any cylindrical radius.

If the disk is optically thick, the local emission corresponds to that of a black body. The ring between r and $r+\Delta r$ then emits a luminosity

$$\Delta L = 2 \times 2\pi r\,\Delta r\,\sigma_{\rm SB} T^4(r)\ , \tag{5.12}$$

where the factor 2 originates from the fact that the disk has two sides. Combining (5.11) and (5.12) yields the radial dependence of the disk temperature,

$$T(r) = \left(\frac{GM_\bullet \dot{m}}{8\pi\sigma_{\rm SB} r^3}\right)^{1/4}\ .$$

A more accurate derivation explicitly considers the dissipation by friction and accounts for the fact that part of the generated energy is used for heating the gas, where the corresponding thermal energy is also partially advected inwards. Except for a numerical correction factor, the same result is obtained,

$$\boxed{T(r) = \left(\frac{3GM_\bullet \dot{m}}{8\pi\sigma_{\rm SB} r^3}\right)^{1/4}}\ , \tag{5.13}$$

which is valid in the range $r \gg r_{\rm S}$. Scaling r with the Schwarzschild radius $r_{\rm S}$, we obtain

$$T(r) = \left(\frac{3GM_\bullet \dot{m}}{8\pi\sigma_{\rm SB} r_{\rm S}^3}\right)^{1/4} \left(\frac{r}{r_{\rm S}}\right)^{-3/4}\ .$$

By replacing $r_{\rm S}$ with (3.43) in the first factor, this can be written as

$$T(r) = \left(\frac{3c^6}{64\pi\sigma_{\rm SB}G^2}\right)^{1/4} \dot{m}^{1/4}\, M_\bullet^{-1/2} \left(\frac{r}{r_{\rm S}}\right)^{-3/4} . \tag{5.14}$$

Interpretation and conclusions. From this analysis, we can immediately draw a number of conclusions. The most surprising one may be the independence of the temperature profile of the disk from the detailed mechanism of the dissipation because the equations do not explicitly contain the viscosity. This fact allows us to obtain quantitative predictions based on the model of a *geometrically thin, optically thick accretion disk*.[4] The temperature in the disk increases inwards $\propto r^{-3/4}$, as expected. Therefore, the total emission of the disk is, to a first approximation, a superposition of black bodies consisting of rings with different radii at different temperatures. For this reason, the resulting spectrum does not have a Planck shape but instead shows a much broader energy distribution. Over a wide range of frequencies, the resulting spectrum from such an optically thick accretion disk is fairly flat, where the lower and upper bound of the frequency interval is determined by the lowest and highest temperature (at the outer and inner radius) of the disk.

Most of the luminosity from a disk comes from the inner parts, and thus depends critically on how far the disk extends inside. Around a black hole, there is a minimum radius $r_{\rm in}$ at which stable circular orbits can exist. For a black hole without rotation, this innermost stable orbit is at $r_{\rm in} = 3r_{\rm S}$, whereas it is smaller for a black hole with angular momentum. Accordingly, the efficiency

$$\epsilon = \frac{L}{\dot{m}c^2} \tag{5.15}$$

with which accreting mass is converted into luminosity depends on the black hole spin. It increases from $\sim 6\,\%$ for a non-rotating black hole to $\sim 29\,\%$ for one with maximum rotation.

For any fixed ratio $r/r_{\rm S}$, the temperature increases with the accretion rate. This again was expected: since the local emission is $\propto T^4$ and the locally dissipated energy is $\propto \dot{m}$, it must be $T \propto \dot{m}^{1/4}$. Furthermore, at fixed ratio $r/r_{\rm S}$, the temperature decreases with increasing mass $M_\bullet$ of the black hole. This implies that the maximum temperature attained in the disk is lower for more massive black holes. This may be unexpected, but it is explained by a decrease of the tidal forces, at fixed $r/r_{\rm S}$, with increasing $M_\bullet$. In particular, it implies that the maximum temperature of the disk in an AGN is much lower than in accretion disks around compact objects of stellar mass. Accretion disks around neutron stars and stellar-mass black holes emit in the hard X-ray part of the spectrum and are known as X-ray binaries. In contrast, the thermal radiation of the disk in an AGN extends to the UV or soft X-ray range only (see below).

Radiatively inefficient accretion. The disk accretion described above requires that the generated energy is emitted locally, which requires the disk to be optically thick. The optical depth of the disk depends on its surface density, which in turn depends on the accretion rate. In a system where the accretion rate $\dot{m}$ is low (in a sense quantified further below), the disk may be optically thin, and the emission process of the heated gas can become inefficient. In this case, the gas cannot efficiently cool, and the thermal energy generated by friction in the disk is advected inwards together with the gas. Such a disk (called 'advection-dominated accretion flow', or ADAF) is rather inefficient in converting rest mass into radiation, and so its corresponding ϵ can be quite small. However, one expects that such an accretion flow may be quite efficient in generating outflows, such that part of the accreted material is ejected in form of jets. Hence, this mode of accretion may play an important role for radio galaxies.

Bondi–Hoyle–Lyttleton accretion. In the absence of radiation pressure, the mean accretion rate is determined by the flux of matter that is added to the outer parts of the accretion disk. This quantity is difficult to estimate and depends on the rate with which gas in a galaxy can be transported inwards. Owing to the angular momentum of the gas, it presumably moves to the central region only through significant perturbations of the gravitational potential from axisymmetry. Details of these processes are not fully understood yet.

However, there is one simple situation where the accretion rate can be estimated analytically, namely the case of spherical accretion from a static medium. Assume a black hole being immersed into a spherically-symmetric gas distribution which for large radii is homogeneous with density ρ_∞ and sound speed $c_{\rm s}$. The gravitational pull by the black hole causes the gas to have an inward-directed velocity. Provided the gas is adiabatic, then the mass accretion rate can be calculated from the equations of fluid dynamics, yielding

[4] The physical mechanism that is responsible for the viscosity is unknown. The molecular viscosity is far too small to be considered as the primary process. Rather, the viscosity is probably produced by turbulent flows in the disk or by magnetic fields, which become spun up by differential rotation and thus amplified, so that these fields may act as an effective friction. In addition, hydrodynamic instabilities may act as a source of viscosity. Although the properties of the accretion disk presented here—luminosity and temperature profile—are independent of the specific mechanism of the viscosity, other disk properties definitely depend on it. For example, the temporal behavior of a disk in the presence of a perturbation, which is responsible for the variability in some binary systems, depends on the magnitude of the viscosity, which therefore can be estimated from observations of such systems.

$$\dot{m} = \frac{4\pi G^2 M_\bullet^2}{c_s^3}\rho_\infty \ . \qquad (5.16)$$

This *Bondi–Hoyle–Lyttleton accretion rate* yields an indication of the mass influx onto the accretion disk, provided the angular momentum of the surrounding gas is sufficiently small. In this case, it can flow in at the rate given by (5.16), until it reaches a radius where the angular momentum becomes important and the gas is forced onto circular orbits, forming an accretion disk. Purely spherical accretion, i.e., where the gas has zero angular momentum and no disk is formed, is very inefficient; only a tiny fraction of the kinetic energy gets dissipated and radiated away.

5.3.3 Superluminal motion

Apparent velocities larger than c. Besides the generation of energy, another piece of evidence for the existence of SMBHs in the centers of AGNs results from observing relative motions of source components at *superluminal* velocities. These observations of central radio components in AGNs are mainly made using VLBI methods since they provide the highest available angular resolution. They measure a time dependence of the angular separation of source components, which often leads to values $> c$ if the angular velocity is translated into a transverse spatial velocity (Fig. 5.16). These superluminal motions caused some discomfort upon their discovery. In particular, they at first raised concerns that the redshift of QSOs may not originate from cosmic expansion. Only if the QSO redshifts are interpreted as being of cosmological origin can they be translated into a distance, which is needed to convert the observed angular velocity into a spatial velocity.

We consider two source components (e.g., the radio core and a component in the jet) which are observed to have a time-dependent angular separation $\theta(t)$. If D denotes the distance of the source, then the apparent relative transverse velocity of the two components is

$$v_{\rm app} = \frac{{\rm d}r}{{\rm d}t} = D\,\frac{{\rm d}\theta}{{\rm d}t} \ , \qquad (5.17)$$

where $r = D\theta$ is the transverse separation of the two components. The final expression in (5.17) shows that $v_{\rm app}$ is directly observable if the distance D is assumed to be known.

Frequently, VLBI observations of compact radio sources yield values for $v_{\rm app}$ that are larger than c! Characteristic values for sources with a dominant core component are $v_{\rm app} \sim 5c$ (see Fig. 5.16). But according to the theory of Special Relativity, velocities $> c$ do not exist. Thus it is

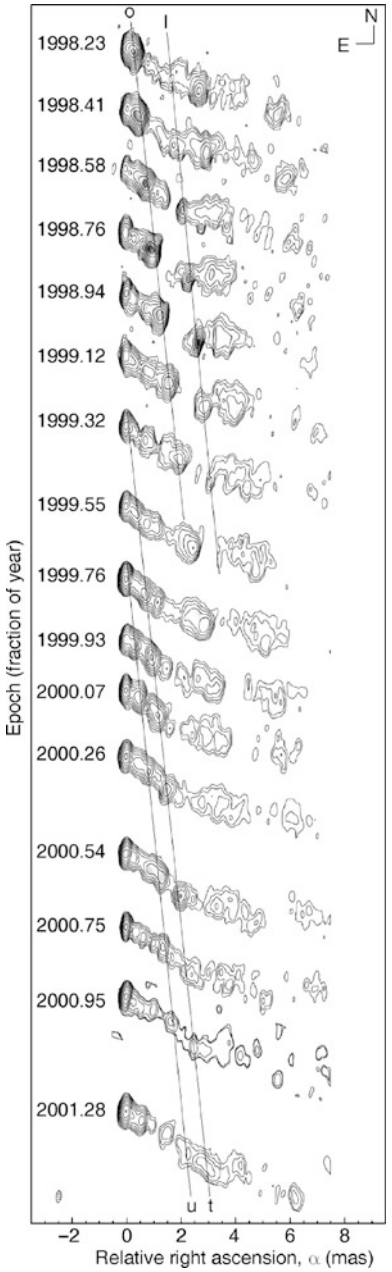


Fig. 5.16 Apparent superluminal velocities of source components in the radio jet of the source 3C120. VLBA observations of this source are presented for 16 different epochs (indicated by the *numbers* at the left of the corresponding radio map), observed at 7 mm wavelength. The ellipse at the *lower left* indicates the beam of the VLBA interferometer and thus the angular resolution of these observations. At the distance of 3C120 of 140 Mpc, a milliarcsecond corresponds to a linear scale of 0.70 pc. The four *straight lines*, denoted by l, o, t, and u, connect the same source components at different epochs. The linear motion of these components is clearly visible. The observed angular velocities of the components yield apparent transverse velocities in the range of $4.1c$ to $5c$. Source: A.P. Marscher et al. 2002, *Observational evidence for the accretion-disk origin for a radio jet in an active galaxy*, Nature 417, 625, Fig. 1

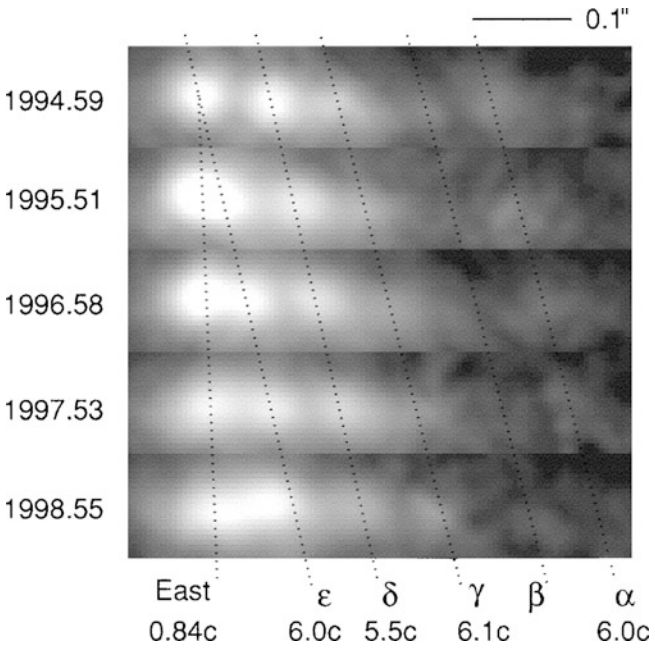


Fig. 5.17 Also at optical wavelengths, apparent superluminal motion was observed. The figure shows the optical jet in M87, based on HST images taken over a period of about 4 years. The angular velocity of the components is up to 23 mas/yr. Assuming a distance of M87 of $D = 16$ Mpc, velocities of up to $\sim 6c$ are obtained for the components. Source: J.A. Biretta et al. 1999, *Hubble Space Telescope Observations of Superluminal Motion in the M87 Jet*, ApJ 520, 621, p. 623, Fig. 2. ©AAS. Reproduced with permission

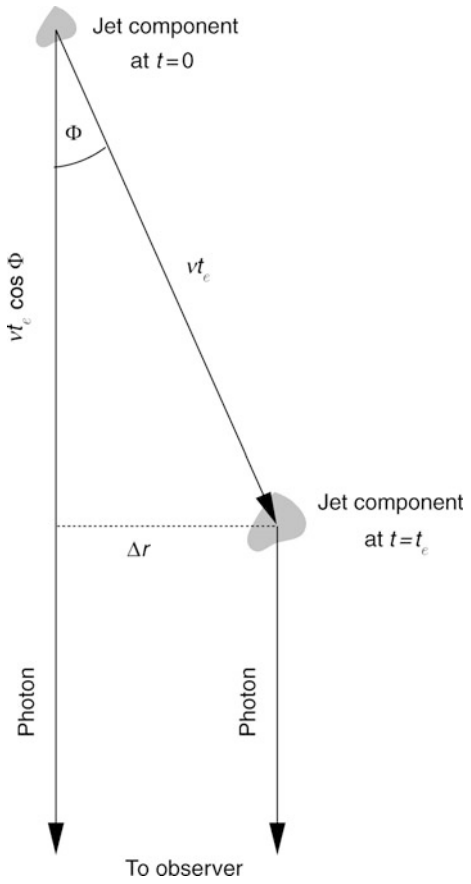


Fig. 5.18 Explanation of superluminal motion: a source component is moving at velocity v and at an angle ϕ relative to the line-of-sight. We consider the emission of photons at two different times $t = 0$ and $t = t_e$. Photons emitted at $t = t_e$ will reach us by $\Delta t = t_e(1-\beta\cos\phi)$ later than those emitted at $t = 0$. The apparent separation of the two source components then is $\Delta r = v t_e \sin\phi$, yielding an apparent velocity on the sky of $v_{app} = \Delta r/\Delta t = v\sin\phi/(1-\beta\cos\phi)$. Adapted from: B.W. Carroll & D.A. Ostlie 1996, *An introduction to Modern Astrophysics*, Reading

not surprising that the phenomenon of superluminal motion engendered various kinds of explanations upon its discovery. By now, superluminal motion has also been seen in optical observations of jets, as is displayed in Fig. 5.17.

For some time, one possibility that had been considered was that the cosmological interpretation of the redshifts may be wrong, because for a sufficiently small D velocities smaller than the speed of light would result from (5.17). However, no plausible alternative explanations for the observed redshifts of QSOs exist, and more than 40 years of QSO observations have consistently confirmed that redshift is an excellent measure for their distances—see Sect. 4.5.1.

However, Relativity only demands that no *signal* may propagate with velocities $> c$. It is easy to construct a thought-experiment in which superluminal velocities occur. For instance, consider a laser beam or a flashlight that is rotating perpendicular to its axis of symmetry. The corresponding light point on a screen changes its position with a speed proportional to the angular velocity and to the distance of the screen from the light source. If we make the latter sufficiently large, it is 'easy' to obtain a superluminal light point on the screen. But this light point does not carry a signal along its track. Therefore, the superluminal motions in compact radio sources may be explained by such a screen effect, but what is the screen and what is the laser beam? We point out that apparent superluminal velocities are seen in the center of the Milky Way in form of X-ray echos—see Sect. 2.6.5—a phenomenon where a screen (of scattering material) explains the effect.

Explanation of superluminal motion. The generally accepted explanation of apparent superluminal motion combines very fast motions of source components with the finite speed of light. For this, we consider a source component moving at speed v at an angle ϕ with respect to the line-of-sight (see Fig. 5.18). We arbitrarily choose the origin of time $t = 0$ to be the time at which the moving component is close to the core component. At time $t = t_e$, the source has a distance $v\,t_e$ from the original position. The observed separation is the transverse component of this distance,

$$\Delta r = v\,t_e \sin\phi\ .$$

Since at time t_e the source has a smaller distance from Earth than at $t = 0$, the light will accordingly take slightly less time to reach us. Photons emitted at times $t = 0$ and $t = t_e$ will reach us with a time difference of

$$\Delta t = t_e - \frac{v\, t_e \cos\phi}{c} = t_e\,(1 - \beta\cos\phi) \ ,$$

where we define

$$\beta := \frac{v}{c} \tag{5.18}$$

as the velocity in units of the speed of light. Equation (5.17) then yields the apparent velocity,

$$\boxed{v_{app} = \frac{\Delta r}{\Delta t} = \frac{v\sin\phi}{1 - \beta\cos\phi}} \ . \tag{5.19}$$

We can directly draw some conclusions from this equation. The apparent velocity v_{app} is a function of the direction of motion relative to the line-of-sight and of the true velocity of the component. For a given value of v, the maximum velocity v_{app} is obtained if

$$(\sin\phi)_{max} = \frac{1}{\gamma} \ , \tag{5.20}$$

where the *Lorentz factor* $\gamma = (1 - \beta^2)^{-1/2}$ was already defined in (5.4). The corresponding value for the maximum apparent velocity is then

$$\left(v_{app}\right)_{max} = \gamma\, v \ . \tag{5.21}$$

Since γ may become arbitrarily large for values of $v \to c$, the apparent velocity can be much larger than c, even if the true velocity v is—as required by Special Relativity—smaller than c. In Fig. 5.19, v_{app} is plotted as a function of ϕ for different values of the Lorentz factor γ. To get $v_{app} > c$ for an angle ϕ, we need

$$\beta > \frac{1}{\sin\phi + \cos\phi} \geq \frac{1}{\sqrt{2}} \approx 0.707 \ .$$

Hence, superluminal motion is a consequence of the finiteness of the speed of light. Its occurrence implies that source components in the radio jets of AGNs are accelerated to velocities close to the speed of light.

In various astrophysical situations we find that the outflow speeds are of the same order as the escape velocities from the corresponding sources. Examples are the Solar wind, stellar winds in general, or the jets of neutron stars, such as in the famous example of SS433 (in which the jet velocity is $0.26\,c$). Therefore, if the outflow velocity of the jets in AGNs is close c, the jets should originate in a region

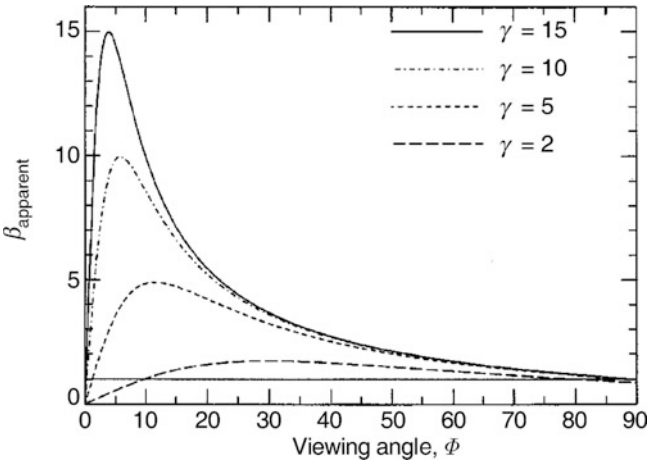


Fig. 5.19 Apparent velocity $\beta_{app} = v_{app}/c$ of a source component moving with Lorentz factor γ at an angle ϕ with respect to the line-of-sight, for four different values of γ. Over a wide range in θ, $\beta_{app} > 1$, thus apparent superluminal motion occurs. The maximum values for β_{app} are obtained if $\sin\theta = 1/\gamma$. Source: C.M. Urry & P. Padovani 1995, *Unified Schemes for Radio-Loud Active Galactic Nuclei*, PASP 107, 803, p. 839, Fig. 21. ©ASP. Reproduced with permission

where the escape velocity has a comparable value. The only objects compact enough to be plausible candidates for this are neutron stars and black holes. And since the central mass in AGNs is considerably larger than the maximum mass of a neutron star, a SMBH is the only option left for the central object. This argument, in addition, yields the conclusion that jets in AGNs must be formed and accelerated very close to the Schwarzschild radius of the SMBH.

The processes that lead to the formation of jets are still subject to intensive research. Most likely magnetic fields play a central role. Such fields may be anchored in the accretion disk, and then spun up and thereby amplified. The wound-up field lines may then act as a kind of spring, accelerating plasma outwards along the rotation axis of the disk (see Sect. 5.5.2 below). In addition, it is possible that rotational energy is extracted from a rotating black hole, a process in which magnetic fields again play a key role. As is always the case in astrophysics, detailed predictions in situations where magnetic fields dominate the dynamics of a system (like, e.g., in star formation) are extremely difficult to obtain because the corresponding coupled equations for the plasma and the magnetic field are very hard to solve.

5.3.4 Further arguments for SMBHs

A black hole is not only the simplest solution of the equations of Einstein's General Relativity, it is also the natural final state of a very compact mass distribution. The occurrence of SMBHs is thus highly plausible from a theoretical point of view. The evidence for the existence of SMBHs in the

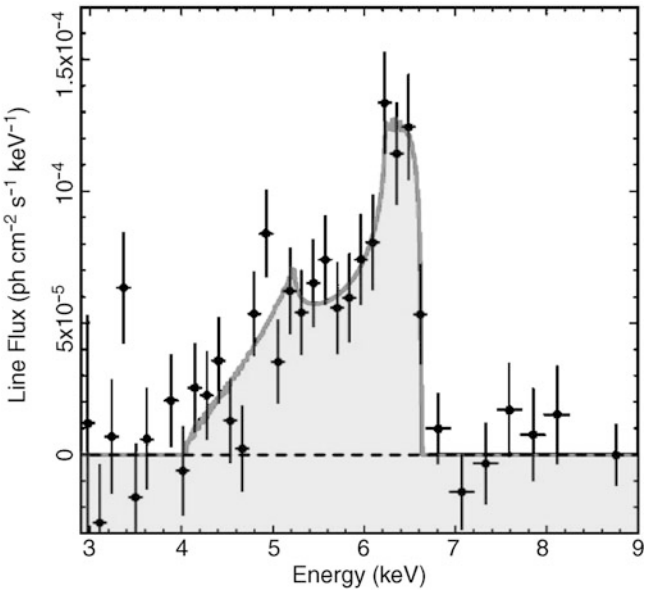


Fig. 5.20 The spectral form of the broad iron line in the Seyfert 1 galaxy MCG-6-30-15 as observed with the ASCA satellite. If the material emitting the line were at rest we would observe a narrow line at $h_P\nu = 6.35\,\text{keV}$. We see that the line is (a) broad, (b) strongly asymmetric, and (c) shifted to smaller energies. A model for the shape of the line, based on a disk around a black hole that is emitting in the radius range $r_S \leq r \leq 20r_S$, is sketched in Fig. 5.21. Source: A.C. Fabian et al. 2000, *Broad Iron Lines in Active Galactic Nuclei*, PASP 112, 1145, Fig. 6. ©ASP. Reproduced with permission

center of galaxies that has been detected in recent years (see Sect. 3.8) provides an additional argument for the presence of SMBHs in AGNs.

Furthermore, we find that the direction of the jets on a milliarcsecond scale, as observed by VLBI, is essentially identical to the direction of jets on much larger scales and to the direction of the corresponding radio lobes. These lobes often have a huge distance from the core, indicating a long lifetime of the source. Hence, the central engine must have some long-term memory because the outflow direction is stable over $\sim 10^7$ yr. A rotating SMBH is an ideal gyroscope, with a direction being defined by its angular momentum vector.

X-ray observations of an iron line of rest energy $h_P\nu = 6.35\,\text{keV}$ in Seyfert galaxies indicates that the emission must be produced in the inner region of an accretion disk, within only a few Schwarzschild radii of a SMBH. An example for this is given in Fig. 5.20. The shape of the line can be explained by a combination of a strong Doppler effect due to high rotation velocities in the disk and by the strong gravitational field of the black hole, as is illustrated in Fig. 5.21.

This iron line is not only detected in individual AGNs, but also in the average spectrum of an ensemble of AGNs. In a deep ($\sim 7.7 \times 10^5$ s) XMM-Newton exposure of the Lockman hole, a region of very low column density of Galactic hydrogen, a large number of AGNs were identified and spectroscopically verified. The X-ray spectrum of these AGNs in the energy ranges of 0.2–3 keV and of 8–20 keV (each in the AGN rest-frame) was modeled by a power law plus intrinsic absorption. The ratio of the measured spectrum of each individual AGN and the fitted model spectrum was then averaged over the AGN population, after transforming the spectra into the rest-frame of the individual sources. As shown in Fig. 5.22, this ratio clearly shows the presence of a strong and broad emission line. The shape of this average emission line can be very well modeled by emission from an accretion disk around a black hole where the radiation originates from a region lying between $\sim$ 3 and $\sim$ 400 Schwarzschild radii. The strength of the iron line indicates a high metallicity of the gas in these AGNs.

The spin of black holes. The spectral shape of the line is affected by the spin of the SMBH. General Relativity predicts that the geometrical properties of space-time around a black hole are determined by its mass and its spin, which affects the properties of the accretion disk in its innermost part as well as the propagation of light rays around the black hole. Furthermore, according to General Relativity, there is a maximum spin a black hole can have. The ratio of the black hole spin to its maximally possible value is called the spin parameter a_{spin}. With sufficiently well-observed spectra, the spin parameter can in fact be estimated, using the model indicated in Fig. 5.21. It is found that a large fraction of SMBHs have a spin parameter $a_{\text{spin}} \gtrsim 0.9$. One would expect this result if the SMBH attained most of its mass through accretion events with almost constant orientation, since accreting matter transfers, beside mass, also angular momentum to the black hole, thus spinning it up. On the other hand, if the mass growth occurred predominantly through merger processes of black holes during the merging of galaxies (see Chap. 10), then smaller values of a_{spin} are expected to result.

5.3.5 A first mass estimate for the SMBH: the Eddington luminosity

Radiation force. As we have seen, the primary energy production in AGNs occurs through accretion of matter onto a SMBH, where the largest part of the energy is produced in the innermost region, close to the Schwarzschild radius. The energy produced in the central region then propagates outwards and can interact with infalling matter by absorption or scattering. Through this interaction of outward-directed radiation with matter, the momentum of the radiation is transferred to the matter, i.e., the infalling matter experiences an outwards-directed radiation force. In order for matter to fall onto the SMBH at all, this radiation force needs to be smaller than the gravitational force. This condition can be

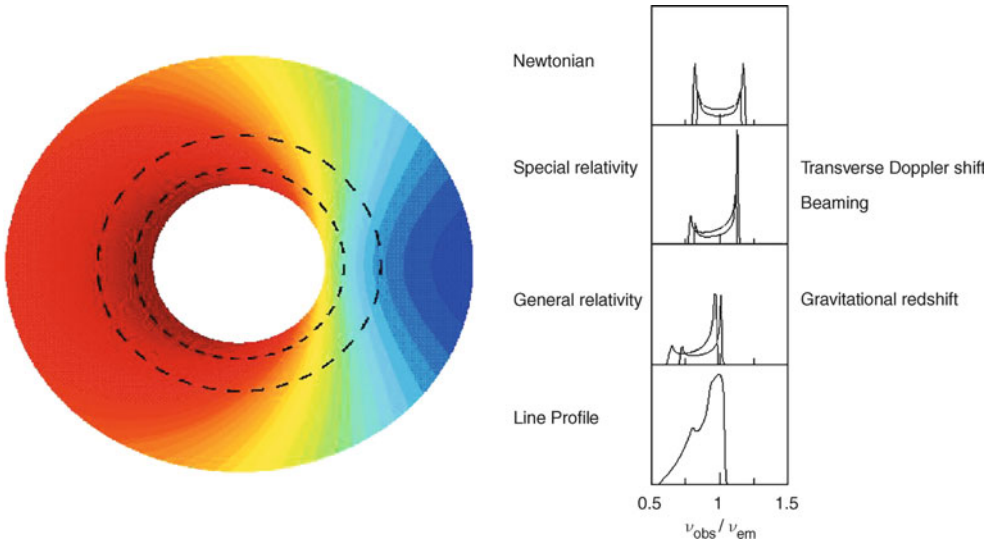


Fig. 5.21 The profile of the broad iron line is caused by a combination of Doppler shift, relativistic beaming, and gravitational redshift. *On the left*, the observed energy of the line as a function of position on a rotating disk is indicated by *colors*. Here, the energy in the right part of the disk which is moving towards us is blueshifted, whereas the left part of the disk emits redshifted radiation. Besides this Doppler effect, all radiation is redshifted because the photons must escape from the deep potential well. The smaller the radius of the emitting region, the larger this gravitational redshift. The line profile we would obtain from a ring-shaped section of the disk (*dashed ellipses*) is plotted in the panels *on the right*. The uppermost panel shows the shape of the line we would obtain if no relativistic effects occurred besides the non-relativistic Doppler effect. Below, the line profile is plotted taking the relativistic Doppler effect and beaming [see (5.37)] into account. This line profile is shifted towards smaller energies by gravitational redshift so that, in combination, the line profile shown at the bottom results. Source: A.C. Fabian et al. 2000, *Broad Iron Lines in Active Galactic Nuclei*, PASP 112, 1145, Fig. 3. ©ASP. Reproduced with permission

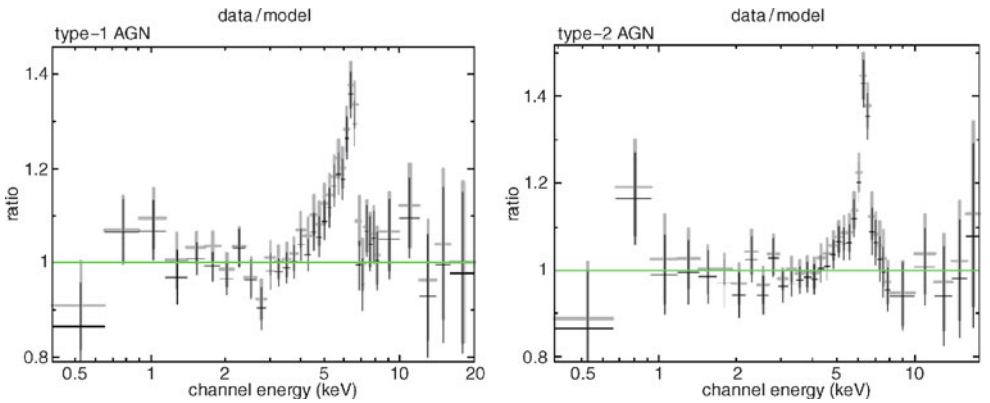


Fig. 5.22 The ratio of the X-ray spectrum of AGNs and a fitted power law averaged over 53 Type 1 AGNs (*left panel*) and 41 Type 2 AGNs (*right panel*). The *gray* and *black* data points are from two different detectors on-board the XMM-Newton observatory. In both AGN samples, a broad relativistic iron line is visible; in the Type 2 AGNs, an additional narrow line component at 6.4 keV can be identified. The line strength indicates that the average iron abundance in these sources is about three times the Solar value. Source: A. Streblyanska et al. 2005, *XMM-Newton observations of the Lockman Hole. III. A relativistic Fe line in the mean X-ray spectra of type-1 and type-2 AGN*, A&A 432, 395, p. 397, Figs. 2, 3. ©ESO. Reproduced with permission

translated into a minimum mass of the SMBH, required for its gravity to dominate the total force at a given luminosity.

We consider a fully ionized gas, so that the interaction of radiation with this infalling plasma is basically due to scattering of photons by free electrons. This is called *Thomson scattering*. The mean radiation force on an electron at radius r is then

$$F_{\rm rad} = \sigma_{\rm T} \frac{L}{4\pi\, r^2\, c}\,, \tag{5.22}$$

where

$$\sigma_{\rm T} = \frac{8\pi}{3}\left(\frac{e^2}{m_{\rm e}c^2}\right)^2 = 6.65\times 10^{-25}\,{\rm cm}^2 \tag{5.23}$$

denotes the *Thomson cross section* (in cgs units). This cross section is independent of photon frequency.[5] To derive (5.22), we note that the flux $S = L/(4\pi r^2)$ is the radiation energy which flows through a unit area at distance r from the central source per unit time. Then S/c is the momentum of photons flowing through this unit area per time, or the radiation pressure, because the momentum of a photon is given by its energy divided by the speed of light. Thus the momentum transfer to an electron per unit time, or the radiation force, is given by $\sigma_T S/c$. From (5.22), we can see that the radiation force has the same dependence on radius as the gravitational force, $\propto r^{-2}$, so that the ratio of the two forces is independent of radius.

Eddington luminosity. For matter to be able to fall in—the condition for energy production—the radiation force must be smaller than the gravitational force. For each electron there is a proton, and these two kinds of particles are electromagnetically coupled. The gravitational force per electron-proton pair is given by

$$F_{\rm grav} = \frac{GM_\bullet m_{\rm p}}{r^2} \,.$$

where we have neglected the mass of the electron since it is nearly a factor of 2000 smaller than the proton mass $m_{\rm p}$. Hence, the condition

$$F_{\rm rad} < F_{\rm grav} \tag{5.24}$$

for the dominance of gravity can be written as

$$\frac{\sigma_T L}{4\pi\, r^2\, c} < \frac{GM_\bullet m_{\rm p}}{r^2} \,,$$

or

$$\boxed{L < \frac{4\pi G\, c\, m_{\rm p}}{\sigma_T} M_\bullet =: L_{\rm edd} \approx 1.26\times 10^{38} \left(\frac{M_\bullet}{M_\odot}\right) {\rm erg/s}} \,, \tag{5.25}$$

where we have defined the *Eddington luminosity* $L_{\rm edd}$ *of a black hole of mass* $M_\bullet$. Since σ_T is independent of photon frequency, the luminosity referred to above is the bolometric luminosity.

A lower limit on $M_\bullet$. For accretion to occur at all, we need $L < L_{\rm edd}$. Remembering that the Eddington luminosity is proportional to $M_\bullet$, we can turn the above argument around: if a luminosity L is observed, we conclude $L_{\rm edd} > L$, or

$$\boxed{M_\bullet > M_{\rm edd} := \frac{\sigma_T}{4\pi G c m_{\rm p}} L \approx 8\times 10^7 \left(\frac{L}{10^{46}\,{\rm erg/s}}\right) M_\odot} \,. \tag{5.26}$$

Therefore, a lower limit for the mass of the SMBH can be derived from the luminosity. For luminous AGNs, like QSOs, typical masses are $M_\bullet \gtrsim 10^8 M_\odot$, while Seyfert galaxies have lower limits of $M_\bullet \gtrsim 10^6 M_\odot$. Hence, the SMBH in our Galaxy could in principle provide a Seyfert galaxy with the necessary energy.

In the above definition of the Eddington luminosity we have implicitly assumed that the emission of radiation is isotropic. In principle, the above argument of a maximum luminosity can be avoided, and thus luminosities exceeding the Eddington luminosity can be obtained, if the emission is highly anisotropic. A geometrical concept for this would be, for example, accretion through a disk in the equatorial plane and the emission of a major part of the radiation along the polar axes. Models of this kind have indeed been constructed. It was shown that the Eddington limit may be exceeded by this, but not by a large factor. However, the possibility of anisotropic emission has another very important consequence. To derive a value for the luminosity from the observed flux of a source, the relation $L = 4\pi D_{\rm L}^2\, S$ is applied, which is explicitly based on the assumption of isotropic emission. But if this emission is anisotropic and thus depends on the direction to the observer, the true luminosity may differ considerably from that which is derived under the assumption of isotropic emission. Later we will discuss the evidence for anisotropic emission in more detail.

Eddington accretion rate. If the conversion of infalling mass into energy takes place with an efficiency ϵ [see (5.15)], the accretion rate can be determined,

$$\dot m = \frac{L}{\epsilon\, c^2} \approx 0.18\, \frac{1}{\epsilon} \left(\frac{L}{10^{46}\,{\rm erg/s}}\right) M_\odot/{\rm yr} \,. \tag{5.27}$$

[5] When a photon scatters off an electron at rest, this process is called Thomson scattering. To a first approximation, the energy of the photon is unchanged in this process, only its direction is different after scattering. This is not really true, though. Due to the fact that a photon with energy E_γ carries a momentum E_γ/c, scattering will impose a recoil on the electron. After the scattering event the electron will thus have a non-zero velocity and a corresponding kinetic energy. Owing to energy conservation the photon energy after scattering is therefore slightly smaller than before. This energy loss of the photon is very small as long as $E_\gamma \ll m_{\rm e}c^2$. When this energy loss becomes appreciable, this scattering process is then called Compton scattering. If the electron is not at rest, the scattering can also lead to net energy transfer to the photon, such as it happens when low-frequency photons propagate through a hot gas (as we will discuss in Sect. 5.4.4 for the case of AGNs, and in Sect. 6.4.4 for galaxy clusters) or through a distribution of relativistic electrons. In this case one calls it the inverse Compton effect. The physics of all these effects is the same, only their kinematics are different.

Since the maximum efficiency is of order $\epsilon \sim 0.1$, this implies accretion rates of typically several Solar masses per year for very luminous QSOs. If L is measured in units of the Eddington luminosity, we obtain with (5.25)

$$\dot{m} = \frac{L}{L_{\rm edd}} \left(\frac{1.26 \times 10^{38}\,{\rm erg/s}}{\epsilon c^2} \right) \left(\frac{M_\bullet}{M_\odot} \right) \equiv \frac{L}{L_{\rm edd}}\, \dot{m}_{\rm edd}\,, \tag{5.28}$$

where in the last step the Eddington accretion rate has been defined,

$$\dot{m}_{\rm edd} = \frac{L_{\rm edd}}{\epsilon c^2} \approx \frac{1}{\epsilon}\, 2 \times 10^{-9}\, M_\bullet\, {\rm yr}^{-1}\,. \tag{5.29}$$

Growth rate of the SMBH mass. The Eddington accretion rate is the maximum accretion rate if isotropic emission is assumed, and it depends on the assumed efficiency ϵ. We can now estimate a characteristic time in which the mass of the SMBH will significantly increase,

$$t_{\rm evo} := \frac{M_\bullet}{\dot{m}} \approx \epsilon \left(\frac{L}{L_{\rm edd}} \right)^{-1} 5 \times 10^8\,{\rm yr}\,, \tag{5.30}$$

i.e., even with efficient energy production ($\epsilon \sim 0.1$), the mass of a SMBH can increase greatly on cosmologically short time-scales by accretion (see problem 5.4). However, this is not the only mechanism that can produce SMBHs of large mass. They can also be formed through the merger of two black holes, each of smaller mass, as would be expected after the merger of two galaxies if both partners hosted a SMBH in its center. This aspect will be discussed more extensively later.

5.4 Components of an AGN

In contrast to stars, which have a simple geometry, we expect several source components in AGNs with different, sometimes very complex geometric configurations to produce the various components of the spectrum; this is sketched in Fig. 5.23. Accretion disks and jets in AGNs are clear indicators for a significant deviation from spherical symmetry in these sources. The relation between source components and the corresponding spectral components is not always obvious. However, combining theoretical arguments with detailed observations has led to quite satisfactory models.

5.4.1 The IR, optical, and UV-continuum

In Sect. 5.3.2 we considered an accretion disk with a characteristic temperature, following from (5.14), of

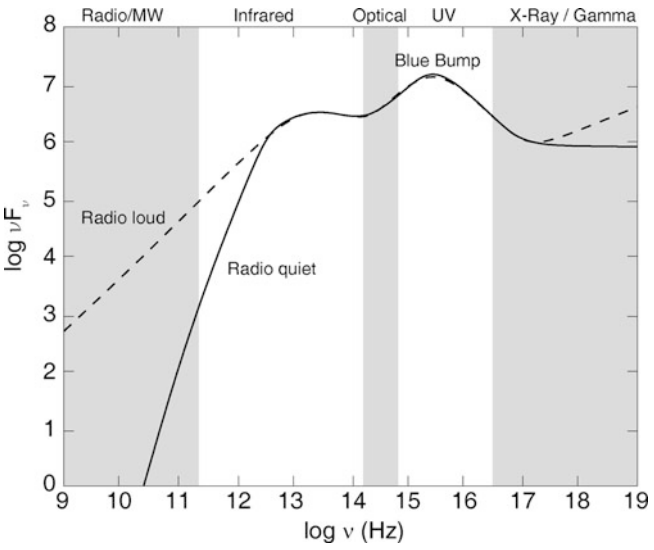


Fig. 5.23 Sketch of the characteristic spectral emission of a QSO. We distinguish between radio-loud (*dashed curve*) and radio-quiet (*solid curve*) QSOs. Plotted is νS_ν (in arbitrary units), so that flat sections in the spectrum correspond to equal energy per logarithmic frequency interval. The most prominent feature is the big blue bump, a broad maximum in the UV up to the soft X-ray domain of the spectrum. Besides this maximum, a less prominent secondary maximum is found in the IR. The spectrum increases towards higher energies in the X-ray domain of the spectrum—typically $\sim 10\,\%$ of the total energy is emitted as X-rays. For blazars, the spectrum can rise at even higher energies, yielding a large fraction of the total flux being radiated at gamma-rays

$$T(r) \approx 6.3 \times 10^5\,{\rm K} \left(\frac{\dot{m}}{\dot{m}_{\rm edd}} \right)^{1/4} \left(\frac{M_\bullet}{10^8 M_\odot} \right)^{-1/4} \left(\frac{r}{r_{\rm S}} \right)^{-3/4}. \tag{5.31}$$

The thermal emission of an accretion disk with this radial temperature profile produces a broad spectrum with its maximum in the UV. The continuum spectrum of QSOs indeed shows an obvious increase towards UV wavelengths, up to the limit of observable wavelengths, $\lambda \gtrsim 1000\,$Å. (This is the observed wavelength; QSOs at high redshifts can be observed at significantly shorter wavelengths in the QSO rest-frame.) At wavelengths $\lambda \leq 912\,$Å, photoelectric absorption by neutral hydrogen in the ISM of the Galaxy sets in, so that the Milky Way is opaque for this radiation. Only at considerably higher frequencies, namely in the soft X-ray band ($h_{\rm P}\nu \gtrsim 0.2\,$keV), does the extragalactic sky become observable again.

If the UV radiation of a QSO originates mainly from an accretion disk, which can be assumed because of the observed increase of the spectrum towards the UV, the question arises whether the thermal emission of the disk is also visible in the soft X-ray regime. In this case, the spectrum in the range hidden from observation, at $13\,{\rm eV} \lesssim h_{\rm P}\nu \lesssim 0.2\,$keV, could be interpolated by such an accretion

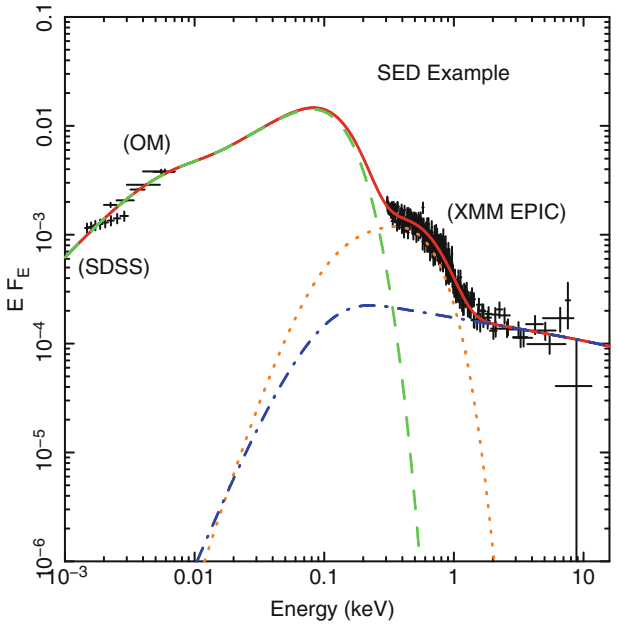


Fig. 5.24 A possible interpretation for the connection between the optical-UV spectrum and the X-ray spectrum of an AGN, interpolating through the unobservable spectral region between 13.6 eV and $\sim$ 0.2 keV. The total spectral energy distribution (*red curve*) in this model is composed of the accretion disk emission (*green dashed curve*) and the X-ray emission through low-temperature Comptonization with high optical depth (*orange dotted curve*) and high-temperature Comptonization at low optical depth (*blue dot-dashed curve*). Source: C. Jin et al. 2012, *A combined Optical and X-ray Spectra Study for Type 1 AGN. III. Broadband SED Properties*, MNRAS 425, 907, Fig. 1. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

disk spectrum. This seems indeed to be the case. The X-ray spectrum of QSOs often shows a very simple spectral shape in the form of a power law, $S_\nu \propto \nu^{-\alpha}$, where $\alpha \sim 0.7$ is a characteristic value (see Sect. 5.4.4 below). However, the spectrum follows this power law only at energies down to $\sim$ 0.5 keV. At lower energies, the spectral flux can be higher than predicted by the extrapolation of the power-law spectrum observed at higher energies. One interpretation of this finding is that the (non-thermal) source of the X-ray emission produces a simple power law, and the additional flux at lower X-ray energies is emission from the innermost part of the accretion disk (see Fig. 5.23).

Perhaps these two spectral properties—the increase of the spectrum towards the UV and the radiation excess in the soft X-ray—have the same origin, being two wings of a broad maximum in the energy distribution, which itself is located in the spectral range unobservable for us. This maximum is called the *big blue bump* (BBB). A description of the BBB is possible using detailed models of accretion disks (see Fig. 5.24 for an example). For this modeling, however, the assumption of a local Planck spectrum at all radii of the disk is too simple because the structure of the accretion disk is more complicated. The spectral properties of an accretion disk have to be modeled by an 'atmosphere' for each radius, similar to that in stars. Indeed, since the radial temperature distribution (5.31) extends to soft X-ray energies only for very low $M_\bullet$, the X-rays (and the BBB) are probably not due to the thermal disk emission, but originates from a hot atmosphere (corona) of the disk, as will be discussed in more detail in Sect. 5.4.4.

Besides the BBB, an additional maximum exists in the MIR (IR-bump). This can be ascribed to thermal emission of warm dust ($T \lesssim 1000$ K). As we will discuss below, other observations provide additional evidence for this dust component, which may be associated with the absorbing torus (see Fig. 5.12).

The optical continuum of blazars is different from that of Seyfert galaxies and QSOs. It often features a spectral pattern that follows, to very good approximation, a power law and is strongly variable and polarized. This indicates that the radiation is predominantly non-thermal. The origin of this radiation thus probably does not lie in an accretion disk. Rather, the radiation presumably has its origin in the relativistic jets which we already discussed for the radio domain, with their synchrotron radiation extending up to optical wavelengths. This assumption was strongly supported by many sources where observations discovered optical and X-ray emission from jets (see Fig. 5.17 and Sect. 5.5.4).

Gravitational microlensing: Microscopy of the accretion disk. In Sect. 2.5 we discussed the Galactic microlensing effect, where a star is lensed by a compact object in our Milky Way. The observational signature of this effect is the flux variation of the source, which occurs due to a time-varying magnification of the background star caused by the lens.

In a strong lensing event where a QSO is mapped into several images by a foreground galaxy (see Sect. 3.11), another kind of microlensing can occur (and in fact had been investigated before Galactic microlensing was discussed). One needs to realize that the mass distribution in a galaxy is composed of a baryonic and a dark matter component, the former being dominated by stars. Thus, the mass distribution responsible for the gravitational light deflection is 'grainy', possessing small-scale structure. Depending on the size of the source being lensed, this graininess can be relevant.

To see why, we first consider the typical length scales of the situation. Consider first a single star of mass M in the lens; then the Einstein angle θ_E of this star is given by (2.82). The corresponding length-scale in the source plane is obtained by projecting this angle onto the source plane,

$$R_\mathrm{E} = D_\mathrm{s}\theta_\mathrm{E} = \sqrt{\frac{4GM}{c^2}\frac{D_\mathrm{s}D_\mathrm{ds}}{D_\mathrm{d}}} \sim 9\times10^{16}\,\mathrm{cm}\sqrt{\frac{M}{M_\odot}}\sqrt{\frac{D_\mathrm{s}}{c/H_0}}\sqrt{\frac{D_\mathrm{ds}}{D_\mathrm{d}}}, \tag{5.32}$$

where we assumed a Hubble constant of $h = 0.7$. Since the angular-diameter distance of QSOs is of order the Hubble radius, and the lens is somewhere in the middle between us and the source, the last two factors in (5.32) are of order unity. The typical stellar mass in an early-type galaxy is of order $M_\odot/2$; thus we conclude that the Einstein radius in the source plane is of the order of a few times 10^{16} cm, or about 10 light-days. Sources of that size or smaller can be significantly magnified by the lensing effect of the star.

This length-scale can be compared to typical sizes of QSOs. The rapid variability of QSOs in the X-rays implies that the X-ray emitting source is typically smaller than R_E; the same applies to the source component emitting UV-radiation. As we shall see below, the size of the region where the broad emission lines originate is typically larger than R_E. From these size considerations, we expect that the UV- and X-ray emission of QSOs can be magnified by stars in the lens galaxy.

There is a major difference between Galactic microlensing and QSO microlensing. We saw in Sect. 3.11 that the probability that a lens in our Galaxy is located close to the line-of-sight to a distant star is tiny—that is one of the reasons why Galactic microlensing surveys are so difficult. This is no longer true in QSO microlensing; the density of stars in the lens galaxy at the location where the multiple images occur is rather large. In fact, the mean separation between stars is not much larger than the Einstein radius of each star, somewhat depending on the local fraction of the surface mass density contained in stars compared to that in the form of dark matter. The consequence of this high microlens density is that the individual stars can no longer be considered as isolated point-mass lenses. Instead, an ensemble of microlenses needs to be considered whose joint lensing action causes the microlensing effect. This phenomenon can be studied with numerical simulations, by 'shooting' light rays through an ensemble of point-mass lenses.

In the upper part of Fig. 5.25, typical magnification patterns are shown, for two different values of the stellar density in the line-of-sight. The magnification pattern is a plot of the magnification $\mu(\boldsymbol{\beta})$ as a function of source position $\boldsymbol{\beta}$. The magnification of a source thus depends on its location relative to the stellar field of the lens (the stars were assumed to be randomly placed in the lens), as well as on the size of the source: The magnification of an extended source is given as the average of the point-source magnification across the brightness profile of the source. We see that characteristic lines of high magnification occur, the so-called caustics, that we already saw for the case of binary lenses (see Fig. 2.40 for an example).

Since the system of source, lens and observer has a relative motion transverse to the line-of-sight, the relative location of the source in the magnification pattern will change in time. Hence, the magnification of each image of a source varies due to microlensing. Synthetic light curves for two different source sizes are shown in the lower part of Fig. 5.25, for the two stellar densities. We see that in the low-density case, strong variations are rare and occur only when the source crosses a caustic. For the high-density case, the frequency of caustic crossings is much higher, and the periods of small flux variations become shorter.

Furthermore, we see from the light curves that there is another relevant scale in the problem: the fastest magnification variations are defined by the time it takes the source to cross a caustic. This time-scale is reflected in the sharp rise (or fall) of the magnification during caustic crossing. Thus, the rise/fall-time is given by the size of the source divided by the transverse velocity. The time-variable magnification of course corresponds to a change of the observed flux of the images.

Indeed, flux monitoring of multiply-imaged QSOs are carried out, predominantly to determine the time delays in lens systems (see Sect. 3.11.4). After accounting for the relative time delay, the light curves of multiply-imaged QSOs are not identical, but they vary independently of each other. This cannot be due to a variation of the source flux—as that would show up in all images with the respective time delay—but must be due to microlensing. Since we have good estimates for typical peculiar cosmic velocities, we can relate time-scales of variations to length-scales in the source plane, and thus estimate the size of the emission region (at a given spectral band).

Results. For any single lens system, the quantitative analysis of microlensing light curves is hampered by uncertainties about the transverse velocity needed to relate observed time-scales to intrinsic length scales. Thus, for any parameter to be estimated from the light curves, one obtains a probability distribution—also because there is stochasticity in the spatial location of the stars. For the lens system $2237 + 0305$, where the lens is at low redshift ($z_\mathrm{d} \approx 0.04$), the effective transverse velocity is probably dominated by that of the Solar System, well known through the CMB dipole. The optical light curves of this system (Fig. 5.26) show pronounced microlensing in all four images. From matching the observed light curves with synthetic ones from ray-shooting simulations, the size of the emitting region can then be estimated. In the left panel of Fig. 5.27, the solid red curve shows the probability distribution of the half-light radius of the optical emission region, here corresponding to a rest wavelength of $\lambda \approx 2000$ Å. The characteristic scale thus obtained is $R \approx 3 \times 10^{16}$ cm. From the available X-ray light curves of this lens system, the size of the X-ray emitting region can be estimated as well, as shown by the other two curves in this figure. The right-hand panel of Fig. 5.27 shows the estimated optical (filled black squares) and (for some systems with

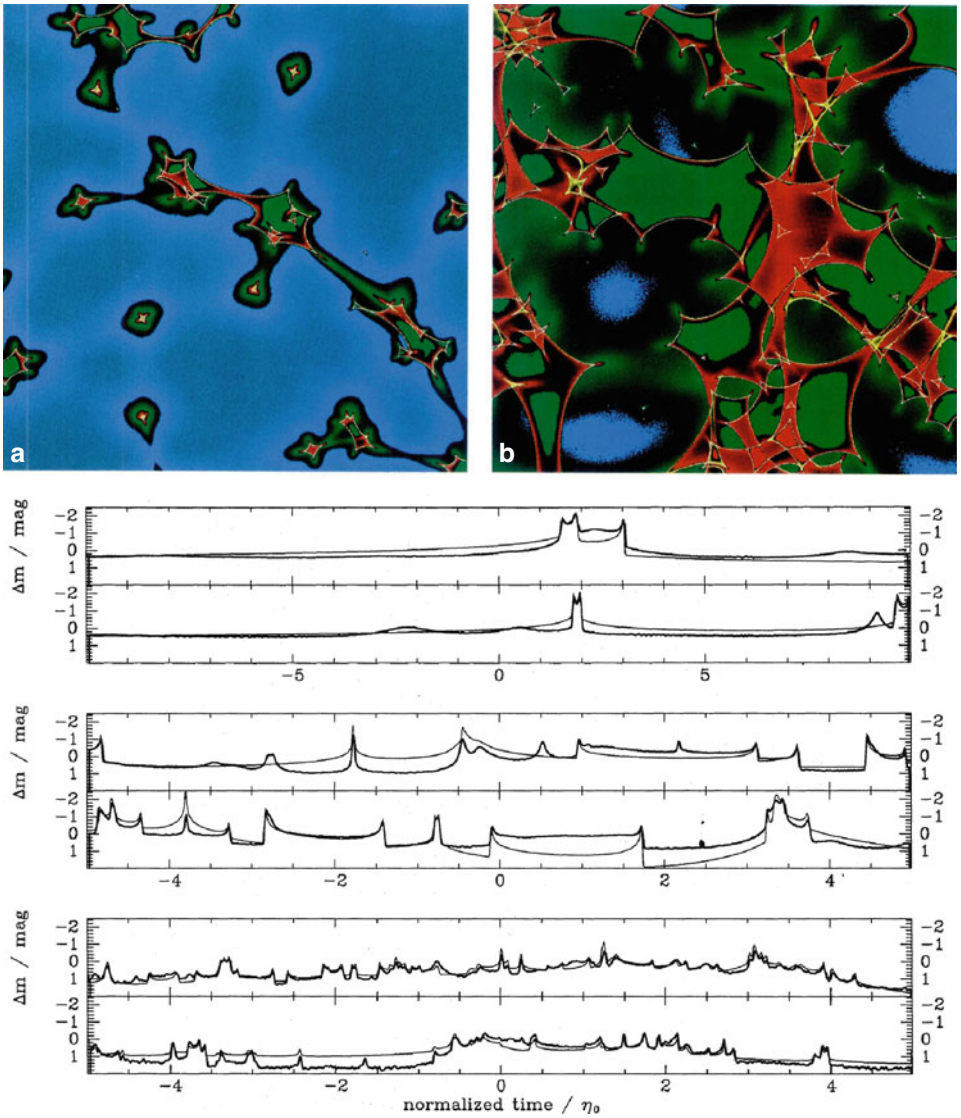


Fig. 5.25 The *upper panels* show magnification patterns from microlensing by stars in a lens galaxy, assuming a relatively low (*left*) or high (*right*) density of microlenses in the line-of-sight. Shown is the magnification as a function of position of a (small) source in the source plane, with *blue* indicating a low magnification, and increasingly higher ones are indicated in *green*, *red* and *yellow*. In the low-density case, one can identify several rather isolated point masses with their characteristic, almost axi-symmetric magnification pattern. However, many regions of larger magnification are connected. In the high-density case, there are no longer isolated point masses, and a complex pattern of magnification occurs. The highest magnifications occur along lines, the so-called caustics (see also Fig. 2.40). In the *lower two panels*, synthetic light curves, Δm as a function of time, normalized by the time is takes the source to move by one Einstein radius of the microlenses, are plotted as they occur if a source moves through the magnification patterns shown in the *upper panels*; the upper (lower) set of light curves correspond to the low (high) density microlensing field. In each case, light curves for two different source sizes are plotted, with the smoother curve corresponding to the large source. Source: J. Wambsganss et al. 1992, *Gravitational microlensing - Powerful combination of ray-shooting and parametric representation of caustics*, A&A 258, 591, Figs. 2,3. ©ESO. Reproduced with permission

available X-ray data) X-ray half-light radius as a function of estimated black hole mass, as obtained from microlensing studies.

With a characteristic size of the emitting region of $\sim 10^{16}$ cm, corresponding to about micro-arcsecond angular resolution (thus the name 'microlensing'), these microlensing studies yield the highest resolution observations available, almost three orders-of-magnitude higher than VLBI. The size of the optical source scales with black hole mass roughly as

$$R_{1/2} \approx 6 \times 10^{15} \left(\frac{M_\bullet}{10^9 M_\odot} \right)^{0.8} , \qquad (5.33)$$

which is shown as dashed curve in the right panel of Fig. 5.27. Given the uncertainty in the parameters of this

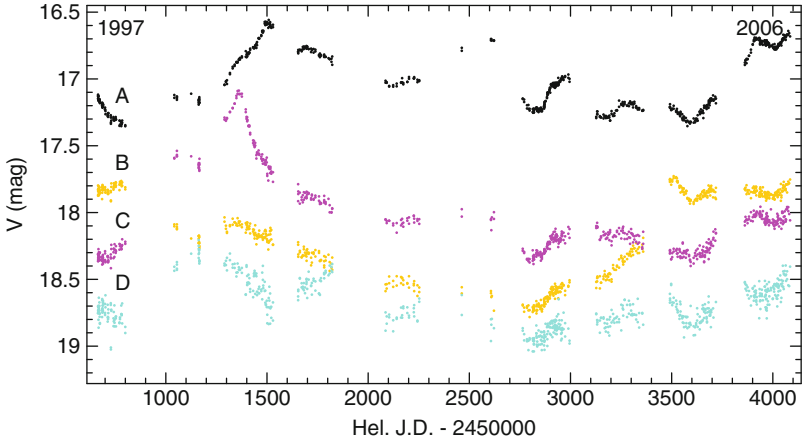


Fig. 5.26 Light curves of the four images of the lens system 2237 + 0305 (Fig. 3.59), taken over ten seasons 1997–2006. The time delay between the images is less than a day, thus all uncorrelated variability is due to microlensing. Source: A. Udalski et al. 2006, *The Optical Gravitational Lensing Experiment. OGLE-III Long Term Monitoring of the Gravitational Lens QSO 2237 + 0305*, Acta Astronomica 56, 293, p. 303. Reproduced by permission of the Copernicus Foundation for Polish Astronomy

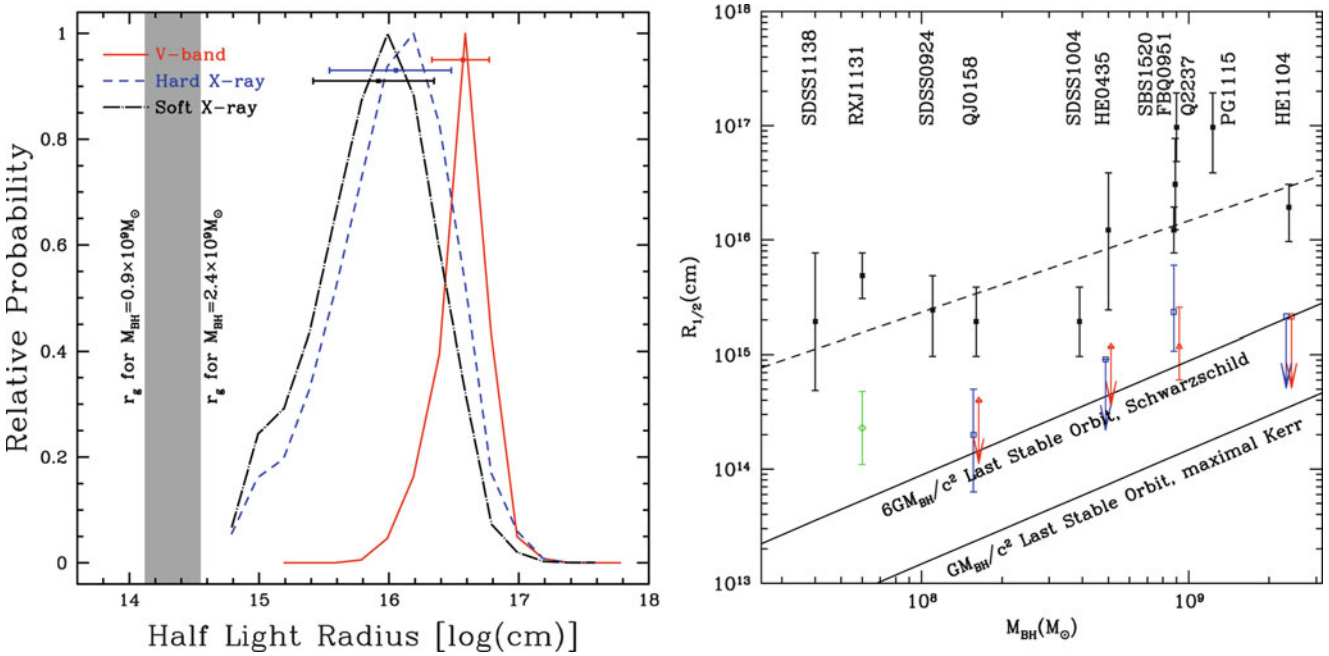


Fig. 5.27 *Left panel:* From an analysis of the microlensing light curves of the lens system 2237 + 0305, the probability distribution for the half-light radius is shown, for the optical emission region by the *solid red curve*, and the soft and hard X-ray emission by the *dash-dotted black* and *dashed blue curves*, respectively. The *grey band* indicates the 'gravitational radius' $r_g = GM_\bullet/c^2$, i.e., half the Schwarzschild radius, for a range of black hole masses between 0.9 and 2.4 × $10^9 M_\odot$, as estimated from the broad emission lines (see Sect. 5.4.2). *Right panel:* Estimates of the half-light radius for 11 multiply-imaged QSOs, plotted against the estimated black hole mass. *Filled black squares* correspond to the optical emitting region, the *other symbols* to X-ray emission, where *arrows* indicate upper limits. The *two solid lines* show the size of the last stable circular orbit for a black hole without spin (Schwarzschild black hole) and one with maximum spin parameter ('maximal Kerr'). Source: A. Mosquera et al. 2013, *The Structure of the X-Ray and Optical Emitting Regions of the Lensed Quasar Q 2237 + 0305*, ApJ 769, 53, p. 5, 7, Figs. 4, 8. ©AAS. Reproduced with permission

power-law fit, this scaling of emission size with $M_\bullet$ is compatible with our model of accretion disks: for a fixed temperature, (5.13) predicts $r^3 \propto \dot{m} M_\bullet \propto (\dot{m}/\dot{m}_{\rm edd})\epsilon^{-1} M_\bullet^2$, where we scaled the accretion rate by the Eddington rate (5.29). Thus, for a given wavelength, standard accretion disk theory predicts $r \propto M_\bullet^{2/3}$, not very different from the scaling in (5.33). However, taken at face value, the prefactor in (5.33) implies that the efficiency ϵ is surprisingly

small, $\epsilon \sim 2\,\%$, compared to the usually expected $\sim 10\,\%$. Whereas the estimate is affected by the aforementioned stochasticity of the effect, as well as factors like inclination angle distribution of the accretion disks, this difference may indicate a deviation from the simple picture of the optical emission coming solely from an accretion disk – for example, some fraction of the optical light may be scattered by gas at larger radius, thus effectively increasing the source size.

Furthermore, we see from Fig. 5.27 that the emission region of X-rays is significantly smaller than the optical emitting region. In fact, the estimated size of the X-ray region is comparable to the radius of the last stable circular orbit in a non-rotating black hole, $R = 3r_S = 6GM_\bullet/c^2$. Hence, in agreement with the size estimates from the broad iron line (Sect. 5.3.4), the X-rays originate from the very central region around the SMBH. Indeed, the estimated X-ray size is compatible with a scaling $R_{1/2} \propto M_\bullet$.

To conclude, microlensing allows us to study the smallest scales in AGNs, and the estimated sizes will become more accurate with increased length of the observed light curves and increasing the sample size.

5.4.2 The broad emission lines

Characteristics of the broad line region. One of the most surprising characteristics of AGNs is the presence of very strong emission lines, except for BL Lac objects which show an almost featureless spectrum. Furthermore, in Type 1 AGN, such as QSOs and Seyfert 1s, the emission lines are very broad. Interpreted as Doppler velocities, the corresponding width of the velocity distribution of the components in the emitting region is of order $\Delta v \lesssim 10\,000\,\mathrm{km/s}$ (or $\Delta\lambda/\lambda \lesssim 0.03$). These lines cannot be due to thermal line broadening because that would imply $k_B T \sim m_p(\Delta v)^2/2 \sim 1\,\mathrm{MeV}$, or $T \sim 10^{10}\,\mathrm{K}$—no emission lines would be produced at such high temperatures because all atoms would be fully ionized (plus the fact that at such temperatures a plasma would efficiently produce e^+e^--pairs, and the corresponding annihilation line at 511 keV should be observable in Gamma radiation). Therefore, the observed line width is interpreted as Doppler broadening. The gas emitting these lines then has large-scale velocities of order $\sim 10\,000\,\mathrm{km/s}$. Velocities this high are indicators of the presence of a strong gravitational field, as would occur in the vicinity of a SMBH. If the emission of the lines occurs in gas at a distance r from a SMBH, we expect characteristic velocities of

$$v_{\mathrm{rot}} \sim \sqrt{\frac{GM_\bullet}{r}} = \frac{c}{\sqrt{2}} \left(\frac{r}{r_S}\right)^{-1/2} ,$$

so for velocities of $v \sim c/30$, we obtain a radial distance of $r \sim 500\, r_S$.

Hence, the Doppler broadening of the broad emission lines can be produced by Kepler rotation at radii of about $500\, r_S$. Although this estimate is based on the assumption of a rotational motion, the infall velocity for free fall does not differ by more than a factor $\sqrt{2}$ from this rotational velocity. Thus the kinematic state of the emitting gas is of no major relevance for this rough estimate if only gravity is responsible for the occurrence of high velocities.

The region in which the broad emission lines are produced is called the *broad line region* (BLR). The density of the gas in the BLR can be estimated from the lines that are observed. To see this, it must be pointed out that allowed and semi-forbidden transitions are found among the broad lines. Examples of the former are Lyα, MgII, and CIV, whereas CIII] and NIV] are semi-forbidden transitions. However, forbidden transitions are essentially absent among the broad lines.[6]

An excited atom can transit into its ground state (or another lower-lying state) either by spontaneous emission of a photon or by losing energy through collisions with other atoms. The probability for a radiative transition is defined by the atomic parameters, whereas the collisional de-excitation depends on the gas density. If the density of the gas is high, the mean time between two collisions is much shorter than the average lifetime of forbidden or semi-forbidden radiative transitions. Therefore the corresponding line photons are not observed.[7] The absence of forbidden lines is then used to derive a lower limit for the gas density, and the occurrence of semi-forbidden lines yields an upper bound for the density. To minimize the dependence of this argument on the chemical composition of the gas, transitions of the same element are preferentially used for these estimates. However, this is not always possible. From the presence of the CIII] line and the non-existence of the [OIII] line in the BLR, combined with model calculations, a density estimate of $n_e \sim 3 \times 10^9\,\mathrm{cm}^{-3}$ is obtained. However, as we shall see shortly, the conditions in the BLR are not uniform, but the BLR extends over a range of scales. The CIII] line originates

[6] The classification into allowed, semi-forbidden, and forbidden transitions is done by means of quantum-mechanical transition probabilities, or the resulting mean time for a spontaneous radiational transition. Allowed transitions correspond to electric dipole radiation, which has a large transition probability, and the lifetime of the excited state is then typically only 10^{-8} s. For forbidden transitions, the time-scales are considerably larger, typically 1 s, because their quantum-mechanical transition probability is substantially lower. Semi-forbidden transitions have a lifetime between these two values. To mark the different kinds of transitions, a double square bracket is used for forbidden transitions, like in [OIII], while semi-forbidden lines are marked by a single square bracket, like in CIII].

[7] To make forbidden transitions visible, the gas density needs to be very low. Such low densities cannot be produced in the laboratory. Forbidden lines are in fact not observed in laboratory spectra; they are 'forbidden'.

from rather large radii. In the inner-most part of the BLR, the electron density is higher, $n_e \sim 10^{11}$ to $10^{12}\,\mathrm{cm}^{-3}$

Furthermore, from the ionization stages of the line-emitting elements, a temperature can be estimated, typically yielding $T \sim 20\,000$ K. Detailed photoionization models for the BLR are very successful and are able to reproduce details of line ratios very well.

From the density of the gas and its temperature, the emission measure can then be calculated (i.e., the number of emitted line photons per unit time and per unit volume element). From the observed line strength and the distance to the AGN, the total number of emitted line photons can be calculated, and by dividing through the emission measure, the volume of the line-emitting gas can be determined. This estimated volume of the gas is much smaller than the total volume ($\sim r^3$) of the BLR. We therefore conclude that the BLR is not homogeneously filled with gas; rather, the gas has a very small filling factor. The gas in which the broad lines originate fills only a small fraction (estimates range from $\sim 10^{-7}$ to ~ 0.1) of the total volume of the BLR; hence, it must be concentrated in clouds.

Geometrical picture of the BLR. From the previous considerations, a picture of the BLR emerges in which it contains gas clouds with a characteristic particle density of $n_e \sim 10^{10}\,\mathrm{cm}^{-3}$. In these clouds, heating and cooling processes take place. Probably the most important cooling process is the observed emission of broad emission lines, with $\sim 25\,\%$ of the cooling due to the iron lines (see Fig. 5.3). Heating of the gas is provided by energetic continuum radiation from the AGN which photoionizes the gas, similar to processes in Galactic gas clouds. The difference between the energy of a photon and the ionization energy yields the energy of the released electron, which is then thermalized by collisions and leads to gas heating. In a stationary state, the heating rate equals the cooling rate, and this equilibrium condition defines the temperature which the clouds attain.

The comparison of continuum radiation and line emission yields the fraction of ionizing continuum photons which are absorbed by the BLR clouds; a value of about $10\,\%$ is obtained. Since the clouds are optically thick to ionizing radiation, the fraction of absorbed continuum photons is also the fraction of the solid angle subtended by the clouds, as seen from the central continuum source. From the filling factor and this solid angle, the characteristic size of the clouds can be estimated, from which we obtain typical values of $\sim 10^{11}$ to $\sim 10^{14}$ cm, depending on the filling factor. In addition, based on these arguments, the number of clouds in the BLR can be estimated. This yields a typical value of $\sim 10^{10}$. An independent argument for a very large number of clouds comes from the fact that the observed line profiles are very smooth. Since the width of the emission line from an individual cloud is very much smaller (of order $\sim 20\,\mathrm{km/s}$ as thermal Doppler width for temperatures of $T \sim 2 \times 10^4$ K) than the observed line width, the broad lines must be produced by the superposition of the line emission from many clouds. If the number of clouds is not much larger than the ratio of observed line width to individual cloud line width, then the observed emission line will contain a kind of 'Poisson noise'; the smoothness of the lines clearly shows that this noise must be small, yielding a lower limit of $\sim 10^7$ clouds in the BLR.[8]

The characteristic velocity of the clouds corresponds to the line width, hence several thousand km/s. However, the kinematics of the clouds is unknown. We do not know whether they are rotating around the SMBH, whether they are infalling or streaming outwards, or whether their motion is rather chaotic. Reasonably good arguments are found for several of these possibilities. The fact that higher ionization lines exhibit a blueshift relative to the narrow emission lines may indicate an outflow of the BLR gas, with the part of the gas streaming away from us being covered by the accretion disk—so that we can see predominantly that part of the flow moving towards us. However, the systematic blueshift could also be produced by extinction of the red wing of the emission line, which would argue for an inflow of the absorbing material. There are some Type 1 AGNs which show broad Balmer emission lines with a double peak, such as would be produced if the emitting region was in a disk-like flow. It is also possible that different regions within the BLR exist with different kinematic properties.

Reverberation mapping: the principle. A direct method to examine the extent of the BLR is provided by *reverberation mapping*. This observational technique utilizes the fact that heating and ionization of the gas in the BLR are both caused by the central continuum source of the AGN. Since the UV radiation of AGNs varies, we expect corresponding variations of the physical conditions in the BLR. In this picture, a decreasing continuum flux should then lead to a lower line flux, as is demonstrated in Fig. 5.28. Due to the finite extent of the BLR, the observed variability in the lines will be delayed in time compared to the ionizing continuum. This delay Δt can be identified with the light travel time across the BLR, $\Delta t \sim r/c$. In other words, the BLR feels the variation in the continuum source only after a delay of Δt. From the observed correlated variabilities of continuum and line emission, Δt can be determined for

[8] Note, however, that this argument essentially pictures the clouds as having some random velocities. It is not unlikely that the picture of 'clouds' is somewhat misleading; instead, the BLR could consist of a turbulent gas, with a large-scale velocity field, in which condensations are present. These condensations then take the roles of the 'clouds' in the simple picture.

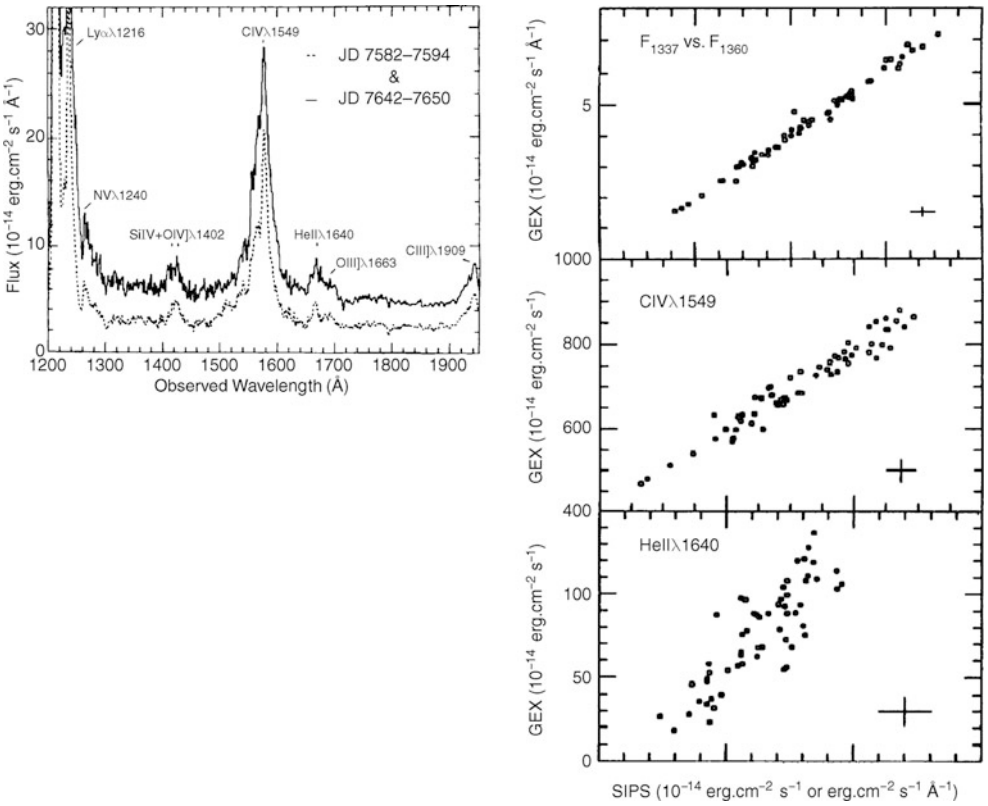


Fig. 5.28 In the *left-hand panel*, the UV spectrum of the Seyfert 1 galaxy NGC 5548 is plotted for two different epochs in which the source radiated strongly and weakly, respectively. It can clearly be seen that not only does the continuum radiation of the source vary but also the strength of the emission lines. The *right-hand panels* show the flux of the continuum at ~ 1300 Å, the CIV line at $\lambda = 1549$ Å, and the HeII line at $\lambda = 1640$ Å, as a function of the near-UV flux at different epochs during an 8-month observational campaign with the IUE. Source: J. Clavel et al. 1991, *Steps toward determination of the size and structure of the broad-line region in active galactic nuclei. I - an 8 month campaign of monitoring NGC 5548 with IUE*, ApJ 366, 64, p. 69, 76, Figs. 1, 2. ©AAS. Reproduced with permission

different line transitions, and so the corresponding values of r can be estimated.[9]

Such analyses of reverberation mapping are extremely time-consuming and complex because one needs to continuously monitor the continuum light and, simultaneously, the line fluxes of an AGN over a long period. The relevant time-scales are typically months for Seyfert 1 galaxies (see Fig. 5.29). To perform such measurements, coordinated campaigns involving many observatories are necessary, because the light curves have to be observed without any gaps, and one should not depend on the local weather conditions at any single observatory.

Reverberation mapping: results. From the results of such campaigns and the correlation of the light curves in the UV continuum and the different line fluxes (Fig. 5.30), the picture of an inhomogeneous BLR is obtained which extends over a large range in r and which consists of different 'layers'. The various emission lines are emitted at different radii, because the ionization structure of the BLR varies with r; the higher the ionization energy of a transition, the smaller the corresponding radius r. For the Seyfert 1 galaxy NGC 5548, one obtains $\Delta t \sim 12$ d for Lyα, about $\Delta t \sim 26$ d for CIII], and about 50 d for MgII. This may not come as a surprise because the ionizing flux increases for smaller r. The fact that lines of higher ionization energy are located closer to the central continuum source implies that they are also broader than low-ionization lines, according to the scaling $v \sim \sqrt{GM_\bullet / r}$.

We can understand this behavior by considering the two processes which are responsible for the ionization state of the gas. On the one hand, ionization is produced by the ionizing flux of the continuum source; thus, the ionization rate is proportional to L/r^2, where L is the luminosity (strictly speaking, L should be taken as the luminosity in ionizing photons, but if we assume that the shape of the continuum does not vary strongly, one can also use the luminosity in some optical or UV band). On the other hand, ions undergo recombination, with the recombination rate being proportional to the number density of electrons. Thus, we expect that the ionization state depends on the ratio of these two rates. One therefore defines the ionization parameter

$$\Xi = \frac{L}{r^2 n_\mathrm{e}} , \tag{5.34}$$

which describes the relative efficiency of ionization and recombination. The larger Ξ, the more abundant are ions

[9] The emissivity of the gas in the BLR reacts very quickly to a change of the ionizing radiation: if the ionizing flux onto a cloud in the BLR decreases, the corresponding line emission from the cloud decreases on the recombination time scale. For a gas density of $n \sim 10^{11}$ cm^{-3}, this time scale is about a minute—that is, almost instantaneously. Thus, the line emission from a cloud depends on the instantaneous ionizing flux at the cloud.

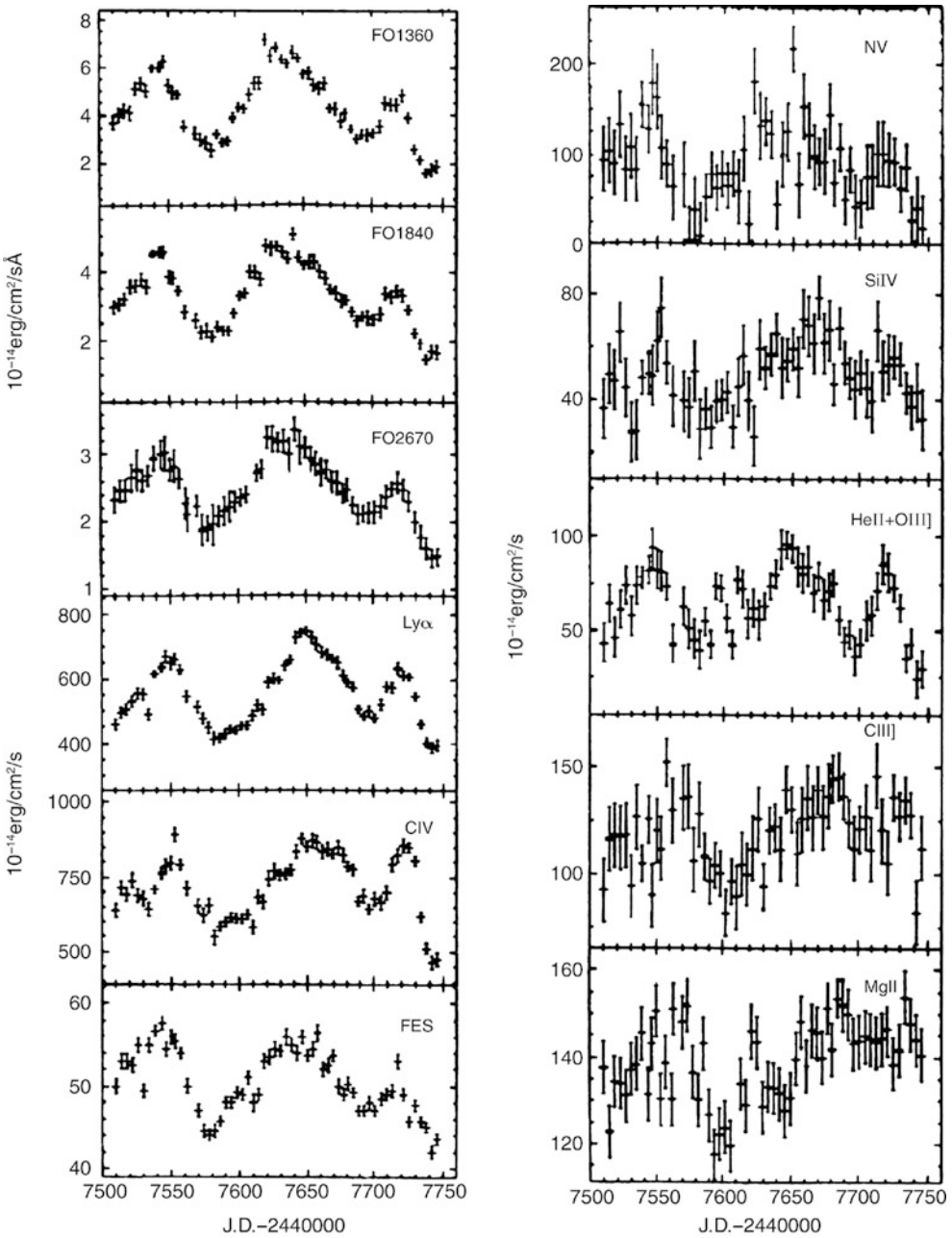


Fig. 5.29 Light curve of NGC 5548 over a period of 8 months at different wavelengths. In the *left-hand panels*, from top to bottom, the continuum at $\lambda = 1350$ Å, $\lambda = 1840$ Å, and $\lambda = 2670$ Å, the broad and strong emission lines Lyα and CIV, as well as the optical light curve are plotted. The *right-hand panels* show the weaker lines NV at $\lambda = 1240$ Å, SiIV at $\lambda = 1402$ Å, HeII+OIII] at $\lambda = 1640$ Å, CIII] at $\lambda = 1909$ Å, and MgII at $\lambda = 2798$ Å. Source: J. Clavel et al. 1991, *Steps toward determination of the size and structure of the broad-line region in active galactic nuclei. I - an 8 month campaign of monitoring NGC 5548 with IUE*, ApJ 366, 64, p. 78, Figs. 3, 4. ©AAS. Reproduced with permission

with high ionization energies, in agreement with the finding from reverberation studies.

If the ionization parameter Ξ is indeed the relevant quantity for the ionization structure of the BLR, then one expects that the size of the region from which a particular line is emitted increases with increasing continuum luminosity. In fact, if one assumes that the density n_e is a constant, one would infer that $r \propto L^{1/2}$ to keep Ξ constant. The Seyfert 1 galaxy NGC 5548 has been monitored for many years, and in addition to short-term, low-amplitude variations of its flux, which is used for the reverberation mapping, its flux varies by larger factors ($\sim$ 2 mag) on longer time scales. Hence, the source has been observed at different levels of activity, i.e., different L. In the left panel of Fig. 5.31, the delay time Δt as measured for the Hβ line is plotted against the continuum luminosity of this source. Clearly, the time lag is correlated with the optical luminosity of this source. The best power-law fit of the relation between time lag ($\propto r$) and luminosity yields a slope of $\sim 0.66 \pm 0.13$, slightly steeper that inferred from the foregoing argument of constant Ξ. However, the uncertainty in the slope is appreciable. In addition, we had to assume that n_e is constant. If n_e decreases with r, a larger slope would be expected. As the bottom line, we see that the extent of the region from which a specific

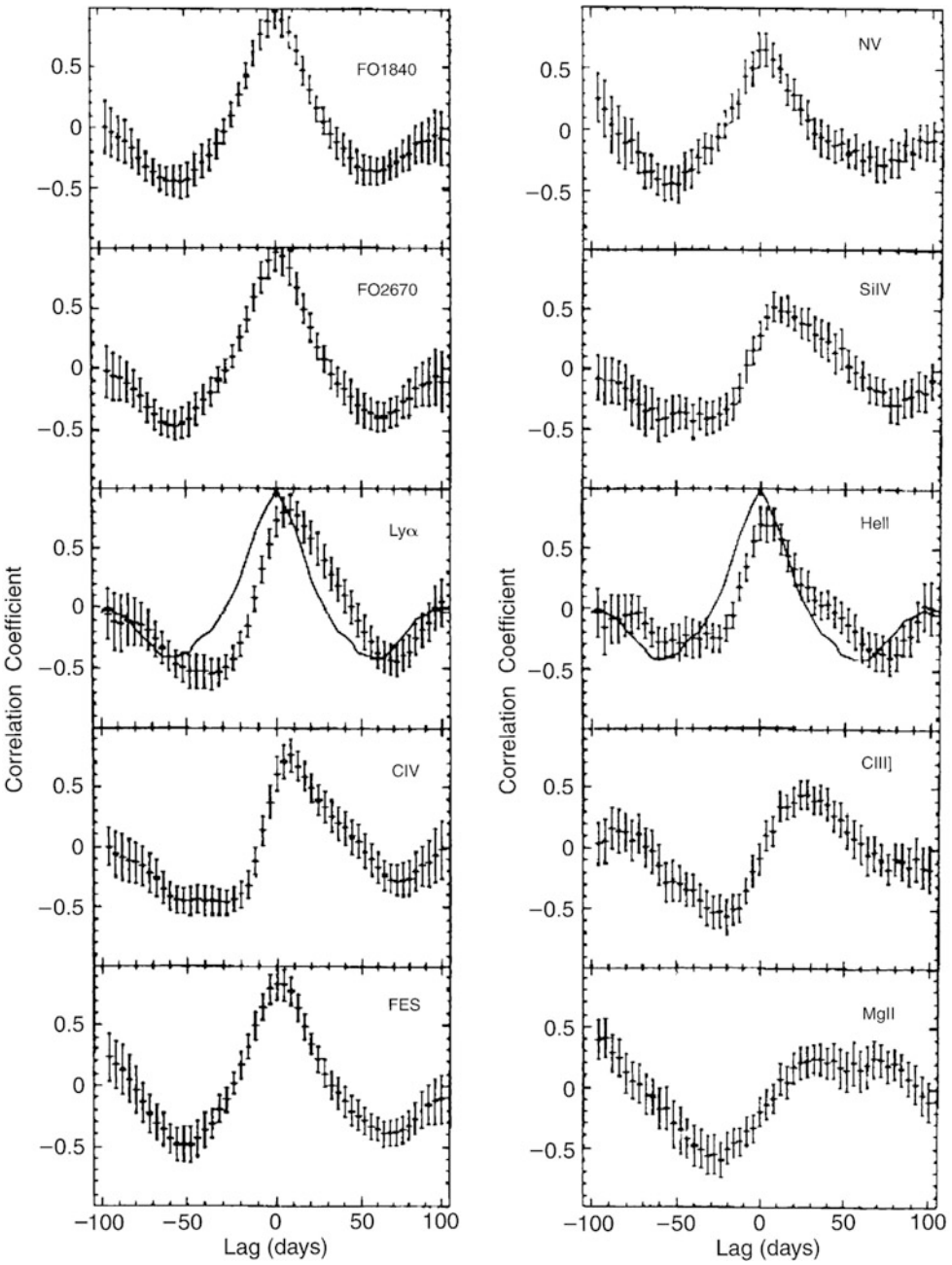


Fig. 5.30 The different light curves from Fig. 5.29 are correlated with the continuum flux at $\lambda = 1350$ Å. The autocorrelation function is shown by the *solid line* in the central panels, the others are cross-correlation functions. We can see that the maximum of the correlation is shifted towards positive times—variations in the continuum flux are not simultaneously followed by the emission lines but appear only after a delay. This delay corresponds to the light travel time from the center of the AGN to the clouds of the BLR where the lines are emitted. The smaller the ionization level of the respective ion, the longer the delay. For example, we obtain a delay of 12 days for Lyα, 26 days for CIII], and about 50 days for MgII, where the latter value could not be measured exactly because the relative flux variations of this line are small and thus the correlation function does not show a very prominent maximum. Source: J. Clavel et al. 1991, *Steps toward determination of the size and structure of the broad-line region in active galactic nuclei. I - an 8 month campaign of monitoring NGC 5548 with IUE*, ApJ 366, 64, p. 79, Figs. 5, 6. ©AAS. Reproduced with permission

line is emitted increases with the source luminosity. This fact has also been demonstrated by comparing the time lags from reverberation mapping of AGNs with different luminosity—the larger L, the larger is the extent of the BLR, scaling roughly as expected from the constancy of Ξ, as seen in the right-hand panel of Fig. 5.31.

The relative flux variations in lines of higher ionization energy are larger, as can also be seen in Fig. 5.14. In addition, lines of higher ionization energy have a mean wavelength systematically shifted bluewards compared to narrower emission lines. As mentioned before, this fact could be interpreted as an indication for outflowing motion of the BLR; however, velocity-resolved reverberation mapping in some Seyfert 1 galaxies hints towards an inflow motion.

Thus, the nature of the clouds in the BLR is to some degree still unknown, as are their geometrical distribution and their kinematic behavior. Their small extent and high temperature imply that they should vaporize on very small time-scales unless they are somehow stabilized. Therefore these clouds need to be either permanently replenished or they have to be stabilized, either by external pressure, e.g., from a very hot but thin medium in the BLR in between the clouds, by magnetic fields, or even gravitationally. One possibility is that the clouds are the extended atmospheres of

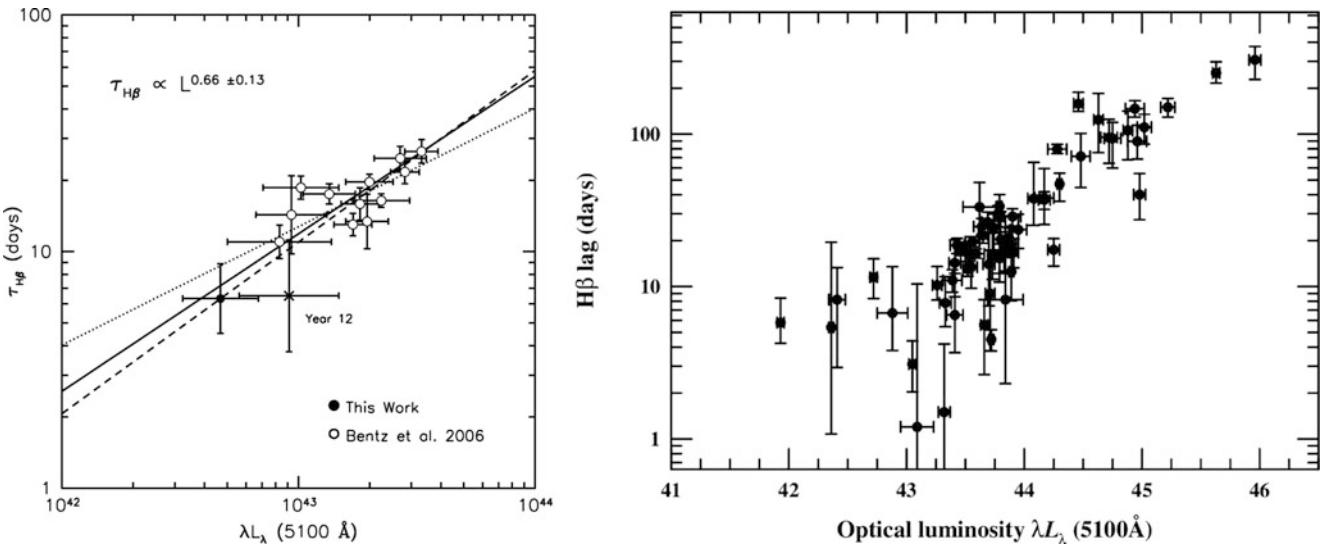


Fig. 5.31 *Left panel*: For the source NGC 5548, the time lag between line- and continuum variations are shown for the Hβ line, as a function of the optical continuum luminosity of the source. The latter was obtained after correcting the optical flux for the contribution by stellar light of the host galaxy. The data were obtained by monitoring this Seyfert 1 galaxy for almost two decades. The *dotted line* indicates the scaling $\delta t \propto L^{1/2}$, whereas the *solid line* is the best fitting power law, with the data point 'Year 12' (for which the sampling during the reverberation mapping campaign has been worse than in other years) excluded. *Right panel*: Measured time lags of the Hβ line as a function of optical luminosity, for a sample of 35 AGNs with reverberation mapping. Note the large range of luminosity. The slope of the best-fitting power law is again 0.6 ± 0.1 as for the case of NGC 5548. Source: *Left*: M.C. Bentz et al. 2007, *NGC 5548 in a Low-Luminosity State: Implications for the Broad-Line Region*, ApJ 662, 205, p. 210, Fig. 5. ©AAS. Reproduced with permission. *Right*: B.M. Peterson 2006, *Emission-Line Variability in Active Galactic Nuclei*, ASPC 360, 191, p. 196. Reproduced by permission of the Astronomical Society of the Pacific

stars; this would, however, imply a very high (probably too high) total mass of the BLR.

5.4.3 Narrow emission lines

Besides the broad emission lines that occur in QSOs, Seyfert 1 galaxies, and broad-line radio galaxies, most AGNs (with exception of the BL Lacs) show narrow emission lines. Their typical width is $\sim 400\,\mathrm{km/s}$. This is considerably narrower than lines of the BLR, but still significantly broader than characteristic velocities in normal galaxies. In analogy to the BLR, the region in which these lines are produced is known as the *narrow line region* (NLR). The strongest line from the NLR is, besides Lyα and CIV, the forbidden [OIII] line at $\lambda = 5007\,\text{Å}$. The existence of forbidden lines implies that the gas density in the NLR is significantly lower than in the BLR.

The gas in the NLR is also assumed to be ionized by UV-radiation from the central continuum source. From estimates analogous to those for the BLR, the characteristic properties of the NLR can then be determined. It should be noted that no reverberation mapping can be applied, since the NLR extends over a region of $\sim 100\,\mathrm{pc}$ for Seyfert 1 galaxies. Because of this large size, no variability of the narrow line intensities is expected on time-scales accessible to observation, and none has been found. The line ratios of allowed and forbidden lines yield $n_\mathrm{e} \sim 10^4\,\mathrm{cm}^{-3}$ for the typical density of the gas in which the lines originate. The characteristic temperature of the gas is likewise obtained from line ratios, $T \sim 15\,000\,\mathrm{K}$, probably slightly lower than in the BLR. The filling factor here is also significantly smaller than unity, about 10^{-2}. Hence, the geometrical picture of clouds in the NLR also emerges. Like in the BLR, the properties of the NLR are not homogeneous but vary with r.

With a size of $r \sim 100\,\mathrm{pc}$, the NLR can be spatially resolved for nearby Seyfert galaxies. The morphology of the NLR is very interesting: it is not spherical, but appears as two cone-shaped regions (Fig. 5.32). It seems as if the ionization of the NLR by the continuum radiation of the AGN is not isotropic, but instead depends strongly on the direction, and is confined largely to a cone-shaped region, called 'ionization cone'.

The BPT diagram. Active galaxies are not the only galaxies which show emission lines; whenever a galaxy undergoes active star formation, its spectrum will contain emission lines from the ionized regions (HII-regions) around newly-born, hot stars. If a galaxy shows broad emission lines, or other clear signs of nuclear activity (such as strong non-thermal radio emission), its identification as an AGN is straightforward. However, for many emission-line galaxies,

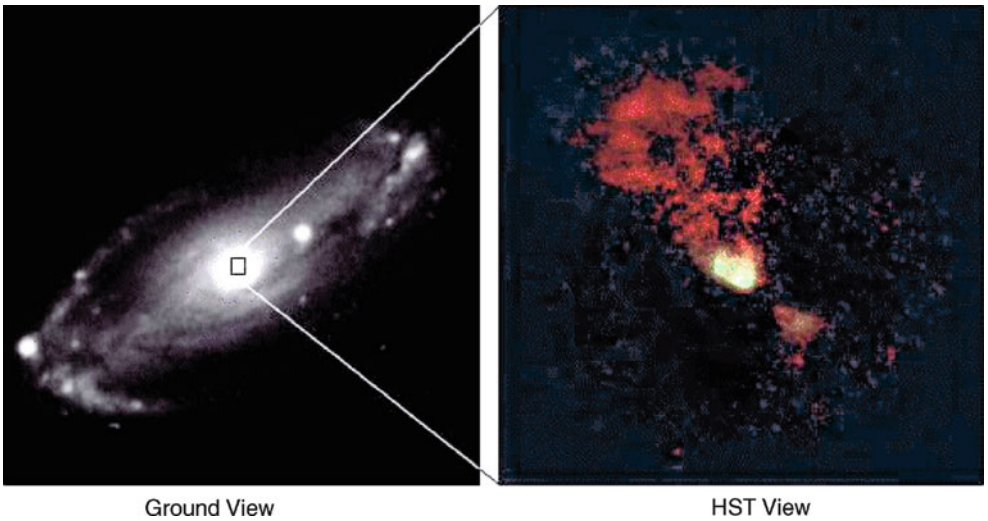


Fig. 5.32 Image of the Seyfert galaxy NGC 5728. On the *left*, a large-scale image showing the disk galaxy; on the *right*, an HST image of its central region taken through a filter with a small bandwidth (narrow-band filter) centered on a narrow emission line. This image shows the spatially resolved NLR. We can see that it is not spherical but consists of two cones ('ionization cones'). From this, it is concluded that the ionizing radiation of the AGN is not isotropic, but is emitted in two preferred directions which appear to be perpendicular, at least in this case, to the disk of the Galaxy. Credit: Allan Sandage, Observatories of the Carnegie Institution of Washington & Andrew S. Wilson, Department of Astronomy, University of Maryland; STScI/NASA

distinguishing between the possible sources of emission lines is non-trivial. For the physical interpretation of a galaxy, this distinction is essential: if a galaxy shows bright emission lines from its central region, it can be due to a burst of star formation, or due to a central AGN.

As pointed out by Baldwin, Phillips, and Terlevich in 1981, the ratios of line strengths can be employed for identifying the different origins of emission lines. The basic idea behind these BPT diagrams is that the source of photons, which ionizes the gas producing the emission lines, is different in these two cases: Massive stars have a clear cut-off in their ionizing spectrum, at the Lyman-limit of helium (corresponding to $\lambda = 228\,\text{Å}$), whereas the non-thermal radiation from AGNs extends to much higher photon energies. One can show that, as a consequence, the ratio of collisionally excited lines to that of lines which are produced in the course of recombination is larger in the case of an AGN-like ionizing radiation field.

Figure 5.33 shows an example of such a BPT diagram, based on SDSS galaxy spectra. As diagnostics, the line ratios [NII]/Hα and [OIII]/Hβ are chosen here; in both cases, the wavelengths of the two lines are quite similar. Thus, these line ratios should only weakly be affected by extinction (of course, ratios of fluxes are independent of the distance to the sources). One sees that the distribution of galaxies in that diagram shows a distinctive pattern: On the one hand, galaxies are distributed along an 'arc' (shown in blue), with an upper envelope indicated by the dashed curve; on the other hand, a second major concentration extends from the lower-right part of the arc towards larger line ratios, shown in grey and red. From numerical modeling, coupled with stellar population synthesis models, it was found that star-forming regions cannot produce line ratios which are above the dotted line. Hence, objects located above the dotted line are essentially powered by AGN radiation. The morphology of the galaxy distribution in the BPT diagram suggests that the galaxies below the dashed curve form a class of its own; therefore, galaxies in the region indicated in blue are considered to be powered solely by star formation. Objects in the region between the two curves, shown in grey, can originate from both, star formation and AGN activity, and they are frequently considered as 'composite' objects.

Furthermore, the distribution of AGN-powered objects in Fig. 5.33 seems to display a further substructure; there seems to be a bimodal distribution in the line ratio of [OIII]/Hβ. Objects with a larger value of this line ratio are found to be Seyfert 2 galaxies (and Seyfert 1 galaxies, if only their narrow line ratios are considered), whereas those with a lower line ratio correspond to LINERs.

The BPT diagram, and variants thereof (in which different line ratios are considered) are a useful tool for classifying emission-line galaxies and are thus extensively used.

5.4.4 X-ray emission

The most energetic radiation of an AGN is expected to be produced in the immediate vicinity of the SMBH. Therefore, the X-ray emission of AGNs is of special interest for probing the innermost regions of these objects, as we have already

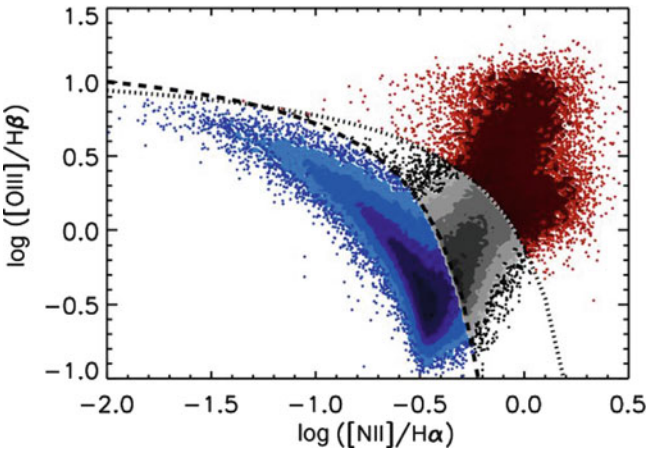


Fig. 5.33 The BPT diagram, a diagnostic for the primary source of emission lines in galaxies. Plotted is the distribution of $\sim$ 240 000 galaxies in the SDSS with $0.02 \leq z \leq 0.35$ for which the indicated emission lines were detected with S/N $>$ 5, with the line ratio [NII]/Hα against [OIII]/Hβ. The *shading* indicates the number density of galaxies, and individual galaxies are plotted as point in regions of low density. The *dashed curve* displays an empirically found division between star-forming galaxies and AGNs, whereas the *dotted curve* shows the division based on theoretical considerations. One frequently considers galaxies below the *dashed curve* (shown in *blue*) as star-formation galaxies, those above the *dotted curve* (*red*) as AGNs, and those between the *curves* (*grey*) as composite objects. Source: L. Trouille, A.J. Barger & C. Tremonti 2011, *The OPTX Project. V. Identifying Distant Active Galactic Nuclei*, ApJ 742, 46, p. 4, Fig. 4. ©AAS. Reproduced with permission

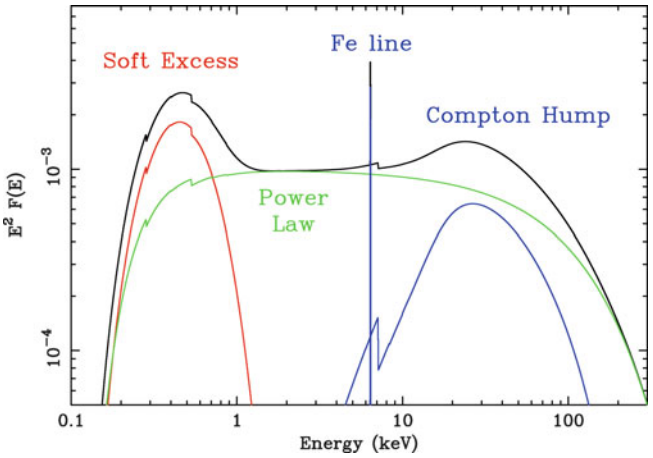


Fig. 5.34 Sketch of a typical X-ray spectrum of an AGN, together with the four components of a standard model, as discussed in the text. Source: A.C. Fabian 2006, ESASP 604, 463, p. 463, Fig. 1. Reproduced by permission of the author

seen from the relativistic iron line shown in Fig. 5.20. In fact, the variability on very short time-scales (see Fig. 5.14) as well as the microlensing results (Fig. 5.27) are a clear indicator of a small extent of the X-ray source.

To a first approximation, the X-ray spectrum in the few keV range is characterized by a power law, $S_\nu \propto \nu^{-\alpha}$, with mean slope $\alpha \sim 0.7$ (see Fig. 5.34). There is a trend that the slope is somewhat steeper for radio-quiet AGNs, and flatter for radio-loud ones. At energies $h_\mathrm{P}\nu \gtrsim 10\,\mathrm{keV}$, the spectrum exceeds the extrapolation of this power law, i.e., it becomes flatter. Towards lower X-ray energies, the spectrum seems to be steeper than the power law; this feature is called the 'soft excess'. The X-ray emission extends up to energies of $\sim$ 100 keV, beyond which there is a spectral cut-off; however, for blazars the spectrum can extend to much higher energies (see below). At low X-ray energies, the observed spectrum is heavily cut off due to photoelectric absorption in the ISM of the Milky Way.

Origin of the X-rays. The decomposition of the continuum spectrum in Fig. 5.34 into a soft excess, a power law, and a Compton hump is based on ideas of how the X-rays in AGNs are generated. Details of the model are yet uncertain; nevertheless, the basic ingredients are probably well established. The origin of the soft excess may be related to the Big Blue Bump, as discussed before; here we will discuss the other components indicated in Fig. 5.34.

Accretion disk corona. Our description of an optically thick, geometrically thin accretion disk in Sect. 5.3.2 led to the picture that at every radius, the disk is characterized by a temperature, and radiates locally almost as a blackbody. It thus can locally be compared to the surface of a star, and stellar atmosphere models have been used to calculate a more accurate spectrum from such disks. In the outer layers of such disks, the gas density is small. Furthermore, we argued that magnetic fields in the disk are probably responsible for the friction which is needed to transport angular momentum through the disk and allow the accretion of material. One may thus expect, in analogy to stars (like in our Sun), that a hot layer of gas forms above the optically thick part of the disk, a corona. The gas in the corona is so thin that it cannot cool efficiently, and hence its temperature can be much higher than that of the disk; perhaps it may even approach the virial temperature, such that $k_\mathrm{B}T \sim GM_\bullet m/r = (2r/r_\mathrm{S})^{-1}\,mc^2$, where m is the mass of a particle. If $m = m_\mathrm{e}$, the temperature of the corona can reach tens or hundreds of keV in the innermost regions of the disk.[10] A sketch of a possible geometry of the corona above the disk is shown in Fig. 5.35.

Inverse Compton effect. As mentioned before, if a photon scatters off an electron at rest, then due to recoil, a (small) fraction of the photon energy is transferred to the electron; this is the Compton effect. However, if the electron has a much larger energy than the photon, part of the electron energy can be transferred to the photon, which is called the

[10] The virial temperature of the protons is much higher, but it is not clear how well electrons and protons are coupled in this hot, thin plasma.

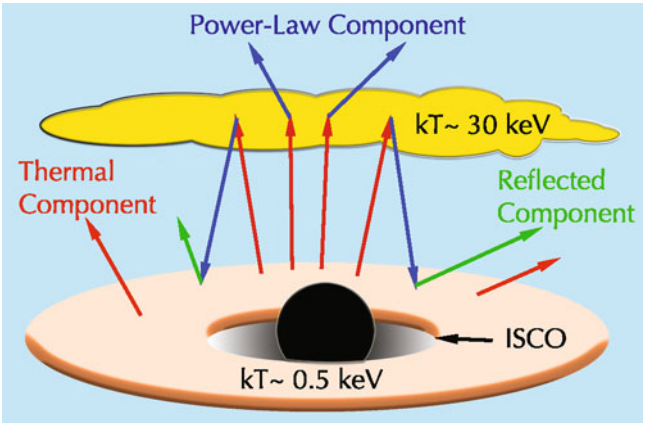


Fig. 5.35 Illustration of an accretion disk and a hot corona, where the possible origin of the various X-ray components of an AGN are indicated; see text for details. Source: L. Gou et al. 2011, *The Extreme Spin of the Black Hole in Cygnus X-1*, ApJ 742, 85, p. 5, Fig. 2. ©AAS. Reproduced with permission

inverse Compton effect. Thus, if photons propagate through a distribution of energetic electrons, they will on average gain energy. For a thermal plasma of temperature T, the mean energy gain of a photon with incoming energy E_γ can be shown to be

$$\Delta E_\gamma = \left(4k_\mathrm{B}T - E_\gamma\right) \frac{E_\gamma}{m_\mathrm{e}c^2} \,, \tag{5.35}$$

so that a photon can either gain or lose energy on average, depending on its energy relative to the electron temperature. If a low-energy photon is scattered multiple times, it can increase its energy is every scattering, until its energy becomes of the same order as the temperature of the gas.

Blackbody photons from the accretion disk propagate through the corona and may scatter off the hot electron distribution. Since the original photon energy is much smaller than $k_\mathrm{B}T$ in the corona, after one scatting is will have an energy

$$E_1 = E_\gamma + \Delta E_\gamma \approx \left(1 + \frac{4k_\mathrm{B}T}{m_\mathrm{e}c^2}\right) E_\gamma \approx E_\gamma \exp\left(\frac{4k_\mathrm{B}T}{m_\mathrm{e}c^2}\right) .$$

A photon may actually scatter more than once, and each time its energy is increased; after N scatterings, the mean photon energy is

$$E_N \approx E_\gamma \exp\left(\frac{4Nk_\mathrm{B}T}{m_\mathrm{e}c^2}\right) .$$

The mean number of scatterings $\langle N\rangle$ a photon undergoes before it leaves the corona depends on the optical depth $\tau_\mathrm{T} = \sigma_\mathrm{T} N_\mathrm{e}$ with respect to electron scattering, where N_e is the electron column density of the corona. If $\tau_\mathrm{T} \lesssim 1$, then $\langle N\rangle \approx \tau_\mathrm{T}$; in the opposite case of $\tau_\mathrm{T} \gg 1$, the mean number of scatterings is $\langle N\rangle \sim \tau_\mathrm{T}^2$. But of course, for a given $\langle N\rangle$, there will be a probability distribution $p(N)$ of the number of scatterings which photons will experience before they leave the corona.

The power-law component. Hence, the emission from the corona will be the sum of photons which scattered N times and which have energy E_N. One can show that the resulting spectrum is a power law in energy whose slope depends on the temperature in the corona and the scattering optical depth τ_T. There is an upper cut-off of this power law, given by $E_\gamma \sim k_\mathrm{B}T$. The power-law component of the X-ray spectrum is interpreted in terms of this thermal Comptonization. The fact that the observed cut-off in the spectrum occurs at $E_\gamma \sim 100\,\mathrm{keV}$ implies that this radiation must originate from the very inner part of the accretion disk, where the virial temperature of the electrons is that high.

The reflection component. Photons scattering in the corona may escape the disk and then form the observable power-law component. However, since the electron scattering is basically isotropic, these photons may be scattered back to the disk. If they have a low energy, they will simply be absorbed. However, if their energy is high, they will scatter inside the disk. Since in this case, the situation $E_\gamma \gg k_\mathrm{B}T$ applies, these photons will lose energy due to scattering. In analogy to what was said above, there is a probability distribution such that photons which hit the disk are scattered N times before they can leave the disk again, and the larger N, the more energy they lose. Furthermore, as they are scattered down in energy, their probability of being absorbed is increased. Thus, this *reflection component* is appreciable only at rather high photon energies. The primary photon energy must be larger than that of the reflection component, indicating that this component must originate from the innermost region of the accretion flow. In addition, line radiation from photoionized gas in the disk, together with the relativistic effects due to relativistic orbital velocities in the inner disk region, may give rise to the soft X-ray excess.

Line emission. Besides this continuum radiation, emission and absorption lines are also found in the X-ray domain, the strongest line being the 6.4 keV iron line (see Sect. 5.3.4). Together with other emission lines at lower energy, it is probably produced in the accretion disk through reflection (note that the energy of this line is much larger than the maximum temperature of a standard accretion disk around SMBHs): A high-energy photon from the corona hits an iron atom in the disk and removes one of its inner (K-shell) electrons. After this K-shell ionization, the electronic K-shell is refilled by a transition of the ion, emitting the 6.4 keV Kα line. The improved sensitivity and spectral resolution of the X-ray telescopes Chandra and XMM-Newton compared to earlier X-ray observatories have greatly advanced the X-ray

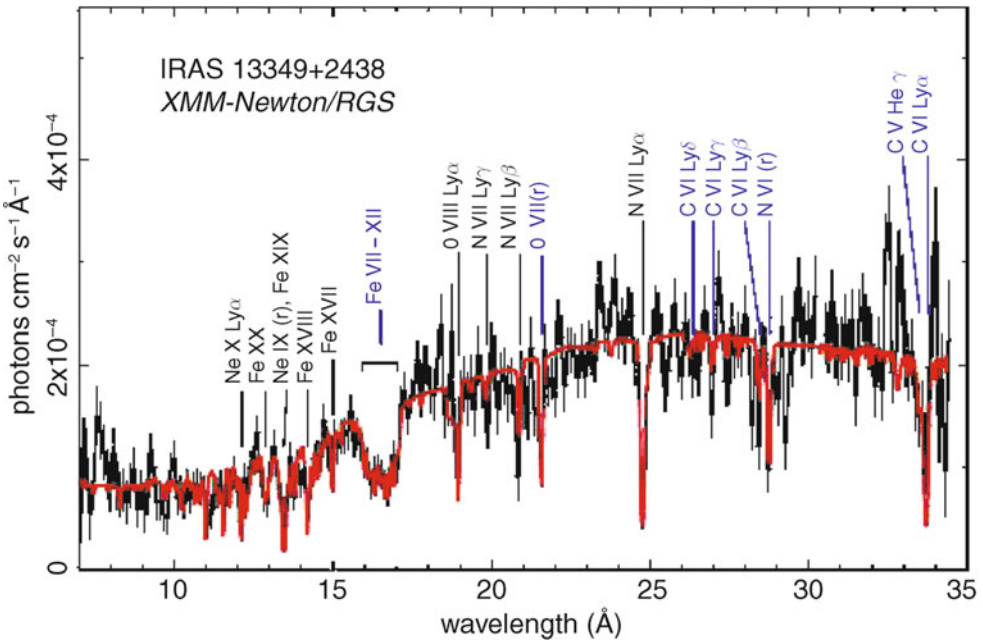


Fig. 5.36 X-ray spectrum of the quasar IRAS 13349 + 2438 ($z = 0.108$), observed by the XMM satellite. Various absorption lines are marked. Source: M. Sako et al. 2001, *Complex resonance absorption structure in the X-ray spectrum of IRAS 13349 + 2438*, A&A 365, L168, p. L170, Fig. 1. ©ESO. Reproduced with permission

spectroscopy of AGNs. Figure 5.36 shows an example of the quality of these spectra.

The X-ray emission of Seyfert 1 and Seyfert 2 galaxies is very different. In the energy range of the ROSAT X-ray satellite ($0.1\,\mathrm{keV} \leq h_\mathrm{P}\nu \leq 2.4\,\mathrm{keV}$), significantly more Seyfert 1 galaxies were discovered than Seyfert 2 galaxies. The origin of this was later uncovered by Chandra and XMM-Newton. In contrast to ROSAT, these two satellites are sensitive up to energies of $h_\mathrm{P}\nu \sim 10\,\mathrm{keV}$ and they have found large numbers of Seyfert 2 galaxies. However, their spectrum differs from that of Seyfert 1 galaxies because it is cut off towards lower X-ray energies. The spectrum indicates the presence of absorbing material with a hydrogen column density of $\gtrsim 10^{22}\,\mathrm{cm}^{-2}$ and in some cases even orders of magnitude higher.[11] This fact will be used in the context of unified models (Sect. 5.5) of AGNs.

5.4.5 The host galaxy

As the term 'active galactic nuclei' already implies, AGNs are considered the central engine of otherwise quite normal galaxies. This nuclear activity is nourished by accretion of matter onto a SMBH. Since it seems that all galaxies (at least those with a spheroidal component) harbor a SMBH, the question of activity is rather one of accretion rate. What does it take to turn on a Seyfert galaxy, and why are most SMBHs virtually inactive? And by what mechanism is matter brought into the vicinity of the SMBH to serve as fuel?

For a long time it was not clear as to whether QSOs are also hosted in a galaxy. Their high luminosity renders it difficult to identify the surrounding galaxy on images taken from the ground, with their resolution being limited by seeing to $\sim 1''$. In the 1980s, the surrounding galaxies of some QSOs were imaged for the first time, but only with the HST, it became possible to obtain detailed images of QSO host galaxies (see Fig. 5.37) and thus to include them in the class of galactic nuclei. In these investigations, it was also found that the host galaxies of QSOs are often heavily disturbed, e.g., by tidal interaction with other galaxies or even by merging processes. These disturbances of the gravitational potential are considered essential for the gas to overcome the angular momentum barrier and to flow towards the center of the galaxy. At the same time, such disturbances seem to increase the star-formation rate enormously, because starburst galaxies are also often characterized by disturbances and interactions. A close connection seems to exist between AGN activity and nuclear starbursts—in fact, they both have in common that they require the presence of gas in the central region of a galaxy. Optical and NIR images of QSOs (see Fig. 5.37) cannot unambiguously answer the question of whether QSO hosts are spirals or ellipticals.

Today it seems established that the hosts of low-redshift luminous QSOs are predominantly massive and bulge-dominated galaxies. This finding is in good agreement with the fact that the black hole mass in 'normal' galaxies scales with the mass of the spheroidal component of the galaxies,

[11] The absorption of X-rays is due to ionization of metals. Whereas the photoelectric effect is also present for hydrogen, the corresponding cross section for X-rays is small, due to their high energy and the strong frequency dependence of the cross section, $\propto \nu^{-3}$ above the energy threshold. Despite the fact that metals have a much smaller abundance than hydrogen and helium, they dominate the optical depth for X-ray absorption. Nevertheless, the absorber is characterized in terms of a hydrogen column density, implicitly assuming that the gas has Solar metallicity.

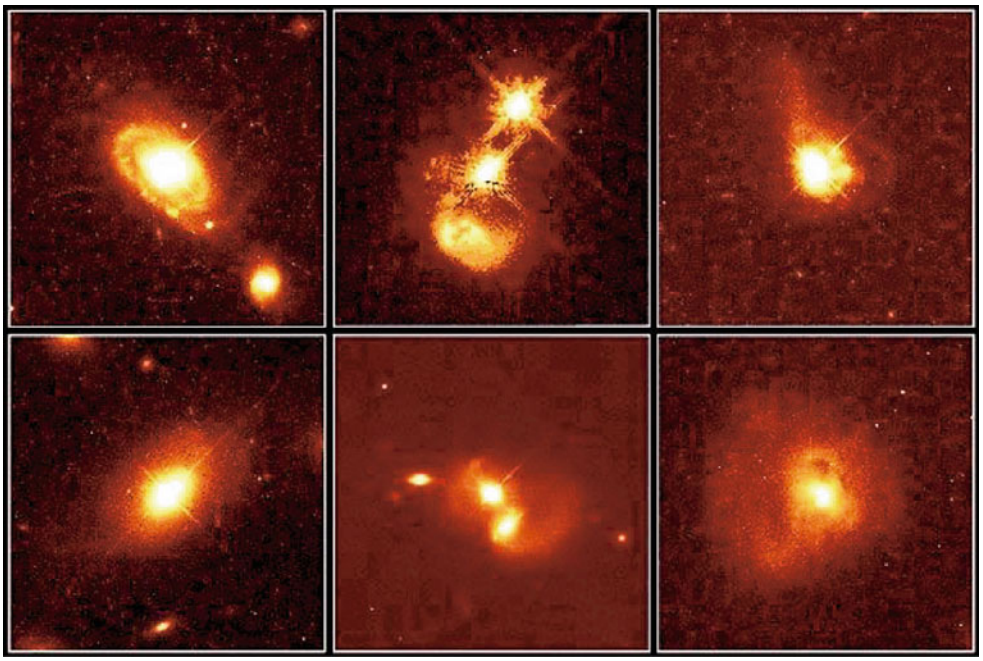


Fig. 5.37 HST images of QSOs. In all cases the host galaxy can clearly be identified, with the QSO itself being visible as a (central) point source in these images. *Top left*: PG 0052 + 251 is located in the center of an apparently normal spiral galaxy. *Bottom left*: PHL 909 seems to be located in the center of a normal elliptical galaxy. *Top center*: the QSO IRAS 04505−2958 is obviously part of a collision of two galaxies and may be provided with 'fuel' by material ripped from the galaxies by tidal forces. Surrounding the QSO core, a region of active star formation is visible. PG 1012 + 008 (*bottom center*) is also part of a pair of merging galaxies. *Top right*: the host galaxy of QSO 0316−346 seems to be about to capture a tidal tail. *Bottom right*: the QSO IRAS 13218 + 0552 seems to be located in a galaxy which just went through a merger process. Credit: J. Bahcall (Institute for Advanced Study, Princeton), M. Disney (University of Wales), and NASA

as the Eddington limit sets a lower bound on $M_\bullet$ for a given luminosity. On the other hand, evolved early-type galaxies are gas-poor, and one might thus not expect that a quiescent elliptical can host a luminous QSO; perhaps the central SMBH is 'switched on' in these galaxies only after some interaction with other galaxies. There are some indications that higher-redshift QSOs are also hosted by massive elliptical galaxies. Figure 5.38 shows three gravitational lens systems where the lens is a (low-luminosity) QSO; from analyzing the lensing geometry, one finds that the overall mass properties (stars plus dark matter) of the three host galaxies are very similar to normal early-type galaxies.

For somewhat less luminous, and thus more abundant AGNs, the situation may be different. Those AGNs seem to have hosts whose color places them mostly in the blue cloud (see Sect. 3.6) or the green valley, but they seem to largely avoid the red sequence. This may indicate a connection between AGNs and star-forming galaxies, both relying on the supply of gas. In contrast, weak AGNs are found preferentially in massive red galaxies, which also seem to be the preferred host for radio-loud AGNs.

Binary QSOs. The connection between the activity of galaxies and the presence of close neighbors is also seen from the clustering properties of QSOs. In surveys for gravitational lens systems, pairs of QSO images have been detected which have angular separations of a few arcseconds and very similar redshifts, but sufficiently different spectra to exclude them being gravitationally lensed images of the same source. The number of binary QSOs thus found is considerably larger than the expectation from the large-scale correlation function of QSOs. This conclusion was further strengthened by an extensive analysis of QSOs in the Sloan Digital Sky Survey. The correlation function of QSOs at separations below $\sim 30h^{-1}$ kpc exceeds that of the extrapolation of the correlation function from larger scales by a factor of 10 or more. Hence it seems that the small-scale clustering of QSOs is very much enhanced, say compared to normal galaxies, which could be due to the triggering of activity by the proximity of the neighbor: in this case, both galaxies attain a perturbed gravitational potential and start to become active.

5.4.6 The black hole mass in AGNs

We now return to the determination of the mass of the central black hole in AGNs. In Sect. 5.3.5, a lower limit on the mass was derived, based on the fact that the luminosity of an

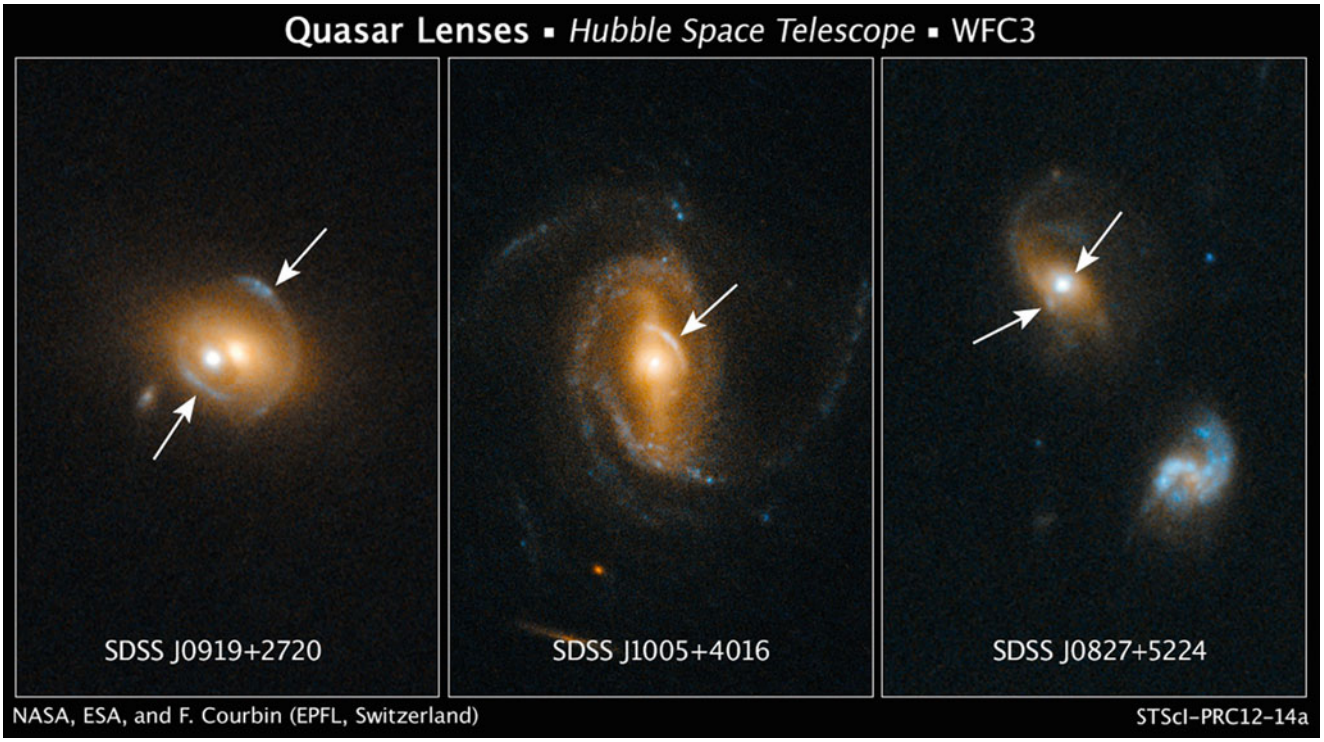


Fig. 5.38 Three QSOs from the SDSS which act as gravitational lenses. Indicated by the *arrows* are the images of star-forming background galaxies ($z_s \sim 0.5$) lensed by the host galaxies of nearby ($z_d \sim 0.2$) QSOs, as obtained from spectroscopy. Analysis of the lensing geometry yields the characteristic mass scale of the lenses, which is very similar to that of early-type gravitational lens galaxies. Credit: NASA, ESA, and F. Courbin (EPFL, Switzerland)

AGN cannot exceed the Eddington luminosity. However, this estimate cannot be very precise, for at least two reasons. The first is related to the anisotropic appearance of an AGN. The observed flux can be translated into a luminosity only on the assumption that the emission from the AGN is isotropic, and we have discussed several reasons why this assumption may not be justified in many cases. Second, we do not have a clear idea what the ratio of AGN luminosity to its Eddington luminosity is. It is clear that this ratio can vary a lot from one system to another. For example, the black hole at the center of our Galaxy could power a luminosity of several 10^{44} erg/s if radiating with the Eddington luminosity—and we know that the true luminosity is many orders of magnitude below this value.

$M_\bullet$ from reverberation mapping. A far more accurate method for estimating the black hole mass in AGNs comes from reverberation mapping which we described in Sect. 5.4.2. The principal quantity that is derived from this technique is the size r of the BLR for a given atomic line or for a given ionization state of a chemical element. Furthermore, the relative line width $\Delta\lambda/\lambda$ can be measured, and can be related to the characteristic velocity dispersion σ in the BLR, $\sigma = c\,\Delta\lambda/\lambda$. Assuming that the gas is virialized, or moving approximately on Keplerian orbits around the black hole, the mass of the latter can be estimated to be

$$M_\bullet = f\, r\, \sigma^2/G\ , \tag{5.36}$$

where $f \sim 1$ for circular orbits and for the observer being located in the plane of the orbit. However, the geometry and kinematics of the BLR may be much more complex than that, and there are orientation effects; the quantity f accounts for this added complexity. If f could be determined, then the black hole mass can be estimated with very reasonable accuracy from reverberation mapping.

One can check whether the functional form of (5.36) is valid, irrespective of the value of f, by studying the relation between time lag Δt (or radius r) and the observed line width *in the same source*, i.e., for a constant value of $M_\bullet$. This was done for the Seyfert 1 galaxy NGC 5548 for which long-term monitoring was carried out (see Fig. 5.31). As shown in Fig. 5.39, one finds indeed a strong correlation between line width and radius, very close to the expected form $\sigma \propto r^{-1/2}$.

The value of f can in principle be estimated from models of the BLR, but these carry considerable uncertainties. A more reliable method is the determination of f empirically. Whereas f cannot be determined for any individual source—and is expected to vary from source to source, e.g.,

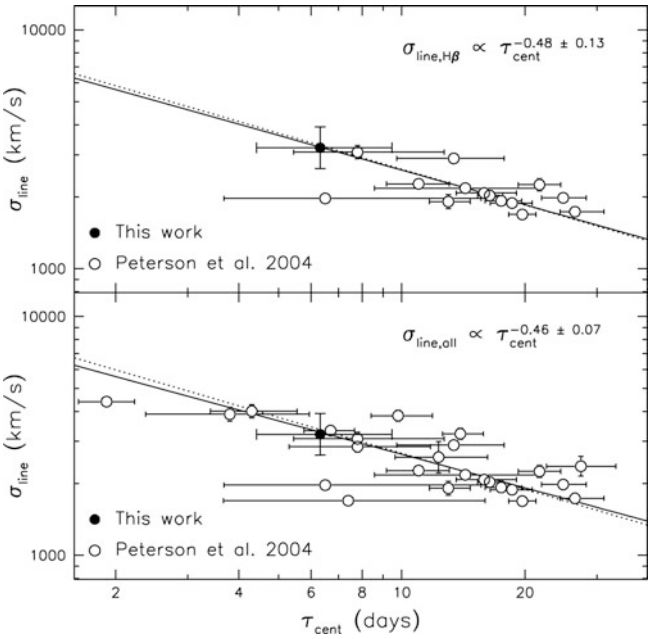


Fig. 5.39 The emission line width as a function of time lags for the Seyfert 1 galaxy NGC 5548. In the *upper panel*, the measurements of the Hβ line are shown, based on the same data used for Fig. 5.31, whereas in the *lower panel* all emission lines are plotted. In both cases, the line width scales with time lag approximately as $\sigma \propto (\Delta t)^{-1/2}$, i.e., obeying a virial relation (5.36). Source: M.C. Bentz et al. 2007, *NGC 5548 in a Low-Luminosity State: Implications for the Broad-Line Region*, ApJ 662, 205, p. 210, Fig. 4. ©AAS. Reproduced with permission

due to different inclination angles relative to the line-of-sight—its mean value can be estimated from a set of AGNs for which the stellar velocity dispersion of the AGN host can be measured. As we have seen in Sect. 3.8.3, there is a well-defined relation between SMBH mass and stellar velocity dispersion. Assuming that the same relation holds for AGNs, then the SMBH mass can be determined from stellar velocity dispersion measurements. This yields a first estimate of $M_\bullet$. From reverberation measurements, (5.36) yields a second, independent mass estimate, which, however, depends on f. Minimizing the difference between these two independent estimates of $M_\bullet$ for a sample of AGNs then yields an estimate for the mean of $f = 5.9 \pm 2.0$; see the left panel of Fig. 5.40.

With the value of f determined, the SMBH mass can now be estimated from (5.36) also for those AGNs for which no measurements of the stellar velocity dispersion σ_* of the host are available. Corresponding results are shown in the right-hand panel of Fig. 5.40, where $M_\bullet$ is plotted as a function of optical continuum luminosity, from which the stellar contribution of the host galaxy was subtracted. Assuming a constant ratio between the optical and bolometric luminosity of 1/9, lines of constant Eddington ratio $\lambda_{\rm Edd} = L_{\rm bol}/L_{\rm edd}$ can be drawn in this figure. We see that the measurements cover a very broad range of luminosities and estimated black hole masses, extending over more than three orders of magnitude. None of the sources has an estimated Eddington ratio larger than unity, but they are concentrated around $\lambda_{\rm Edd} \sim 0.1$. Also, only a single source has an estimated $\lambda_{\rm Edd}$ smaller than 0.01. It thus seems that AGNs with broad emission lines are accreting at a fairly high rate.

However, reverberation mapping is a fairly expensive observing technique. Furthermore, the effort required for this technique increases with AGN luminosity, since the size of the BLR, and thus the time delay and the necessary length of the monitoring campaign, increases with the black hole mass. We might therefore want to look at alternative methods for estimating $M_\bullet$.

$M_\bullet$ from scaling relations. We saw from Fig. 5.40 that nearby AGNs, for which reverberation data are available, obey the same relation (3.49) between $M_\bullet$ and the stellar velocity dispersion σ_* of the host as is obtained for inactive galaxies. This scaling relation then yields a useful estimate of the black hole mass from the stellar velocity dispersion. Unfortunately, even this method cannot be applied to a broad range of AGNs, since the velocity dispersion of stars cannot be measured in AGNs which are either too luminous—since then the nuclear emission outshines the stellar light, rendering spectroscopy of the latter impossible—or too distant, so that a spatial separation of nuclear light from stellar light is no longer possible.

However, the scaling relation between the size of the BLR for a given transition and the optical continuum luminosity, $r \propto L^\beta$, with $\beta \approx 0.6$ (see Fig. 5.31), is very useful for estimating black hole masses. From the continuum luminosity, the size r is estimated using this scaling relation; combined with the measured emission line width σ, the relation (5.36) can be applied. In particular, this method can be extended to luminous and high-redshift sources.

The Eddington ratio. Once an estimate for $M_\bullet$ is obtained, the Eddington luminosity can be calculated and compared with the observed luminosity. The ratio of these two is the Eddington ratio, $\lambda_{\rm Edd} \equiv L_{\rm bol}/L_{\rm Edd}$. For the estimate of $\lambda_{\rm Edd}$, the observed luminosity in the optical band needs to be translated into a bolometric luminosity, which can be done with the help of the average spectral energy distribution of AGNs of a given class. If one can ignore strongly beamed emission, $\lambda_{\rm Edd}$ should be smaller than unity, which seems to be indeed the case as seen from Fig. 5.40. Whereas the various steps in deriving $\lambda_{\rm Edd}$ involve statistical and systematic uncertainties by factors $\gtrsim 2$, the results indicate that most broad-line AGNs have an Eddington ratio of order $\lambda_{\rm Edd} \sim 0.1$.[12]

[12]There might be a trend that radio-loud QSOs have a somewhat larger $\lambda_{\rm Edd}$, but these correlations are controversial and might be based on

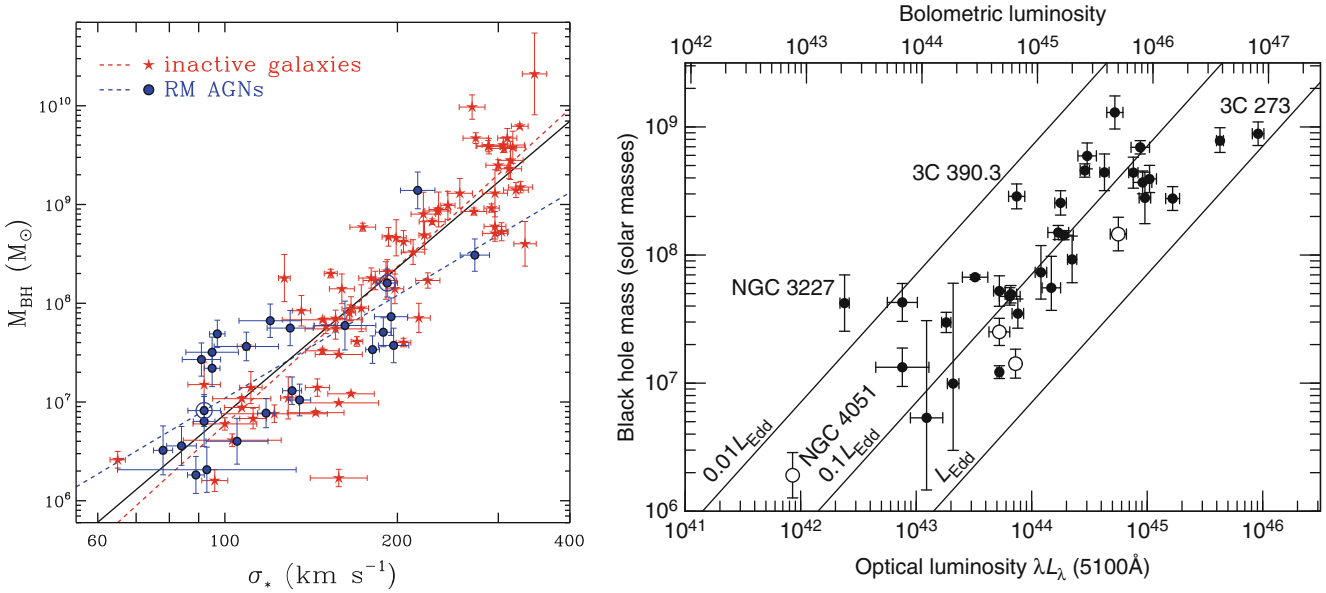


Fig. 5.40 *Left panel*: The SMBH mass $M_\bullet$ as a function of stellar velocity dispersion, for a sample of nearby normal, quiescent galaxies (shown in *red*) and for AGNs for which reverberation mapping data are available (*blue symbols*). The two *dashed lines* correspond to the best power-law fits for these two samples, where the virial factor $f \approx 5.9 \pm 2$ yields the best agreement between these two samples. The *black solid curve* is the best power-law fit to both data sets. *Right panel*: The relation between black hole mass $M_\bullet$ and continuum luminosity, for a sample of 35 AGNs with reverberation mapping, using a value of $f = 5.5$. The upper axis is an estimate of the bolometric luminosity under the assumption that $L_{bol} = 9\lambda L_\lambda(5100\,\text{Å})$. *Diagonal lines* indicate constant Eddington ratio $\lambda_{Edd} = L_{bol}/L_{Edd}$ of 0.01, 0.1 and 1. Source: Left: J.-H. Woo et al. 2013, *Do quiescent and active galaxies have different M_{BH}-σ relations?*, arXiv:1305.2946, Fig. 4. Reproduced by permission of the author. Right: B.M. Peterson et al. 2004, *Central Masses and Broad-Line Region Sizes of Active Galactic Nuclei. II. A Homogeneous Analysis of a Large Reverberation-Mapping Database*, ApJ 613, 682, p. 696, Fig. 16. ©AAS. Reproduced with permission

The fact that λ_{Edd} is confined to a fairly narrow interval implies that the luminosity of a QSO can be used to estimate $M_\bullet$, just by setting $M_\bullet = \lambda_{Edd} M_{Edd}(L)$. This mass estimate has a statistical uncertainty of at least a factor of three in individual sources, but requires only the measurement of the continuum luminosity. However, it is not clear whether the mean Eddington ratio is approximately constant in redshift; it may well be that there is a cosmological evolution of the properties of AGNs. We will come back to this point in Sect. 10.6.2.

The Galactic center black hole. The Eddington ratio of the SMBH in the Galactic center is many orders of magnitude smaller than unity, at least at the current epoch; in fact, with its total luminosity of 5×10^{36} erg/s, $\lambda_{Edd} \sim 10^{-8}$. Such a small value indicates that the SMBH in our Galaxy is starved; the accretion rate must be very small. However, one can estimate a minimum mass rate with which the SMBH in the Galactic center is fed, by considering the mass-loss rate of the stars in its vicinity. This amounts to $\sim 10^{-4} M_\odot$/yr, enough material to power an accretion flow with $L \sim 10^{-2} L_{Edd}$. The fact that the observed luminosity is so much smaller than this value leads to two implications. The first of these is that there must be other modes of accretion which are far less efficient than that of the geometrically thin, optically thick accretion disk. Such models for accretion flows were indeed developed, such as the ADAF briefly described in Sect. 5.3.2. The second conclusion is that the central mass concentration must indeed be a black hole—a black hole is the only object which does not have a surface. If, for example, one would postulate a hypothetical object with $M \sim 4 \times 10^6 M_\odot$ which has a hard surface (like a scaled-up version of a neutron star), the accreted material would fall onto the surface, and its kinetic and internal energy would be deposited there. Hence, this surface would heat up and radiate thermally. Since we have strict upper limits on the radius of the object, coming from mm-VLBI observations, we can estimate the minimum luminosity such a source would have. This estimate is again several orders of magnitude larger than the observed luminosity from Sgr A*, firmly ruling out the existence of such a solid surface.

Evolution of the $M_\bullet$ scaling relations. As we have seen in Sect. 3.8.3, the black hole mass in normal, nearby galaxies is correlated with the bulge (or spheroidal) luminosity. As this component of galaxies consists of an old stellar population, its luminosity is very closely related to its stellar mass. Esti-

selection effects. On the other hand, radio galaxies have a lower value of λ_{Edd} than QSOs.

mating the black hole mass from the continuum luminosity of the QSOs, and observing the spheroidal luminosity of their host galaxies (which requires the high angular resolution of HST), one can investigate whether such a scaling relation already existed at earlier epochs, i.e., at high redshifts.

The results from such studies indicate that the ratio of black hole mass and stellar mass of the spheroidal component of the host galaxy evolves with redshift, in the sense that $M_\bullet/M_*$ was larger in the past. Furthermore, the scatter in the scaling relation increases with redshift. There is some debate on how strong these evolutionary effects are: Whereas evolution as strong as $M_\bullet/M_* \propto (1+z)^2$ is claimed, the scatter in the relation impacts on the sample of QSOs with well measured stellar and black hole masses by selection effects which bias the apparent redshift evolution. A milder evolution of the form $M_\bullet/M_* \propto (1+z)^{0.7}$ was deduced from QSOs in the COSMOS field, when these selection effects are accounted for.

5.5 Family relations of AGNs

5.5.1 Unified models

In Sect. 5.2, several different types of AGNs were mentioned. We saw that many of their properties are common to all types, but also that there are considerable differences. Why are some AGNs seen as broad line radio galaxies, others as BL Lac objects? The obvious question arises as to whether the different classes of AGNs consist of rather similar objects which differ in their appearance due to geometric or light propagation effects, or whether more fundamental differences exist. At the beginning of Sect. 5.2, we summarized a classification scheme for AGNs, called unified model. In this section, we will collect the various differences and similarities of the various classes of AGNs, and provide evidence for the unified scheme, which will be explained in more detail below.

Common properties. Common to all AGNs is a SMBH in the center of the host galaxy, the supposed central engine, and also an accretion disk that is feeding the black hole. This suggests that a classification can be based on $M_\bullet$ and the accretion rate $\dot{m}$, or perhaps more relevantly the ratio $\dot{m}/\dot{m}_{\rm edd}$. $M_\bullet$ defines the maximum (isotropic) luminosity of the SMBH in terms of the Eddington luminosity, and the ratio $\dot{m}/\dot{m}_{\rm edd}$ describes the accretion rate relative to its maximum value. Furthermore, the observed properties, in particular the seemingly smooth transition between the different classes, suggest that radio-quiet QSOs and Seyfert 1 galaxies basically differ only in their central luminosity. From this, we would then deduce that they have a similar value of $\dot{m}/\dot{m}_{\rm edd}$ but differ in $M_\bullet$. An analogous argument may be valid for the transition from BLRGs to radio-loud quasars.

The difference between these two classes may be due to the nature of the host galaxy. Radio galaxies (and maybe radio-loud quasars?) are situated in elliptical galaxies, Seyfert nuclei (and maybe radio-quiet quasars?) preferentially in spirals. A correlation between the luminosity of the AGN and that of the host galaxy also seems to exist. This is to be expected if the luminosity of the AGN is strongly correlated with the respective Eddington luminosity, because of the correlation between the SMBH mass in normal galaxies and the properties of the galaxy (Sect. 3.8.3). Another question is how to fit blazars and Seyfert 2 galaxies into this scheme.

Anisotropic emission. In the context of the SMBH plus accretion disk model, another parameter exists that affects the observed characteristics of an AGN, namely the inclination, i.e., the angle between the rotation axis of the disk and the direction from which we observe the AGN. We should mention that in fact there are many indications that the radiation of an AGN is not isotropic and thus its appearance depends on the viewing angle. Among these are the observed ionization cones in the NLR (see Fig. 5.32) and the morphology of the radio emission, as the radio lobes define a preferred direction. Furthermore, our discussion of superluminal motion has shown that the observed superluminal velocities are possible only if the direction of motion of the source component is close to the direction of the line-of-sight. The X-ray spectrum of many AGNs shows intrinsic (photoelectric) absorption caused by high column density gas, where this effect is mainly observed in Seyfert 2 galaxies. Because of these clear indications it seems obvious to examine the dependence of the appearance of an AGN on the viewing direction. For example, the observed difference between Seyfert 1 and Seyfert 2 galaxies may simply be due to a different orientation of the AGN relative to the line-of-sight.

Broad emission lines in polarized light. In fact, another observation of anisotropic emission provides a key to understanding the relation between AGN types, which supports the above idea. The galaxy NGC 1068 has no visible broad emission lines and is therefore classified as a Seyfert 2 galaxy. Indeed, it is considered an archetype of this kind of AGN. However, the optical spectrum of NGC 1068 in polarized light shows broad emission lines (Fig. 5.41) such as one would find in a Seyfert 1 galaxy. Obviously the galaxy must have a BLR, but it is only visible in polarized light. The photons that are emitted by the BLR are initially unpolarized. But polarization may be induced through scattering of the

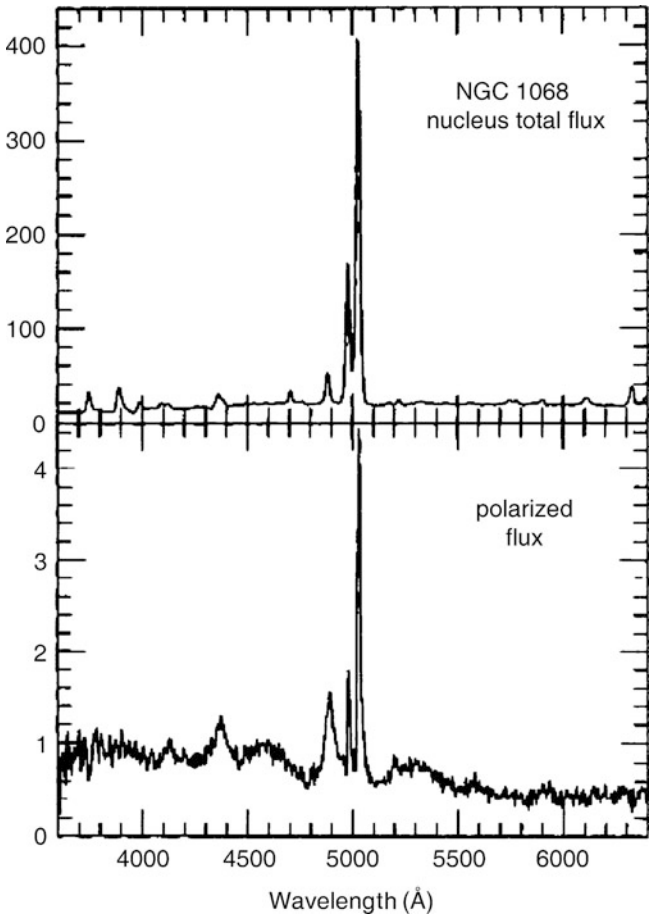


Fig. 5.41 Spectrum of the Seyfert 2 galaxy NGC 1068. The *top panel* displays the total flux which, besides the continuum, also shows narrow emission lines, in particular [OIII] at $\lambda = 5007$ Å and $\lambda = 4959$ Å. However, in polarized light (*bottom panel*), broad emission lines (like Hβ λ4861 Å and Hγ λ4340 Å) typical of a Seyfert 1 galaxy are also visible. Therefore, it is concluded that the BLR becomes visible in light polarized via scattering; the BLR is thus visible only indirectly. Source: J.S. Miller et al. 1991, *Multidirectional views of the active nucleus of NGC 1068*, ApJ 378, 47, p. 50, Fig. 6. ©AAS. Reproduced with permission

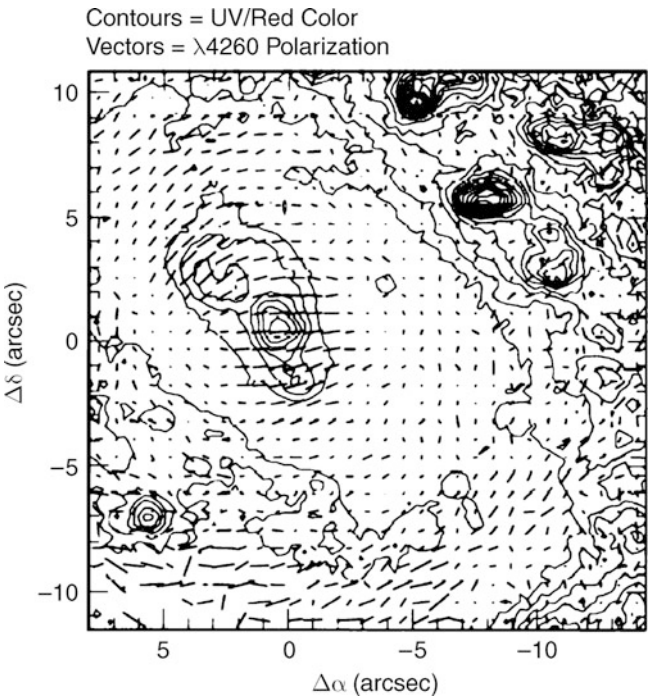


Fig. 5.42 The contours show the color of the optical emission in the Seyfert 2 galaxy NGC 1068, namely the flux ratio in the U- and R-bands. The sticks indicate the strength and orientation of the polarization in B-band light. The center of the galaxy is located at $\Delta\alpha = 0 = \Delta\delta$. At its bluest region (*center left*), the polarization of the optical emission is strongest and is perpendicular to the direction to the center of the galaxy; this is the direction of polarization expected for local scattering by electrons. Hence, where the scattering is strongest, the largest fraction of direct light from the AGN is also observed, and the optical spectrum of AGNs is considerably bluer than the stellar light from galaxies. Source: R.W. Pogge & M.M. De Robertis 1993, *Extended near-ultraviolet continuum emission and the nature of the polarized broad-line Seyfert 2 galaxies*, ApJ 404, 563, p. 568, Fig. 4. ©AAS. Reproduced with permission

light, where the direction perpendicular to the directions of incoming and scattered photons defines a preferred direction, which then yields the polarization direction.

The interpretation of this observation (see Fig. 5.12) now is that NGC 1068 has a BLR but our direct view of it is obscured by absorbing material. However, this absorber does not fully engulf the BLR in all directions but only within a solid angle of $<4\pi$ as seen from the central core. If photons from the BLR are scattered by dust or electrons in a way that we are able to observe the scattered radiation, then the BLR would be visible in this scattered light. Direct light from the AGN completely outshines the scattered light, which is the reason why we cannot identify the latter in the total flux. By scattering, however, this radiation is also polarized. Thus in observations made in polarized light, the (unpolarized) direct radiation is suppressed and the BLR becomes visible in the scattered light.[13]

This interpretation is additionally supported by a strong correlation of the spatial distribution of the polarization and the color of the radiation in NGC 1068 (see Fig. 5.42). We can conclude from this that the differences between Seyfert 1 and Seyfert 2 galaxies originate in the orientation of the accretion disk and thus of the absorbing material relative to the line-of-sight.

From the abundance ratio of Seyfert 1 to Seyfert 2 galaxies (which is about 1:2), the fraction of solid angle in which the view to the BLR is obscured, as seen from the AGN, can be estimated. This ratio then tells us that about 2/3 of the

[13]Not all Seyfert 2 galaxies show broad emission lines in polarized flux, which may be either due to the fact that there is no appropriate scattering medium which makes the BLR visible for us, or that some of the sources intrinsically lack a BLR. The discussion about the possible existence of such 'true Seyfert 2' objects has not yet come to a clear conclusion.

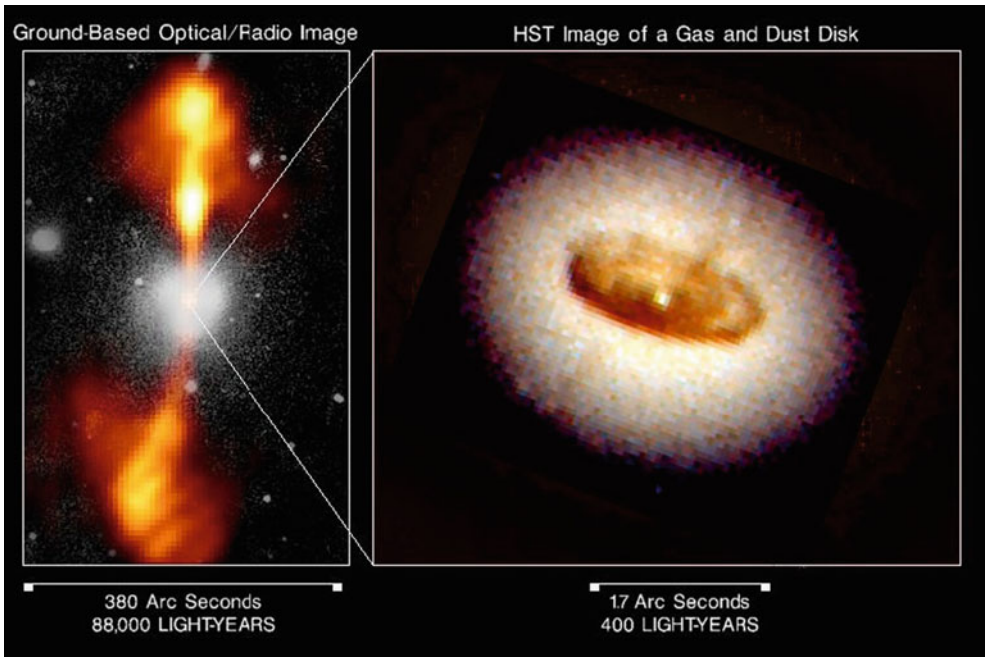


Fig. 5.43 The elliptical galaxy NGC 4261. The *left-hand panel* shows an optical image of this galaxy together with the radio emission (shown in *orange*). An HST image showing the innermost region of the galaxy is shown on the *right*. The jet is virtually perpendicular to the central disk of gas and dust, which is in agreement with the theoretical picture in the context of a unification model. Credit: National Radio Astronomy Observatory, California Institute of Technology, Walter Jaffe/Leiden Observatory, Holland Ford/JHU/STScI, and NASA

solid angle is covered by an absorber. Such a blocking of light may be caused by dust imbedded in a gas distribution. It is assumed that the dust is located in the plane of the accretion disk in the form of a thick 'torus' (see Figs. 5.12 and 5.43 for a view of this geometry).

Originally, the 'torus' was imagined as a rather big and smooth distribution of gas and dust. However, from the properties of the infrared emission from QSOs, it is now believed to be considerably smaller and instead clumpy. The 'torus' is now visioned as a region around the central AGN, perhaps an order-of-magnitude larger than the BLR, which is filled with rather dense clouds. The dust in the clouds absorbs radiation from the AGN, gets heated, and reradiates the energy in form of infrared radiation. Models assuming a size distribution of the clouds can reproduce the observed spectral shapes in the infrared spectral regime. Whereas there is considerable variation between individual objects, one finds that some 25 % of the total power emitted by an AGN emerges in the infrared regime. Taken at face value, this would imply that the material in the torus blocks the light in a quarter of all directions; however, this estimate is probably too simple. Instead, one expects a broad probability distribution of absorbing optical depths for light rays traversing the torus region.

Search for Type 2 QSOs. If the difference between Seyfert galaxies of Type 1 and Type 2 is caused merely by their orientation, and if likewise the difference between Seyfert 1 galaxies and QSOs is basically one of absolute luminosity, then the question arises as to whether a luminous analog for Seyfert 2 galaxies exists, a kind of Type 2 QSO. Until around 2000, such Type 2 QSOs had not been observed, from which it was concluded that either no dust torus is present in QSOs due to the high luminosity (and therefore no Type 2 QSOs exist) or that Type 2 QSOs are not easy to identify.

This question is now settled: the current X-ray satellites Chandra and XMM-Newton have identified the population of Type 2 QSOs. Due to the high column density of hydrogen which is distributed in the torus together with the dust, low-energy X-ray radiation is almost completely absorbed by the photoelectric effect if the line-of-sight to the center of these sources passes through the obscuring torus. These sources were therefore not visible for ROSAT ($E \leq 2.4\,\mathrm{keV}$), but the energy ranges of Chandra and XMM-Newton finally allowed the X-ray detection and identification of these Type 2 QSOs.

Other candidates for Type 2 QSOs are the ultra-luminous infrared galaxies (ULIRGs), in which extreme IR-luminosity is emitted by large amounts of warm dust which is heated either by very strong star formation or by an AGN. Since ULIRGs have total luminosities comparable to QSOs, the latter interpretation is possible. In fact, distinguishing between the two possibilities is not easy for individual ULIRGs, and in many sources indicators of both strong star formation and non-thermal emission (e.g., in the form of X-ray emission) are found. This discovery indicates that in many objects, the processes of strong star formation and accretion onto a SMBH are linked. For both processes, large amounts of gas are necessary, and the fact that both starburst galaxies and AGNs are often found in interacting galaxies, where the disturbance in the gravitational field provides the conditions for a gas flow into the center of the galaxy, suggests a link between the two phenomena.

Next we will examine how blazars fit into this unified scheme. A first clue comes from the fact that all blazars are radio sources. Furthermore, in our interpretation of superluminal motion (Sect. 5.3.3) we saw that the appearance and apparent velocity of the central source components depend on the orientation of the source with respect to us, and that it requires relativistic velocities of the source components. To obtain an interpretation of the blazar phenomenon that fits into the above scheme, we first need to discuss an effect that results from Special Relativity.

5.5.2 Beaming

Due to relativistic motion of the source components relative to us, another effect occurs, known as *beaming*. Because of beaming, the relation between source luminosity and observed flux from a moving source depends on its velocity with respect to the observer. One aspect of this phenomenon is the Doppler shift in frequency space: the measured flux at a given frequency is different from that of a non-moving source because the measured frequency corresponds to a Doppler-shifted frequency in the rest-frame of the source. Another effect described by Special Relativity is that a moving source which emits isotropically in its rest-frame has an anisotropic emission pattern, with the angular distribution depending on its velocity. The radiation is emitted preferentially in the direction of the velocity vector of the source (thus, in the forward direction), so that a source will appear brighter if it is moving towards the observer. In Sect. 4.3.2, we already mentioned the relation (4.46) between the radiation intensity in the rest-frame of a source and in the system of the observer. Due to the strong Doppler shift, this implies that a source moving towards us appears brighter by a factor (*Doppler factor*)

$$\mathcal{D}_+ = \left(\frac{1}{\gamma(1 - \beta \cos\phi)} \right)^{2+\alpha} \qquad (5.37)$$

than the source at rest, where α is the spectral index. Here, $\beta = v/c$, ϕ is the angle between the velocity vector of the source component and the line-of-sight to the source, and the Lorentz factor $\gamma = (1 - \beta^2)^{-1/2}$ was already defined in Sect. 5.3.3. Even at modest relativistic velocities ($\beta \sim 0.9$) this can already be a considerable factor, i.e., the radiation from the relativistic jet may appear highly amplified. Another consequence of beaming is that if a second jet exists which is moving away from us (the so-called counter-jet), its radiation will be weakened by a factor

$$\mathcal{D}_- = \left(\frac{1}{\gamma(1 + \beta \cos\phi)} \right)^{2+\alpha} \qquad (5.38)$$

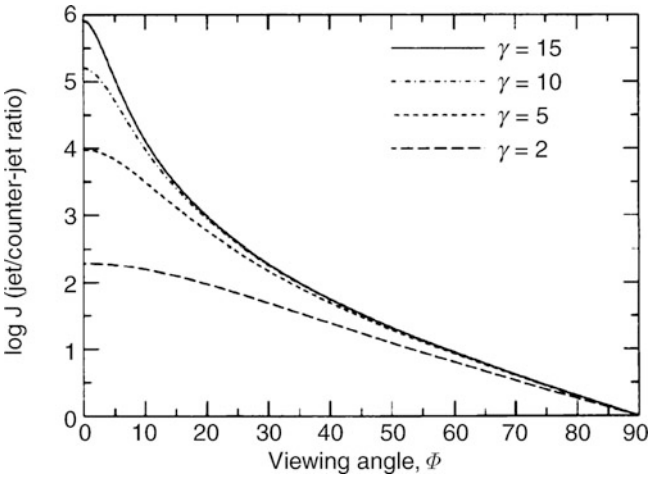


Fig. 5.44 The logarithm of the flux ratio of jet and counter-jet (5.39) is plotted as a function of the angle ϕ for different values of the Lorentz factor γ. Even at relatively small values of γ, this ratio is large if ϕ is close to 0, but even at $\phi \sim 30°$ the ratio is still appreciable. Hence the plot shows the Doppler favoritism and explains why, in most compact radio AGNs, one jet is visible but the counter-jet is not. Source: C.M. Urry & P. Padovani 1995, *Unified Schemes for Radio-Loud Active Galactic Nuclei*, PASP 107, 803, p. 839, Fig. 22. ©ASP. Reproduced with permission

relative to the stationary source. Obviously, $\mathcal{D}_-$ can be obtained from $\mathcal{D}_+$ by replacing ϕ by $\phi + \pi$, since the counter-jet is moving in the opposite direction. In particular, the flux ratio of jet and counter-jet is

$$\frac{\mathcal{D}_+}{\mathcal{D}_-} = \left(\frac{1 + \beta \cos\phi}{1 - \beta \cos\phi} \right)^{2+\alpha} , \qquad (5.39)$$

and this factor may easily be a hundred or more (Fig. 5.44). The large flux ratio (5.39) for relativistic jets is the canonical explanation for VLBI jets being virtually always only one-sided. This effect is also denoted as 'Doppler favoritism'—the jet pointing towards us is observed preferentially because of the beaming effect and the resulting amplification of its flux.

Beaming and the blazar phenomenon. If we observe a source from a direction very close to the jet axis and if the jet is relativistic, its radiation can outshine all other radiation from the AGN because $\mathcal{D}_+$ can become very large in this case. Especially if the beamed radiation extends into the optical/UV part of the spectrum, the line emission may also become invisible relative to the jet emission, and the source will appear to us as a BL Lac object. If the line radiation is not outshined completely, the source may appear as an OVV. The synchrotron nature of the optical light is also the explanation for the optical polarization of blazars since synchrotron emission can be polarized, in contrast to thermal emission.

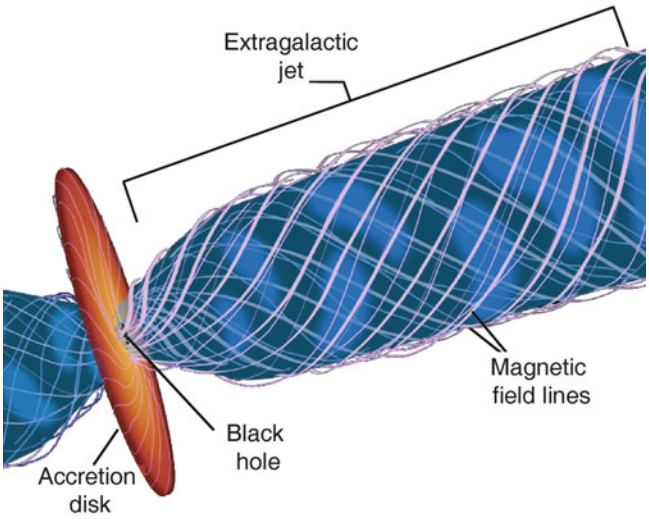


Fig. 5.45 Illustration of the relativistic jet model. The acceleration of the jet to velocities close to the speed of light is probably caused by a combination of very strong gravitational fields in the vicinity of the SMBH and strong magnetic fields which are rotating rapidly because they are anchored in the accretion disk. Shock fronts within the jet lead to acceleration processes of relativistic electrons, which then strongly radiate and become visible as 'blobs' in the jets. By rotation of the accretion disk in which the magnetic field lines are frozen in, the field lines obtain a characteristic helical shape. It is supposed that this process is responsible for the focusing (collimation) of the jet. Credit: NASA/ESA and Ann Feild, Space Telescope Science Institute

The strong beaming factor also provides an explanation for the rapid variability of blazars. If the velocity of the emitting component is close to the speed of light, $\beta \lesssim 1$, even small changes in the jet velocity or its direction may noticeably change the Doppler factor $\mathcal{D}_+$. Such small changes in the direction are expected because there is no reason to expect a smooth outflow of material along the jet at constant velocity. In addition, we argued that, very probably, magnetic fields play an important role in the generation and collimation of jets. These magnetic fields are toroidally spun-up, and emitting plasma can, at least partially, follow the field lines along helical orbits (see Fig. 5.45).

Hence beaming can explain the dominance of radiation from the jet components if the gas is relativistic, and also the absence or relative weakness of emission lines. At the same time, it provides a plausible scenario for the strong variability of blazars. The relative strength of the core emission and the extended radio emission depends heavily on the viewing direction. In blazars, a dominance of the core emission is expected, which is exactly what we observe.

5.5.3 Beaming on large scales

A consequence of this model is that the jets on kpc scales, which are mainly observed by the VLA, also need to be at least semi-relativistic: kpc-scale jets are in most cases also one-sided, and they are always on the same side of the core as the VLBI jet on pc scales. Thus, if the one-sidedness of the VLBI jet is caused by beaming and the corresponding Doppler favoritism of an otherwise intrinsically symmetric source, the one-sidedness of large-scale jets should have the same explanation, implying relativistic velocities for them as well. These do not need to be as close to c as those of the components that show superluminal motion, but their velocity should also be at least a few tenths of the speed of light. In addition, it follows that the kpc-scale jet is moving towards us and is therefore closer to us than the core of the AGN; for the counter-jet we have the opposite case. This prediction can be tested empirically, and it was confirmed in polarization measurements. Radiation from the counter-jet crosses the ISM of the host galaxy, where it experiences additional Faraday rotation (see Sect. 2.3.4). It is in fact observed that the Faraday rotation of counter-jets is systematically larger than that of jets. This can be explained by the fact that the counter-jet is located behind the host galaxy and we are thus observing it through the gas of that galaxy.

5.5.4 Jets at higher frequencies

Optical jets. In Sect. 5.1.2, we discussed the radio emission of jets, and Sect. 5.3.3 described how their relativistic motion is detected from their structural changes, i.e., superluminal motion. However, jets are not only observable at radio frequencies; they also emit at much shorter wavelengths. Indeed, the first two jets were detected in optical observations, namely in QSO 3C273 (Fig. 5.46) and in the radio galaxy M87 (Fig. 5.47), as a linear source structure pointing radially away from the core of the respective galaxy. With the commissioning of the VLA (Fig. 1.26) as a sensitive and high-resolution radio interferometer, the discovery and examination of hundreds of jets at radio frequencies became possible.

The HST, with its unique angular resolution, has detected numerous jets in the optical (see also Fig. 5.17). They are situated on the same side of the corresponding AGNs as the main radio jet. Optical counterparts of radio counter-jets have not been detected thus far. Optical jets are always shorter, narrower, and show more structure than the corresponding radio jets. The spectrum of optical jets follows a power law (5.2) similar to that in the radio domain, with an index α that describes, in general, a slightly steeper spectrum. In some cases, linear polarization in the optical jet radiation of $\sim 10\,\%$ was also detected. If we also take into account that the positions of the knots in the optical and in the radio jets agree very well, we inevitably come to the conclusion that the optical radiation is also synchrotron emission. This

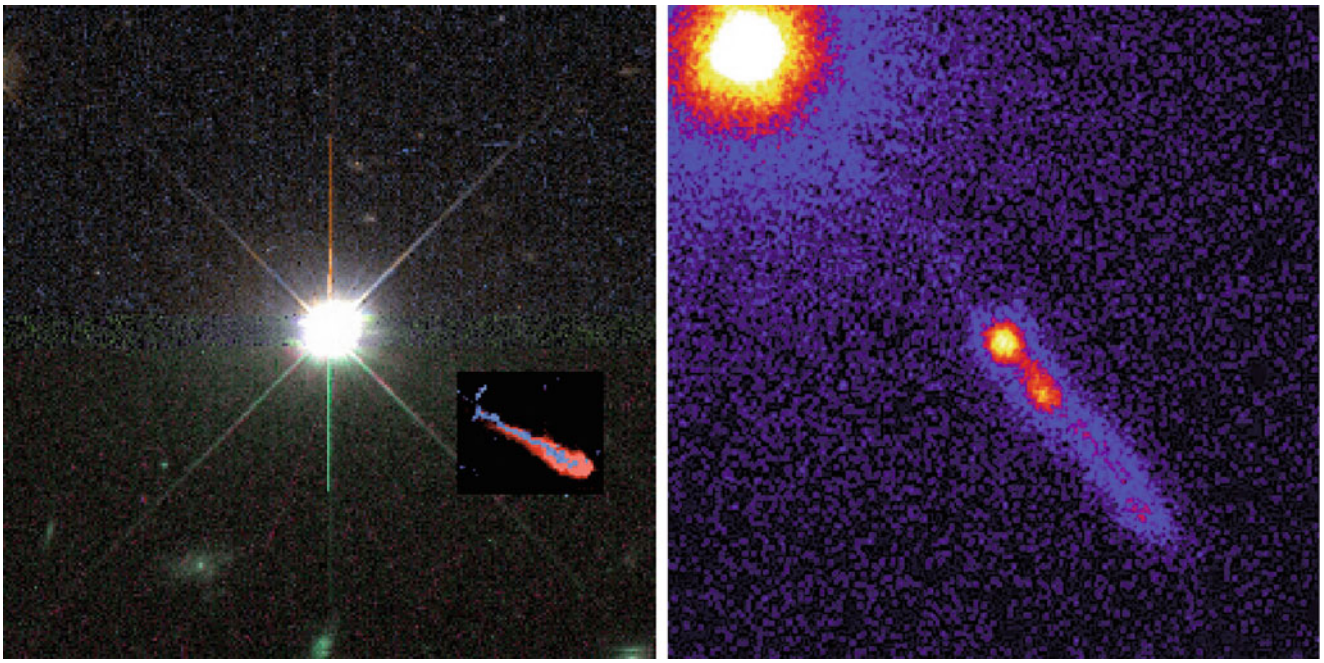


Fig. 5.46 Jets are visible not only in the radio domain but in some cases also at other wavelengths. *On the left*, an HST image of the quasar 3C273 is shown, with the point-like quasar in the center and jet-shaped optical emission (shown in *blue*) that spatially coincides with the radio jet (displayed in *red*). *On the right*, an X-ray image of this quasar taken by the Chandra satellite. The jet is also visible at very high energies. Credit: J. Bahcall/IAS Princeton and NASA; NASA/CXC/SAO/H. Marshall et al.

conclusion is further supported by a nearly constant flux ratio of radio and optical radiation along the jets.

As was mentioned in Sect. 5.1.3, the relativistic electrons that produce the synchrotron radiation lose energy by emission. In many cases, the cooling time (5.6) of the electrons responsible for the radio emission is longer than the time of flow of the material from the central core along the jet, in particular if the flow is (semi-)relativistic. It is thus possible that relativistic electrons are produced or accelerated in the immediate vicinity of the AGN and are then transported away by the jet. This is not the case for those electrons producing the optical synchrotron radiation, however, because the cooling time for emission at optical wavelengths is only $t_{\rm cool} \sim 10^3\,(B/10^{-4}\,{\rm G})^{-3/2}\,{\rm yr}$.[14] Even if the electrons responsible for the optical emission are transported in a (semi-)relativistic jet, they cannot travel more than a distance of ~ 1 kpc before losing their energy. The observed length of optical jets is much larger, though. For this reason, the corresponding electrons cannot be originating in the AGN itself but instead must be produced locally in the jet. The knots in the jets, which are probably shock fronts in the outflow, are thought to represent the location of the acceleration of relativistic particles. Quantitative estimates of the cooling time are hampered by the unknown beaming factor (5.37). Since optical jets are all one-sided, and in most cases observed in radio sources with a flat spectrum, a very large beaming factor is generally assumed. Transforming back into the rest-frame of the electrons yields a lower frequency and a lower luminosity. Since the latter is utilized for estimating the strength of the magnetic fields (by assuming equipartition of energy, for instance), this also changes the estimated cooling time.

The cooling time of electrons emitting at radio frequencies is much longer. The energy-dependent cooling time causes a spectral break in the electron distribution, which shows up accordingly in the synchrotron spectrum, as can be seen in Fig. 5.48 for the case of Centaurus A. The break frequency is highest close to the nucleus, and decreases as one moves away from it.

X-ray radiation of jets. The Chandra satellite discovered that many of the jets which had been identified in the radio are also visible in X-ray light (see Figs. 5.46, 5.48 and 5.50); in fact, currently about 50 X-ray jets are known which are spatially related to corresponding radio structures. This discovery came as a real surprise, since the strong correlation of the spatial distribution of radio, optical, and X-ray emission implies that they must all originate from the same regions in the jets, i.e., that the origins of the emission must be linked to each other. As we have discussed, radio and optical radiation originate from synchrotron emission, the

[14] This dependence of the cooling time $t_{\rm cool} = E/\dot{E}$ on the magnetic field strength follows from (5.3), which at a fixed frequency yields $\gamma = E/(m_{\rm e}c^2) \propto B^{-1/2}$, and the energy loss (5.5), which reads $\dot{E} \propto \gamma^2 B^2 \propto B$ at fixed frequency.

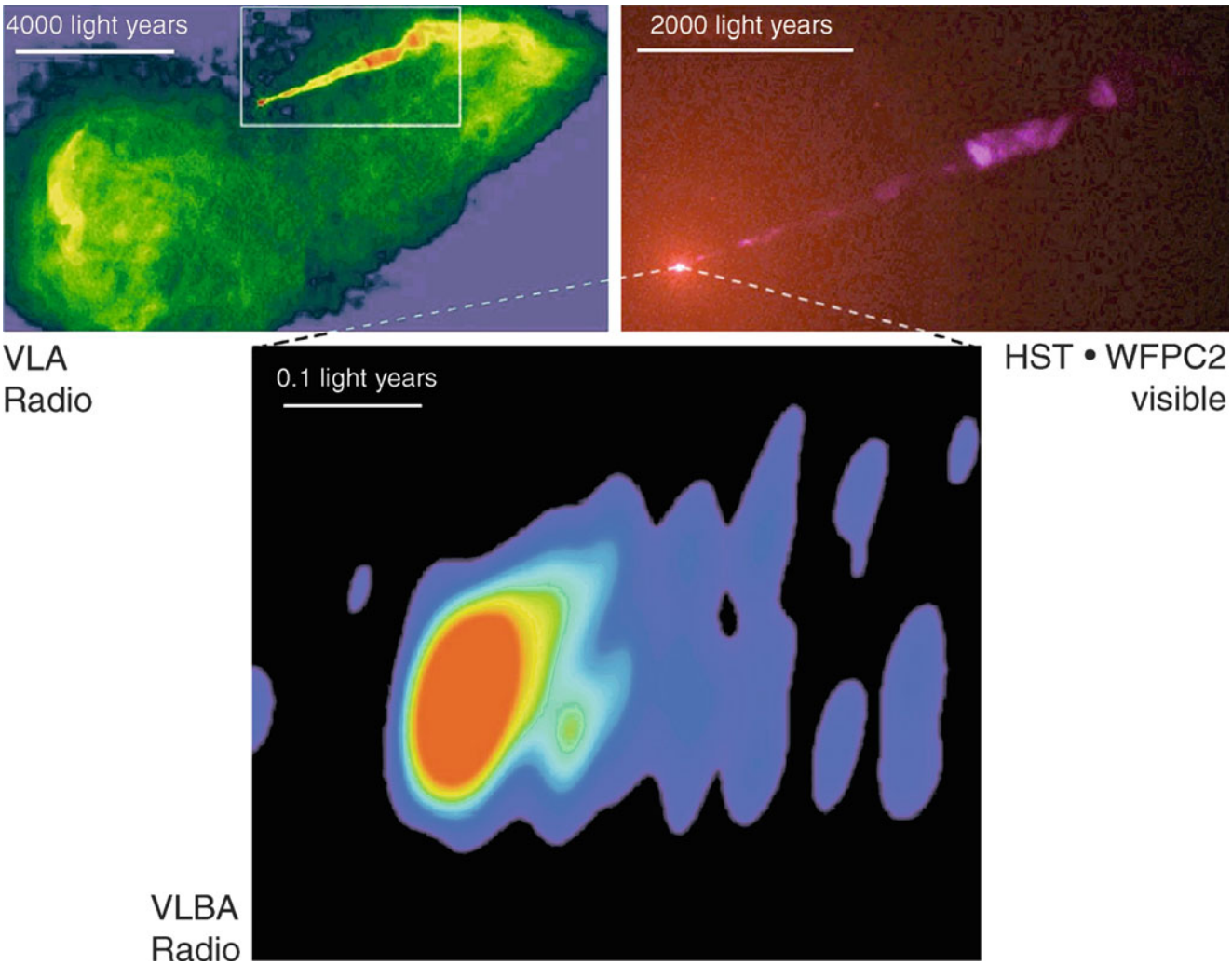


Fig. 5.47 *Top left*: A radio map of M87, the central galaxy in the Virgo cluster of galaxies. *Top right*: An HST image of the region shown in the *inset* of the *left-hand panel*. The radio jet is also visible at optical wavelengths. The *lower image* shows a VLBI map of the region around the galaxy core; the jet is formed within a few 10^{17} cm from the core of the galaxy, which contains a black hole of $M_\bullet \sim 6 \times 10^9 M_\odot$. Very close to the center the opening angle of the jet is significantly larger than further out. This indicates that the jet only becomes collimated at a larger distance. With VLBI observations at 1.3 mm, it was shown that the base of the jet is as compact as 40 μarcsec—corresponding to about $6r_S$! Credit: NASA, National Radio Astronomy Observatory/National Science Foundation, John Biretta (STScI/JHU), and Associated Universities, Inc.

emission by relativistic electrons moving in a magnetic field. This electron population can also produce X-ray photons. However, it is less clear which emission process generates the X-ray emission. The two principal possibilities are (i) synchrotron emission as well, or (ii) inverse Compton scattering. In the first case, the spectrum of the electrons must extend to extremely high energies: If the radio emission is due to synchrotron emission of electrons with $\gamma \sim 10^4$ in a magnetic field with $\sim 100\,\mu$G [cf. (5.3)], then in order to generate X-rays in the same magnetic field, the Lorentz factor must be $\sim 10^8$—corresponding to an electron energy of $E = \gamma m_e c^2 \sim 100$ TeV. It is currently unclear which acceleration processes may account for these high energies.

In the case of inverse Compton scattering, low-energy photons are scattered to much higher energies by collisions with relativistic electrons—a photon of frequency ν typically has a frequency $\nu' \approx \gamma^2 \nu$ after being scattered by an electron of energy $\gamma m_e c^2$ (see also Sect. 5.4.4 for the case of inverse Compton scattering by a thermal distribution of electrons). Since the characteristic Lorentz factors of electrons causing the synchrotron radiation of radio jets may reach values of $\gamma \sim 10^4$, these electrons may scatter, by inverse Compton scattering, radio photons into the X-ray domain of the spectrum. If the radio photons are those produced in situ by synchrotron emission, this effect is also called synchrotron self-Compton radiation. In particular, if the low-frequency spectrum is a power law, then the inverse

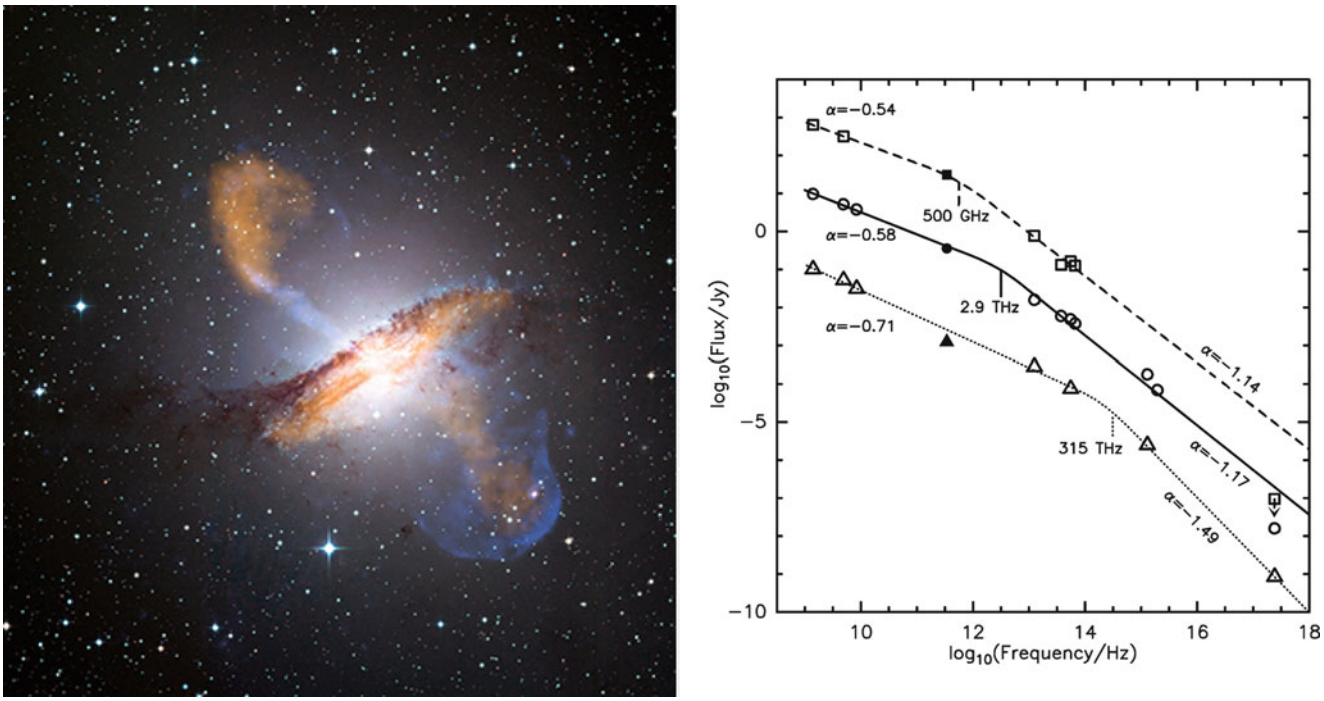


Fig. 5.48 *Left panel*: A composite image of the radio galaxy Centaurus A (NGC 5128), one of the first active galaxies discovered. This image is 16′ across, and combines optical data from the Wide Field Imager at the MPG/ESO 2.2 m telescope at La Silla (shown in *white*), the X-ray data from the Chandra observatory (*blue*), and the sub-millimeter data from the APEX telescope in Chile (*orange*). The dust lane, clearly seen in the optical image, emits strongly in the sub-millimeter regime, due to dust heated by starlight. Furthermore, APEX and Chandra display the AGN activity in this source, due to jets and lobes, which are located in a direction roughly perpendicular to the dust lane. Thus, the origin of the sub-millimeter emission from the dust lane and the jet/lobe regions is quite different: in the former case it is due to warm dust with a temperature of $T_{\rm d} \sim 30$ K, as obtained by combining the APEX data with far-IR data from the ISO satellite, whereas in the latter case it is due to synchrotron emission. *Right panel*: By combining the data shown in the image with radio and optical data of the jet region, one finds that the spectrum can be well described by a broken power law, from the radio to the X-rays, which is the expected spectral behavior of a synchrotron source where the population of relativistic electrons cools due to emission. The three sets of data, and the corresponding lines, are the spectra at different locations along the jet/lobe, with the lowest one being closest to the galaxy nucleus, and the *dashed* one being furthest away. As can be seen, the break frequency, i.e., the frequency where the spectrum steepens, decreases as one gets further out into the lobe, again as expected from a cooling electron distribution. Credit: *Left*: X-ray: NASA/CXC/CfA/R. Kraft et al.; Sub-millimeter: MPIfR/ESO/APEX/A. Weiss et al.; optical: ESO/WFI. *Right*: A. Weiss et al. 2008, *LABOCA observations of nearby, active galaxies*, A&A 490, 77, p. 85, Fig. 11. ©ESO. Reproduced with permission

Compton spectrum will be a power law as well. Alternatively, relativistic electrons can also scatter optical photons from the AGN, for which less energetic electrons are required. The omnipresent CMB may also be considered as a photon source for the inverse Compton effect, and in many cases the observed X-ray radiation is probably Compton-scattered CMB radiation.

It is usually very difficult to distinguish between these alternatives, and it is by no mean clear that the high-energy emission has the same origin in all sources. For at least one source, there is good evidence that the synchrotron self-Compton effect is at play. For that, we start with a simple thought experiment: suppose that in a synchrotron source the number density of relativistic electron would suddenly double. In this case, the synchrotron emission would simply double as well, since it is linear in the electron number. However, the synchrotron self-Compton emission would increase by a factor of four, since both the number of low-energy (synchrotron) photons, as well as the number of electrons doubles, and hence there are four times as many scatterings between low-energy photons and electrons. Hence, for synchrotron self-Compton, one expects that there is a quadratic dependence of the Comptonized flux on the synchrotron flux.

This argument, however, assumes that the source is optically thin with respect to Compton scattering. In the most compact sources, this may not be the case, and the Comptonized photons may be upscattered again. This higher-order scattering emission would then, to first order, have a cubic dependence on the electron density—and hence a cubic dependence on the synchrotron emissivity.

The BL Lac object PKS 2155−304 was observed simultaneously in the optical, X-ray and TeV gamma regime when it had a major flare; this flare, which happened on a time-scale of ~ 1 h, led to flux variations by more than a factor of 20 in the TeV region, whereas the variations were merely a factor

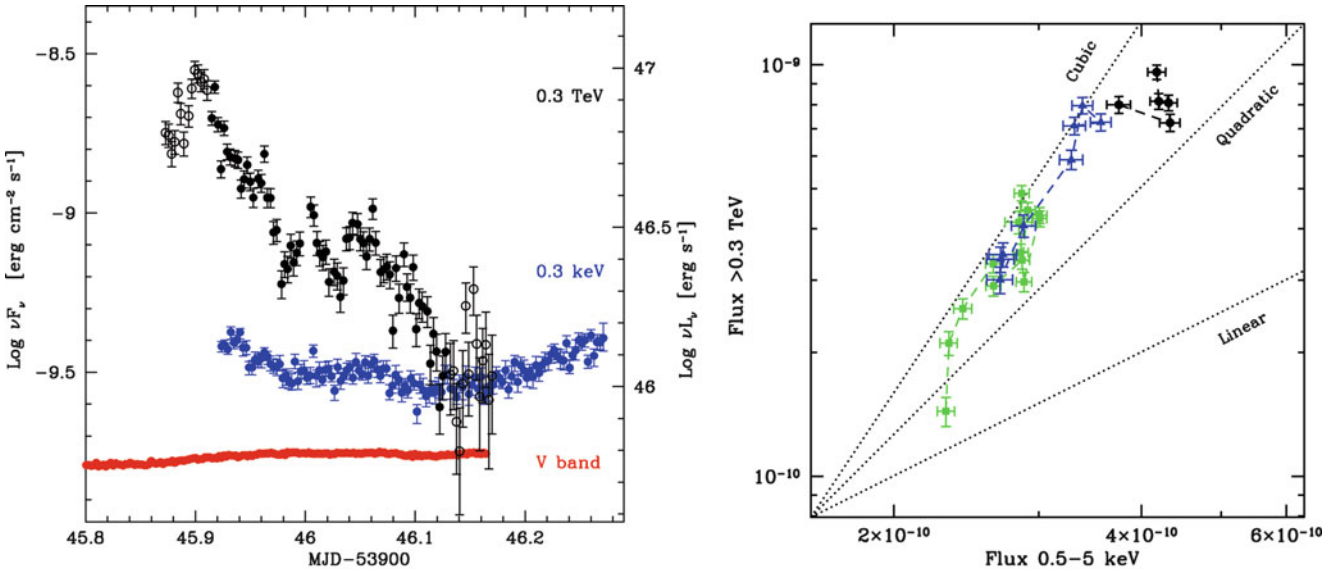


Fig. 5.49 A spectacular flare of the BL Lac object PKS 2155−304. In the *left panel*, the light curves of the source in the optical (*red*), X-ray (*blue*) and high-energy gamma radiation (*black*) at $E = 0.3$ TeV are shown. Note the very much enhanced range of fluxes in the gamma rays. In the *right panel*, the TeV flux is plotted against the X-ray flux, for several time intervals during the flare. As can be seen, the TeV flux varies approximately as the cube of the X-ray flux. Source: F. Aharonian et al. 2009, *Simultaneous multiwavelength observations of the second exceptional γ-ray flare of PKS 2155−304 in July 2006*, A&A 502, 749, p. 754, 762, Figs. 3, 13. ©ESO. Reproduced with permission

of 2 in the X-ray flux (and 15 % in the optical)—see the left panel of Fig. 5.49. The TeV light curve showed similar features as the X-ray (at 0.3 keV) light curve, however with quite a different scaling: correlating the TeV flux with that of the X-ray flux, a behavior $S_{\mathrm{TeV}} \propto S_{\mathrm{X}}^3$ was found (right panel of Fig. 5.49)—i.e., a behavior that is predicted for the case that the TeV radiation is second-order inverse Compton scattering and the X-ray emission is due to synchrotron radiation. Although the actual processes in the source are most likely much more complicated than outlined in this simple thought experiment, it nevertheless indicates that the synchrotron self-Compton process is acting in this source.

The inverse Compton model cannot, however, be applied to all X-ray jets without serious problems occurring. For instance, variability in X-ray emission was observed in the knots of M87, indicating a very short cooling time for the electrons. Since the electrons must have a much larger Lorentz factor γ if the radiation, at a given frequency, originates from synchrotron emission, compared to the case that the X-rays are produced by inverse Compton scattering, their cooling time t_{cool} (5.6) would be much shorter as well. In such sources, which are typically FR I radio sources, the variability therefore argues for the synchrotron process to be responsible for the X-ray emission. The implied very short cooling time-scales then leads to an increased necessity for a local acceleration of the electrons. On the other hand, the required energies for the electrons are very high, ~ 100 TeV.

Detecting radio jets at X-ray frequencies seems to be a frequent phenomenon: about half of the flat-spectrum radio QSO with jet-like extended radio emission also show an X-ray jet. All of those are one-sided, although the corresponding radio images often show lobes also on the other side of the X-ray jets, reinforcing the necessity for Doppler favoritism also in the X-ray waveband (Fig. 5.50).

TeV emission from blazars. Blazars emit at the highest photon energies yet observed, in the TeV $= 10^{12}$ eV spectral range, as can be observed with ground-based Cherenkov telescopes (see Sect. 1.3.6). Furthermore, they are the dominant population of extragalactic sources observed in the GeV-range. At TeV energies, they can exhibit strong flux variations on time-scales of minutes (see Fig. 5.49 for an example). The broad-band spectral energy distribution of these TeV blazars show two dominant broad peaks, one located in the X-ray regime of the spectrum, the other at TeV energies, with approximately equal energy output in both. As mentioned before, the correlated variability in both spectral ranges clearly argues for a synchrotron self-Compton origin of the high-energy gamma radiation.

The fact that we can see this emission is another clear argument for the highly relativistic nature of the source. If we ignore relativistic effects for a moment, the small time-scale of variability yields an upper bound on the source size. Together with the observed flux, and thus an estimated luminosity, one can calculate the energy density inside the source. This density is so high that no TeV photon could escape the emitting region—they would scatter with lower energy photons to produce electron-positron pairs. In order to avoid

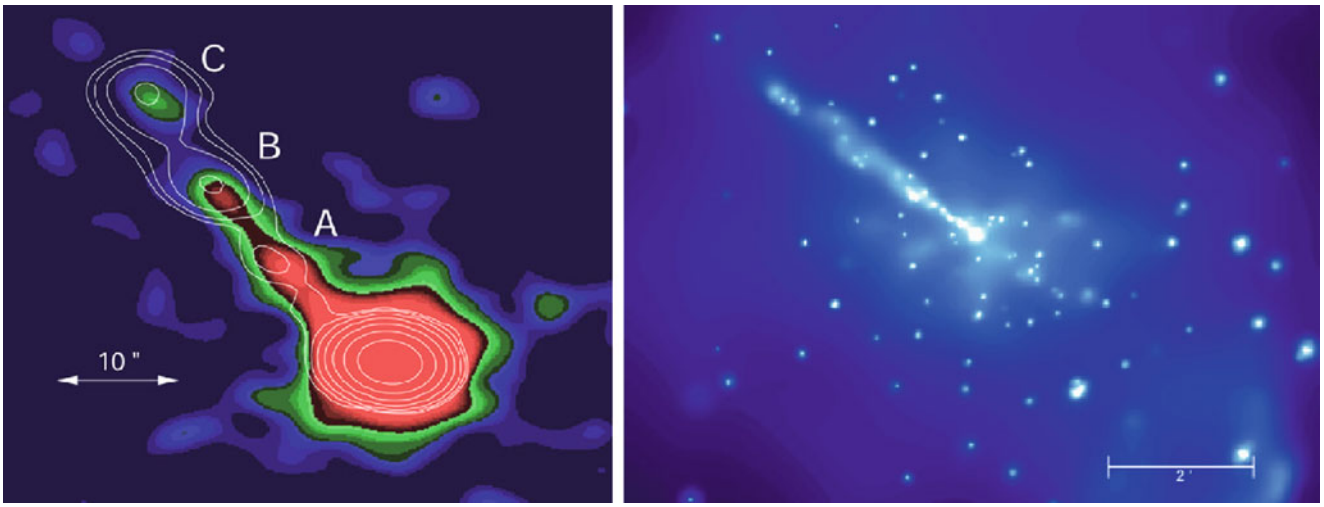


Fig. 5.50 X-ray images of AGN jets. *On the left*, a Chandra image of the jet in the QSO PKS 1127−145, with overlaid contours of radio emission (1.4 cm, VLA). The direction of the jet and its substructure are very similar at both wavelengths, suggesting an interpretation in which the radiation is caused by the same population of relativistic electrons. *On the right*, a Chandra image of the active galaxy Centaurus A. Here the jet is visible, as well as a large number of compact sources interpreted to be X-ray binaries. Credit: *Left*: X-ray: NASA/CXC/A. Siemiginowska/CfA & J. Bechtold/University of Arizona; radio: Siemiginowska et al./VLA. *Right*: NASA/SAO/R. Kraft et al.

this conclusion, the intrinsic luminosity must be smaller and the source size larger. Both is the case if we allow for a relativistic source component for which significant beaming occurs, with a Doppler factor $\mathcal{D} = [\gamma(1 - \beta \cos\phi)]^{-1}$ of order 10 or larger. This argument also implies that we are located in a direction very close to the jet axis.

5.5.5 Unified models—summary

After discussing the various components of the unified model, we summarize here our current picture of the AGN phenomenon, referring back to Fig. 5.12.

In common of all AGNs is a central SMBH, powered by an accretion disk. Surrounding gas is photoionized and radiates emission lines, with the broad lines coming from a region with a size of $\sim 10^3$ Schwarzschild radii around the black hole (and that of the highest ionization species from an even smaller region), and the narrow lines from a much more extended region. The strong infrared emission in all AGNs argues for the presence of obscuring matter, which is concentrated towards the plane in which the accretion occurs—the torus. There is a large spread in the ratio of radio-to-optical luminosity of AGNs, which presumably is related to the different efficiency of jet formation.

Our classification of an AGN depends strongly on the viewing direction, or the inclination angle. Type 1 AGNs are those where our view to the central engine is not blocked by the absorbing material of the torus, and we can see the BLR. At higher inclination, the view towards the BLR is no longer possible, and the object appears as a Type 2 AGN. Depending on the luminosity of the AGN and its radio-to-optical ratio, the Type 1's are either QSOs, Seyfert 1 galaxies or broad line radio galaxies, whereas for Type 2's, they are Type 2 QSOs, Seyfert 2 galaxies or narrow line radio galaxies. Blazars are sources in which we look directly into the relativistic jet, to within a few degrees of the jet axis, where phenomena of Special Relativity (like beaming) can explain the unusual spectral properties and rapid variability of these sources.

Refinement. The fact that the torus is probably a collection of optically thick clouds modifies this simple picture in a slight way. The question of whether a sight-line to the central AGN is absorbed is now one of probability: With the torus being in the equatorial plane, it is much more likely that our view to the AGN is absorbed if we are in this plane, compared to being close to the symmetry axis. However, there is a finite probability that we see through the torus (and thus classify the source as Type 1) even if we are close to the equatorial plane, and conversely, there is a finite probability to have an absorbed sight-line to the AGN even if being close to the symmetry axis. These refinements of the unification scheme are needed to understand some objects with observed properties which can place them in either (or neither) category.

Quasar mode vs. radio mode. Geometry and orientation are not the only effects which determine the appearance of an AGN. For example, the spin of the black hole may play an important role in the behavior of an accretion flow. Furthermore, the efficiency of an accretion flow to launch a jet probably depends on the accretion rate. If this is

sufficiently low, the accretion disk will become optically thin, and can not longer radiate efficiently (see Sect. 5.3.2). In this case, most of the energy is advected inwards, and there are strong indications that this provides a very favorable situation for launching a jet. Indeed, there is a very strong correlation between the ratio of the emitted fluxes in the radio and optical spectral range, and the Eddington ratio $L/L_{\rm edd}$. Sources for which the latter is high, i.e., where the accretion occurs through an optically thick disk, emit only a small fraction of their luminosity at radio wavelengths, indicating rather inefficient jet production. This can also be seen in Fig. 5.13, in which the radio-to-optical luminosity increases from the bottom right to the upper left corner. The luminous QSOs (star symbols) are located towards the bottom right corner, whereas the FR I galaxies are near the upper left one.

It thus seems that, depending on the accretion rate, a black hole can either shine through the 'QSO mode', where the luminosity is dominated by quasi-thermal radiation from the accretion disk, or through a 'radio mode', in which the disk does not radiate efficiently, and a large fraction of the energy is channeled into an outflow, visible in form of a jet.

Our attempts at finding a unification scheme for the different classes of AGNs have been quite successful. The scheme of unification is generally accepted, even though some aspects are still subject to discussion and require further studies.

5.5.6 Tidal disruption events

Our AGN model connects the activity with the accretion of gas onto a SMBH in the center of a galaxy. Despite the fact that all galaxies (with a spheroidal stellar component) host a SMBH, only those where gas can flow inwards and accrete show AGN activity.

Disruption of a star. Situated in a galactic nucleus, from time to time a star on its orbit may come close to the SMBH—perhaps even too close for its survival. We have seen in Sect. 2.3.6 that tidal gravitational forces can disrupt a system of particles—or a star. The condition for this to happen is [cf. (2.47)]

$$\frac{M_*}{r_*^3} \lesssim \frac{M_\bullet}{R^3} \,, \tag{5.40}$$

where M_* and r_* denote the stellar mass and radius, and R is its distance to the SMBH. Once a star satisfies this condition, it will be disrupted in the tidal field of the SMBH.

Expectations. The consequences of such an event were studied theoretically already in the 1980s: Following this process, about half of the stellar mass will be forced into a bound orbit around the black hole, while the other half is ejected. The bound mass will then be accreted onto the black hole; the initial phase of this accretion process proceeds rapidly, whereas the accretion rate at later times is expected to decreases with time as $t^{-5/3}$, if t is measured from the time of disruption.

This process is expected to have clear observational signatures: The initial accretion event should show up as a bright flare, perhaps even close to the Eddington luminosity of the SMBH. At later times, the luminosity should decrease roughly in proportion to the accretion rate, i.e., $\propto t^{-5/3}$. This flare should occur in the center of a galaxy, even if that galaxy has shown no sign of nuclear activity before.

These events should be quite rare; from the density of stars in the center of galaxies and stellar dynamics, typical event rate are estimated to be of order $\sim 10^{-5}$ per galaxy per year. Hence, they may only be detected by monitoring a large number of galaxies.

Detection of stellar disruption events. Not surprisingly, the first such tidal disruption events (TDEs) were found in the X-rays—since in this spectral regime, radiation from the AGN sticks out most clearly from the emission of the host galaxy. Repeated observations by the ROSAT satellite led to the discovery of the first TDEs in the late 1990s. The soft X-ray luminosity at peak brightness is huge, reaching $10^{44}\,{\rm erg/s}$ or even higher. Long-term monitoring shows that the decline of the luminosity is consistent with the $t^{-5/3}$-law predicted from the accretion of tidal debris—for the first events, the decline could be followed over more than a decade, showing a decrease from the peak flux by more than three orders of magnitude. From modeling the events, typical SMBH masses of 10^6 to $10^8\,M_\odot$ are derived, as expected from the scaling relation between $M_\bullet$ and the properties of the stellar population of the host galaxy. Despite the small number statistics of events (of order a dozen have been observed by now), the event rate is compatible with the theoretical expectations.

One of the most recent events was discovered by the Swift satellite, both in soft and hard X-rays. The peak flux was at least 10^4 times larger than the X-ray flux 20 years before the TDE, and at least a factor of 100 larger than a year before the event, as obtained from upper flux limits derived from previous X-ray observations of the galaxy (at $z = 0.35$). In fact, the peak flux corresponds to a luminosity of $\sim 10^{48}\,{\rm erg/s}$—corresponding to the Eddington luminosity of a SMBH with $M_\bullet \sim 10^{10} M_\odot$. However, the actual black hole mass must be far smaller: strong variability of the X-ray flux was seen on a time-scale of 100 s, which implies that the light-travel time across the Schwarzschild radius cannot be larger than this value. This argument yields an upper bound of $M_\bullet \leq 7 \times 10^6 M_\odot$. Hence, during its peak the apparent

luminosity of the flare was much larger than the Eddington luminosity of the SMBH.

The solution of this apparent discrepancy is provided from radio observations of this event; from the temporal and spectral properties seen at radio frequencies, it is suspected that the TDE launched a relativistic jet. The associated beaming can explain the huge apparent X-ray luminosity. Curiously, integrating the X-ray luminosity over the first 50 days of the flare yields a total energy of $\sim 10^{53}$ erg—which is close to the energy generated by accreting $1 M_\odot$ with an efficiency of 10 %.

TDEs not only provide a clear confirmation that inactive SMBHs can be revived, but they will allow us to study details of the accretion physics, provided they are followed with sufficient time coverage and depth. The onset of the formation of a jet is just one example of what we can learn from these events. The future eROSITA survey will obtain several all-sky surveys and is ideally placed to discover many such events.

5.6 Properties of the AGN population

AGNs, and QSOs in particular, are visible out to very high redshifts. Since their discovery in 1963, QSOs have held the redshift record most of the time. Only in recent years have QSOs and galaxies been taking turns in holding the record. Today, several hundred QSOs are known with $z \geq 4$, and the number of those with $z > 5$ continues to grow since a criterion was found to identify these objects. This leads to the possibility that QSOs could be used as cosmological probes, and thus to the question of what we can learn about the Universe from QSOs. For example, one of the most exciting questions is how does the QSO population evolve with redshift—was the abundance of QSO at high redshifts, i.e., at early epochs of the cosmos, similar to that today, or does it evolve over time?

5.6.1 The K-correction

To answer this question, we must know the luminosity function of QSOs, along with its redshift dependence. As we did for galaxies, we define the luminosity function $\Phi(L, z)\,\mathrm{d}L$ as the spatial number density of QSOs with luminosity between L and $L + \mathrm{d}L$. Φ normally refers to a comoving volume element, so that a non-evolving QSO population would correspond to a z-independent Φ. One of the problems in determining Φ is related to the question of which kind of luminosity is meant here. For a given observed frequency band, the corresponding rest-frame radiation of the sources depends on their redshift. For optical observations, the measured flux of nearby QSOs corresponds to the rest-frame optical luminosity, whereas it corresponds to the UV luminosity for higher-redshift QSOs. In principle, using the bolometric luminosity would be a possible solution; however, this is not feasible since it is *very* difficult to measure the bolometric luminosity (if at all possible) due to the very broad spectral energy distribution of AGNs. Observations at all frequencies, from the radio to the gamma domain, would be required, and obviously, such observations can only be obtained for selected individual sources.

Of course, the same problem occurs for all sources at high redshift. In comparing the luminosity of galaxies at high redshift with that of nearby galaxies, for instance, it must always be taken into account that, at given observed wavelength, different spectral ranges in the galaxies' rest-frames are measured. This means that in order to investigate the optical emission of galaxies at $z \sim 1$, observations in the NIR region of the spectrum are necessary.

Frequently the only possibility is to use the luminosity in some spectral band and to compensate for the above effect as well as possible by performing observations in several bands. For instance, one picks as a reference the blue filter which has its maximum efficiency at ~ 4500 Å and measures the blue luminosity for nearby objects in this filter, whereas for objects at redshift $z \sim 0.7$ the intrinsic blue luminosity is obtained by observing with the I-band filter, and for even larger redshifts observations need to be extended into the near-IR. The observational problems with this strategy, and the corresponding corrections for the different sensitivity profiles of the filters, must not be underestimated and are always a source of systematic uncertainties. An alternative is to perform the observation in only one (or a few) filters and to approximately correct for the redshift effect.

In Sect. 4.3.3, we defined various distance measures in cosmology. In particular, the relation $S = L/(4\pi D_\mathrm{L}^2)$ between the observed flux S and the luminosity L of a source defines the luminosity distance D_L. Here both the flux and the luminosity refer to bolometric quantities, i.e., flux and luminosity integrated over all frequencies. Due to the redshift, the measured spectral flux S_ν is related to the spectral luminosity $L_{\nu'}$ at a frequency $\nu' = \nu(1 + z)$, where one finds

$$S_\nu = \frac{(1+z) L_{\nu'}}{4\pi D_\mathrm{L}^2} \,. \tag{5.41}$$

We write this relation in a slightly different form,

$$S_\nu = \frac{L_\nu}{4\pi D_\mathrm{L}^2} \left[\frac{L_{\nu'}}{L_\nu} (1+z) \right] , \tag{5.42}$$

where the first factor is of the same form as in the relation between the bolometric quantities while the second factor corrects for the frequency shift. This factor is denoted the

K-correction. It obviously depends on the spectrum of the source, i.e., to determine the K-correction for a source its spectrum needs to be known. Furthermore, this factor depends on the filter used. Since in optical astronomy magnitudes are used as a measure for brightness, (5.42) is usually written in the form

$$m_{\rm int} = m_{\rm obs} + K(z) \quad \text{with} \quad K(z) = -2.5 \log \left[\frac{L_{\nu'}}{L_\nu} (1+z) \right] , \tag{5.43}$$

where $m_{\rm int}$ is the magnitude that would be measured if the spectrum of the source would not be shifted in wavelength by redshift, and $m_{\rm obs}$ describes the brightness actually observed. The K-correction is not only relevant for QSOs but for all objects at high redshift, in particular also for galaxies.

5.6.2 The luminosity function of QSOs

Construction of the luminosity function. By counting QSOs, we obtain the number density $N(>S)$ of QSOs with a flux larger than S. We find a relation of roughly $N(>S) \propto S^{-2}$ for large optical fluxes S, whereas the source counts are considerably flatter for smaller fluxes. The flux at which the transition from steep counts to flatter ones occurs corresponds to an apparent magnitude of about $B \sim 19.5$. About 10 QSOs per square degree are found brighter than this break magnitude.

From QSO number counts, combined with measurements of QSO redshifts, the luminosity function $\Phi(L, z)$ can be determined. As already defined above, $\Phi(L, z)\,{\rm d}L$ is the number density *in a comoving volume element* of QSOs at redshift z with a luminosity between L and $L + {\rm d}L$.

Two fundamental problems exist in determining the luminosity function. The first is related to the above discussion of wavelength shift due to cosmological redshift: a fixed wavelength range in which the brightness is observed corresponds to different wavelength intervals in the intrinsic QSO spectra, depending on their redshift. We need to correct for this effect if the number density of QSOs above a given luminosity in a certain frequency interval is to be compared for local and distant QSOs. One way of achieving this is by assuming a universal spectral shape for QSOs; over a limited spectral range (e.g., in the optical and the UV ranges), this assumption is indeed quite well satisfied. This universal spectrum is obtained by averaging over the spectra of a larger number of QSOs (Fig. 5.2). By this means, a useful K-correction of QSOs as a function of redshift can then be derived.

The second difficulty in determining $\Phi(L, z)$ is to construct QSO samples that are 'complete'. Since QSOs are point-like they cannot be distinguished from stars by morphology on optical images, but rather only by their color properties and subsequent spectroscopy. However, with the star density being much higher than that of QSOs, this selection of QSO candidates by color criteria, and subsequent spectroscopic verification, is very time-consuming. Only more recent surveys, which image large areas of the sky in several filters, were sufficiently successful in their color selection and subsequent spectroscopic verification, so that very large QSO samples could be compiled. An enormous increase in statistically well-defined QSO samples was achieved by two large surveys with the 2dF spectrograph and the Sloan Digital Sky Survey which we discuss in the context of galaxy redshift surveys in Sect. 8.1.2.

The optical QSO luminosity function. The luminosity function that results from such analyses is typically parametrized as

$$\Phi(L, z) = \frac{\Phi^*}{L^*(z)} \left[\left(\frac{L}{L^*(z)} \right)^{\gamma_1} + \left(\frac{L}{L^*(z)} \right)^{\gamma_2} \right]^{-1} ; \tag{5.44}$$

i.e., for fixed z, Φ is a double power law in L. At $L \gg L^*(z)$, the second term in the square brackets in (5.44) dominates if $\gamma_2 > \gamma_1$, yielding $\Phi \propto L^{-\gamma_2}$. On the other hand, the first term dominates for $L \ll L^*(z)$, so that $\Phi \propto L^{-\gamma_1}$. Typical values for the exponents are $\gamma_1 \approx 1.5$, $\gamma_2 \approx 3.5$. The characteristic luminosity $L^*(z)$ where the L-dependence changes, strongly depends on redshift. A good fit to the data for $z \lesssim 2$ is achieved by

$$L^*(z) = L_0^* (1+z)^k , \tag{5.45}$$

with $k \approx 3.45$, where the value of k depends on the assumed density parameters $\Omega_{\rm m}$ and Ω_Λ. This approximation is valid for $z \lesssim 2$, whereas for larger redshifts $L^*(z)$ seems to vary less with z. The normalization constant is determined to be $\Phi^* \approx 5.2 \times 10^3\, h^3\, {\rm Gpc}^{-3}$, and L_0^* corresponds to roughly $M_{\rm B} = -20.9 + 5 \log h$. The luminosity function as determined from the combined 2dF and SDSS surveys is plotted in Fig. 5.51.

From this luminosity function, a number of conclusions can be drawn. The luminosity function of QSOs is considerably broader than that of galaxies, which we found to decrease exponentially for large L, compared to the power-law behavior we see here. The strong dependence of the characteristic luminosity $L^*(z)$ on redshift, which is seen in Fig. 5.51 as a systematic shift of the turnover luminosity towards fainter values as z decreases, clearly shows a very significant cosmological evolution of the QSO luminosity function. For example, at $z \sim 2$, $L^*(z)$ is about 50 times larger than today. Furthermore, for high luminosities, $\Phi \propto [L^*(z)]^{\gamma_2 - 1}\, L^{-\gamma_2}$. This means that the spatial number density of luminous QSOs was more than 1000 times larger at $z \sim 2$ than it is today, which can also be seen directly by

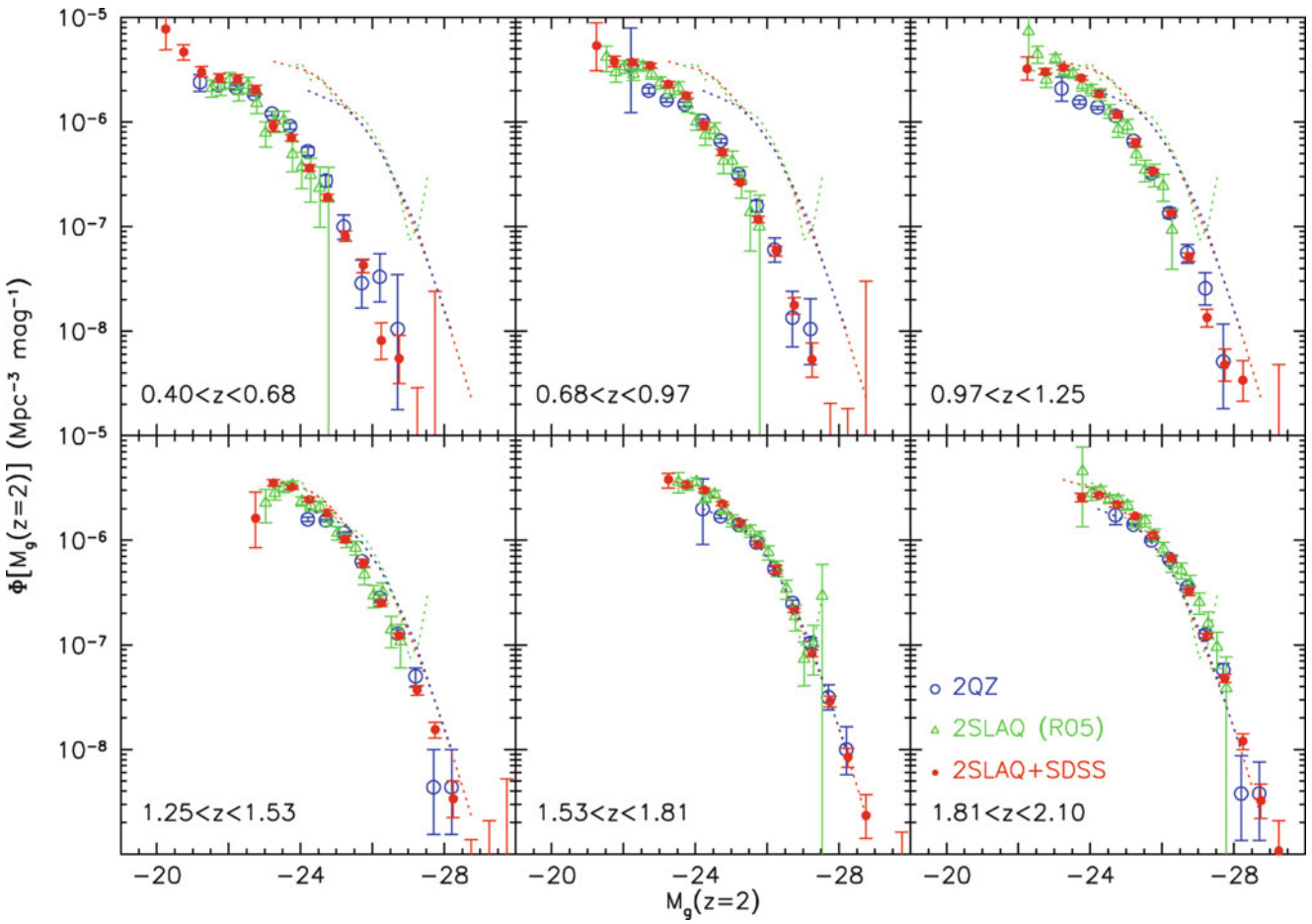


Fig. 5.51 The luminosity function from the combined 2dF and SDSS QSO surveys (*red points*), compared to earlier surveys, plotted for six different redshift intervals as indicated. The *dotted curves* in each panel show the fit to the data of the redshift bin $1.53 < z < 1.81$. For obtaining these results, several corrections were necessary, including K-correction and to account for the light from the host galaxy. Source: S.M. Croom et al. 2009, *The 2dF-SDSS LRG and QSO survey: the QSO luminosity function at* $0.4 < z < 2.6$, MNRAS 399, 1755, p. 1764, Fig. 11. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

comparing the data points in the lowest-redshift panel with the dotted curve in Fig. 5.51—the low-redshift luminosity function does not extend to the very bright luminosities for which the luminosity function at high redshifts was measured. The number density of very luminous QSOs at low redshifts is so small that essentially none of them are contained in the survey volume from which the results in Fig. 5.51 were derived. This point is better illustrated in Fig. 5.52 which shows the evolution of the number density of QSOs as a function of redshift, for different bins in luminosity. High-L QSOs are very rare in the current Universe, and in particular, the ratio of the number density of high- to-low luminosity QSOs strongly decreases towards lower redshifts.

AGN selection in X-rays. Finding QSOs at higher redshifts with optical methods is more difficult, as we discussed before—the observed optical colors change as the strong emission lines move into the optical bands. Furthermore, optical surveys are not best suited for finding low-luminosity AGNs; for them, the optical light of the host galaxy renders the AGN less pronounced, and thus it is more difficult to identify it as such without spectroscopy.

Obtaining a complete census of the AGN population is much easier by X-ray selection, for a number of reasons. First, all known AGN-types emit X-rays; furthermore, the fraction of the total energy that is emitted in the form of X-rays is less dependent on the AGN type than is the case for the optical emission. Second, the X-ray emission from galaxies is weak, and thus the AGN sticks out clearly. In fact, in a high Galactic latitude field, $\sim$ 90 % of all X-ray sources are AGNs, the rest being galaxy clusters. The fact that far more optically-selected AGNs are known than X-ray-selected ones is due to the small sky areas over which deep X-ray surveys have been carried out.

Indeed, one can estimate that optical surveys miss $\sim$ 80 % of the AGNs at any fixed bolometric luminosity, whereas X-ray surveys are more efficient. Observing at hard X-rays (2–20 keV), one obtains essentially a complete census of AGNs for bolometric luminosities $L_{\rm bol} \gtrsim 10^{45}$ erg/s, and

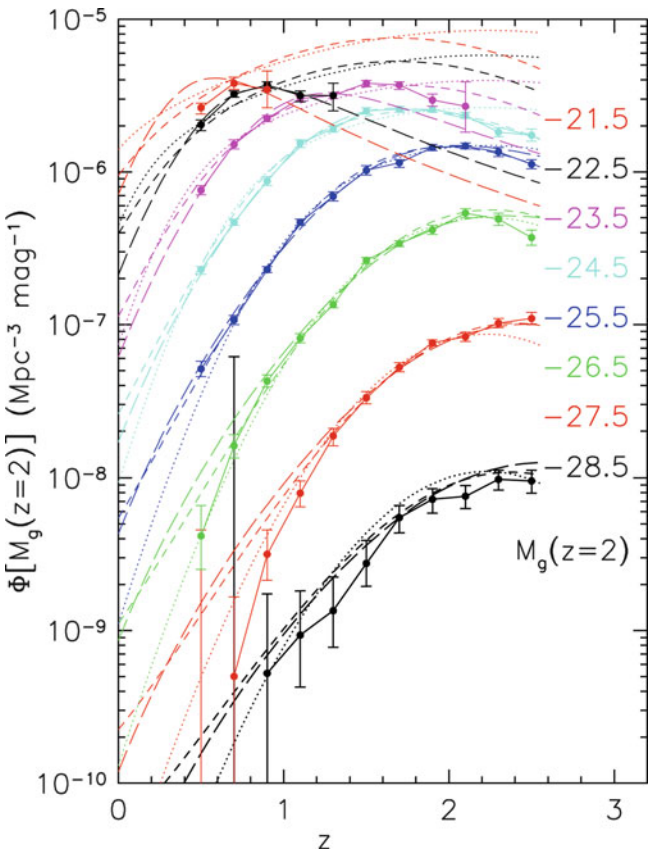


Fig. 5.52 The comoving spatial density of QSOs per magnitude interval, as obtained from the combined 2dF and SDSS QSO surveys, plotted for eight different bins in absolute magnitude. The *solid curves* connect the data points, whereas the other *curves* show various evolution models. As already seen from Fig. 5.51 the density of luminous QSOs steadily decreases from $z \sim 2$ until today. This decrease is also seen for somewhat less luminous QSOs, however it is less steep. In particular, for the less luminous ones, it appears as if the maximal space density is at redshifts smaller than 2. The redshift of the peak density of QSOs decreases towards lower luminosity sources. Source: S.M. Croom et al. 2009, *The 2dF-SDSS LRG and QSO survey: the QSO luminosity function at* $0.4 < z < 2.6$, MNRAS 399, 1755, p. 1764, Fig. 10. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

a moderate incompleteness of $\sim$ 30 % for a factor of 100 less luminous sources. Despite the higher sensitivity of soft X-ray (0.5–2 keV) detectors, the incompleteness here is larger, of order 50 % (again depending on luminosity). The reason for this behavior is the photoelectric absorption in the source, perhaps from the gas in the torus, combined with the frequency dependence of this effect. Thus, the emission in the soft X-ray band is more sensitive to this absorption, whereas the hard X-rays are much less affected. Only if the column density of hydrogen approaches $N_{\rm H} \sim 10^{24}\,{\rm cm}^{-2}$, and the source starts to become optically thick with respect to Compton scattering ('Compton thick AGN'), are the sources difficult to detect in hard X-rays.

We thus conclude that the high AGN completeness of X-ray observations makes this the preferred band for selection. As a drawback, however, one requires in addition an optical identification and spectrum, to obtain the redshift of the source.

Bolometric luminosity function. Given the broad-band spectral energy distribution of AGNs, together with observational constraints on the distribution of absorber column densities, the luminosity in one spectral band can be used to estimate the bolometric luminosity of sources. These estimates can be cross-checked by requiring that the bolometric luminosity determined, say, from the X-ray flux agrees with that obtained from the optical data, and that the distribution of the bolometric luminosity functions obtained from both band mutually agree. In this way, it is possible to construct the bolometric luminosity function of AGNs.

In Fig. 5.53 the parameters of the double power-law fit (5.44) to the luminosity function are displayed, now for the bolometric luminosity function. The bright- and faint-end slopes γ_i were allowed to depend on redshift as well. As the upper left and middle panel of Fig. 5.53 show, the data are not compatible with constant slopes; in particular, towards very high redshifts, the bright end of the luminosity function flattens. The overall normalization Φ^* is essentially constant.

There is a dramatic evolution in the break luminosity L^* with redshift, as seen in the lower left panel, increasing by a factor $\sim$ 30 from today to redshift $z \sim 2$, and then decreasing towards even higher redshifts. This drastic change is accompanied by a strong evolution of the bolometric luminosity density of AGNs, obtained by integrating the luminosity function (5.44) over L, as shown in the lower middle panel: The energy output from AGNs at redshift $z \sim 2$ was an order of magnitude larger than it is today, and decreases towards very high redshift. The drastic decrease in the space density of luminous optical QSOs, seen in Fig. 5.52, is matched by the model: the abundance of the most luminous AGNs shows a very strong drop from its maximum at $z \sim 2.5$ towards lower and higher redshifts.

Hence we conclude that the AGN population displays a remarkable cosmic evolution. AGN activity was peaked at redshifts $z \sim 2.5$, and the decrease towards earlier and later times is very pronounced. Given that the width of the peak, $\delta z \sim 1$, corresponds to a cosmologically relatively short time-scale, a kind of 'quasar epoch' happened in our Universe, in the sense that the (luminous) QSO population seems to have quickly formed and then largely became extinct again. For less luminous AGNs, the peak density occurs at lower redshifts, as can be seen also from Fig. 5.52. It thus appears that the most luminous AGNs formed first, and the less luminous ones at later epochs.

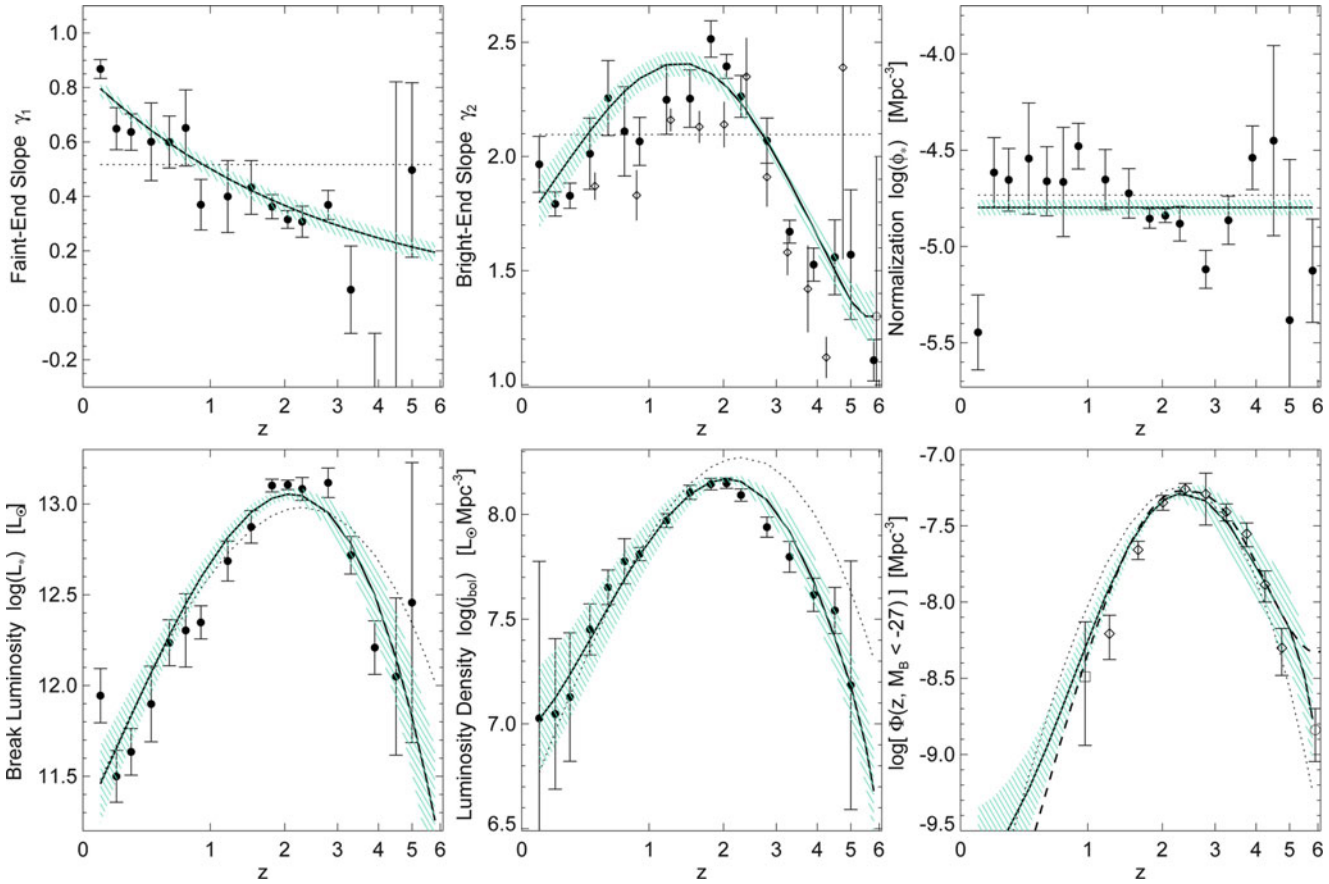


Fig. 5.53 Redshift-dependent parameters of the double power-law fit (5.44) to the bolometric luminosity function of AGNs. The *top row* shows the faint- and bright-end slopes γ_1, γ_2, and the normalization Φ^*. The *bottom row* shows the break luminosity L^*, the luminosity density and the abundance of optically luminous AGNs. The *solid curves* show the parameter fit to the data, the *dotted curves* correspond to a simple model in which the abundance is kept fixed, but the luminosity of each source evolves in redshift. This simple model (pure luminosity evolution) does not yield an acceptable fit. Source: P.F. Hopkins et al. 2007, *An Observational Determination of the Bolometric Quasar Luminosity Function*, ApJ 654, 731, p. 742, Fig. 8. ©AAS. Reproduced with permission

Interpretation. There are several possible interpretations of the QSO luminosity function and its redshift dependence. One of them is that the luminosity of any one QSO varies in time, parallel to the evolution of $L^*(z)$—this would correspond to a pure luminosity evolution model, indicated by the dotted curves in Fig. 5.53, which is seen to provide a rather poor fit to the data. Most likely this interpretation is wrong, also because it implies that a luminous QSO will always remain luminous. Although the efficiency of energy conversion into radiation is much higher for accretion than for thermonuclear burning, an extremely high mass would nevertheless accumulate in this case. This would then be present as the mass of the SMBH in local QSOs.[15] However, estimates of $M_\bullet$ in QSOs rarely yield values larger than $\sim 3 \times 10^9 M_\odot$.

However, it is by no means clear that a given source will be a QSO throughout its lifetime: a source may be active as a QSO for a limited time, and later appear as a normal galaxy again. It is likely that virtually any massive galaxy hosts a potential AGN. This is clearly supported by the fact that apparently all massive galaxies harbor a central SMBH. If the SMBH is fed by accreting matter, this galaxy will then host an AGN. However, if no more mass is provided, the nucleus will cease to radiate and the galaxy will no longer be active. Our Milky Way may serve as an example of this effect, since although the mass of the SMBH in the center of the Galaxy would be sufficient to power an AGN luminosity of more than 10^{44} erg/s considering its Eddington luminosity (5.25), the observed luminosity is lower by many orders of magnitude.

AGNs are often found in the vicinity of other galaxies. One possible interpretation is that the neighboring galaxy disturbs the gravitational field of the QSO's host, such that the flow of matter towards its central regions is favored where it is accreted onto the central black hole—and 'the

[15] Compare the mass estimate in Sect. 5.3.1 where, instead of 10^7 yr, the lifetime to be inserted here is the age of the Universe, $\sim 10^{10}$ yr.

monster starts to shine'. If this is the case, the luminosity function (5.44) does not provide information about individual AGNs, but only about the population as a whole.

Interpreting the redshift evolution then becomes obvious. The increase in QSO density with redshift in the scenario described above originates from the fact that at earlier times in the Universe, interactions between galaxies and merger processes were significantly more frequent than today. On the other hand, the decrease at very high z is to be expected because the SMBHs in the center of galaxies first need to form, and this obviously happens in the first $\sim 10^9$ yr after the Big Bang. We will see later (Sect. 9.6.2) that the star-formation history of the Universe displays a similar behavior as that of AGNs. In Chap. 10, we will consider models how this behavior can be understood in terms of the evolution of galaxies and their central SMBHs.

Black hole demography. Supermassive black holes grow in mass by accretion. Whereas the population of supermassive black holes can also be changed by merging processes, i.e., in the aftermath of galaxy mergers, the corresponding central black holes will merge as well, the total black hole mass is largely conserved in this case, modulo some general relativistic effects. The accretion is related to the energy release in AGNs; thus one might ask whether the total mass density of black holes at the present epoch is compatible with the integrated AGN luminosity. In other words, can the mass density of black holes be accounted for by the total accretion luminosity over cosmic time, as seen in the AGN population?

The first of these numbers is obtained from the scaling relation between SMBH mass and the properties of the spheroidal components in galaxies, as discussed in Sect. 3.8.3. This yields a value of the spatial mass density of SMBHs in the mass range $10^6 \le M_\bullet/M_\odot \le 5 \times 10^9$ of $\sim 4 \times 10^5 M_\odot/\mathrm{Mpc}^3$, with about a 30 % uncertainty. About a quarter of this mass is contributed by SMBHs in the bulges of late-type galaxies; hence, the total SMBH mass density is dominated by ellipticals.

The overall accreted mass is obtained from the redshift-dependent luminosity function of AGNs, by assuming an efficiency ϵ of the conversion of mass into energy. Indeed, the local mass density of SMBHs is matched if the accretion efficiency is $\epsilon \sim 0.10$, as is expected from standard accretion disk models. It therefore seems that the population of SMBHs located in normal galaxies at the present epoch have undergone an active phase in their past, causing their mass growth.

This argument may be slightly incomplete, in that some fraction of the energy released during the accretion process is converted into kinetic energy, as seen by powerful jets in AGN. This fraction is largely undetermined at present, but may not be negligible. In this case, the true ϵ needs to be somewhat higher than 0.1, which is only possible for black holes which rotate rapidly. In fact, the observed profile of the iron emission line from AGNs indicates black hole rotation. On the other hand, the mass growth of black holes is dominated by AGNs with a high Eddington ratio L/L_{edd}, and we have argued that most of them emit a relatively small fraction of their luminosity in the radio regime—and thus their jet power is relatively low.

A more detailed comparison between the SMBH and AGN populations reveals that the characteristic Eddington ratio is $\lambda_{\mathrm{Edd}} \sim 0.3$. With this value, combined with (5.30), one can estimate the mean time-scale over which a typical SMBH was active in the past, yielding $t_{\mathrm{act}} \sim 2 \times 10^8$ yr. Hence, the SMBH of a current day massive galaxy was active during about 2 % of its lifetime.

5.7 Quasar absorption lines

The optical/UV spectra of QSOs are characterized by strong emission lines. In addition, they also show absorption lines, which we have not mentioned thus far. Depending on the redshift of the QSOs, the wavelength range of the spectrum, and the spectral resolution, QSO spectra may contain a large variety of absorption lines. In principle, several different explanations for their occurrence exist. They may be caused by absorbing material in the AGN itself or in its host galaxy, so they have an intrinsic origin. Alternatively, they may arise during the long journey between the QSO and us due to intervening gas along the line-of-sight. We will see that different kinds of absorption lines exist, and that both of these possibilities indeed occur. The analysis of those absorption lines which do not have their origin in the QSO itself provides information about the gas in the Universe. For this purpose, a QSO is basically a very distant bright light source used for probing the intervening gas.

This gas can either be in intergalactic space or being correlated with foreground galaxies. In the first case, we expect that this gas is metal-poor and thus consists mainly of hydrogen and helium. Furthermore, in order to cause absorption, the intergalactic medium must not be fully ionized, but needs to contain a fraction of neutral hydrogen. Gas located closer to galaxies may be expected to also contain appreciable amounts of metals which can give rise to absorption lines.

The identification of a spectral line with a specific line transition and a corresponding redshift is, in general, possible only if at least two lines occur at the same redshift. For this reason, doublet transitions are particularly valuable, such as those of MgII ($\lambda = 2795$ Å and $\lambda = 2802$ Å), and CIV ($\lambda = 1548$ Å and $\lambda = 1551$ Å). The spectrum of virtually any QSO at high redshift z_{em} shows narrow absorption lines of CIV and MgII at redshifts $z_{\mathrm{abs}} < z_{\mathrm{em}}$. If the spectral coverage extends to shorter wavelengths than the observed Lyα emission line of the QSO, numerous narrow absorption lines show up at $\lambda_{\mathrm{obs}} \lesssim \lambda_{\mathrm{obs}}(\mathrm{Ly}\alpha) = (1 + z_{\mathrm{em}})1216$ Å.

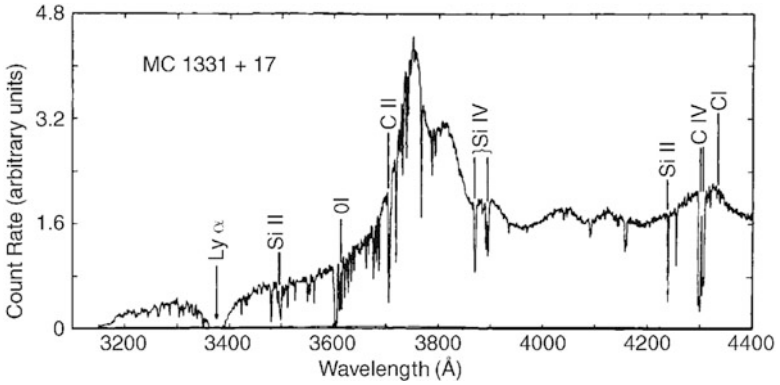


Fig. 5.54 Spectrum of the QSO 1331 + 17 at $z_{em} = 2.081$ observed by the Multi-Mirror Telescope in Arizona. In the spectrum, a whole series of absorption lines can be seen which have all been identified with gas at $z_{abs} = 1.776$. The corresponding Lyα line at $\lambda \approx 3400$ Å is very broad; it belongs to the damped Lyα lines. Source: F.H. Chaffee et al. 1988, *Molecules at early epochs. III - The Lyman-alpha disk system toward 1331 + 170*, ApJ 335, 584, p. 586, Fig. 1. ©AAS. Reproduced with permission

This set of absorption lines is denoted as the *Lyman-α forest*. In about 15 % of all QSOs, very broad absorption lines are found, the width of which may even considerably exceed that of the broad emission lines.

Classification of QSO absorption lines. The different absorption lines in QSOs are distinguished by classes according to their wavelength and width.

- *Metal systems:* In general these are narrow absorption lines, of which MgII and CIV most frequently occur (and which are the easiest to identify). However, in addition, a number of lines of other elements exist (Fig. 5.54). The redshift of these absorption lines is $0 < z_{abs} < z_{em}$; therefore they are caused by intervening matter along the line-of-sight and are not associated with the QSO. Normally a metal system consists of many different lines of different ions, all at the same redshift. From the line strength, the column density of the absorbing ions can be derived. For an assumed chemical composition and degree of ionization of the gas, the corresponding column density of hydrogen can then be determined. Estimates for such metal systems yield typical values of $10^{17}\,\mathrm{cm}^{-2} \lesssim N_H \lesssim 10^{21}\,\mathrm{cm}^{-2}$, where the lower limit depends on the sensitivity of the spectral observation.
- *Associated metal systems:* These systems have characteristics very similar to those of the aforementioned intervening metal systems, but their redshift is $z_{abs} \sim z_{em}$. Since such systems are over-abundant compared to a statistical z-distribution of the metal systems, these systems are interpreted as being related to the QSO itself. Thus the absorber is physically associated with the QSO and may be due, for example, to absorption in the QSO host galaxy or in a companion galaxy.
- *Lyα forest:* The large set of lines at $\lambda < (1 + z_{em})\,1216$ Å, as shown in Fig. 5.55, is interpreted to be Lyα absorption by hydrogen along the line-of-sight to the QSO. The statistical properties of these lines are essentially the same for all QSOs and seem to depend only on the redshift of the Lyα lines, but not on z_{em}. This interpretation is confirmed by the fact that for nearly any line in the Lyα forest, the corresponding Lyβ line is found if the quality and the wavelength range of the observed spectra permit this. The Lyα forest is further subdivided, according to the strength of the absorption, into narrow lines, Lyman-limit systems, and damped Lyα systems. Narrow Lyα lines are caused by absorbing gas of neutral hydrogen column densities of $N_H \lesssim 10^{17}\,\mathrm{cm}^{-2}$. Lyman-limit systems derive their name from the fact that at column densities of $N_H \gtrsim 10^{17}\,\mathrm{cm}^{-2}$, neutral hydrogen almost totally absorbs all radiation with $\lambda \lesssim 912$ Å (in the rest-frame of the gas), i.e., those photons which can ionize hydrogen (Fig. 5.56). If such a system is located at z_{limit} in the spectrum of a QSO, the observed spectrum at $\lambda < (1 + z_{limit})\,912$ Å is almost completely suppressed. Damped Lyα systems occur if the column density of neutral hydrogen is $N_H \gtrsim 2 \times 10^{20}\,\mathrm{cm}^{-2}$. In this case, the absorption line becomes very broad due to the extended damping wings of the Voigt profile.[16]

[16]The Voigt profile $\phi(\nu)$ of a line, which specifies the spectral energy distribution of the photons around the central frequency ν_0 of the line, is the convolution of the intrinsic line profile, described by a Lorentz profile,

$$\phi_L(\nu) = \frac{\Gamma/4\pi^2}{(\nu - \nu_0)^2 + (\Gamma/4\pi)^2}\,,$$

and the Maxwellian velocity distribution of atoms in a thermal gas of temperature T. From this, the Voigt profile follows,

$$\phi(\nu) = \frac{\Gamma}{4\pi^2} \int_{-\infty}^{\infty} \mathrm{d}v\, \frac{\sqrt{m/2\pi k_B T}\, \exp\left(-mv^2/2k_B T\right)}{(\nu - \nu_0 - \nu_0 v/c)^2 + (\Gamma/4\pi)^2}\,, \qquad (5.46)$$

where the integral extends over the velocity component along the line-of-sight. In these equations, Γ is the intrinsic line width which results from the natural line width (related to the lifetime of the atomic states)

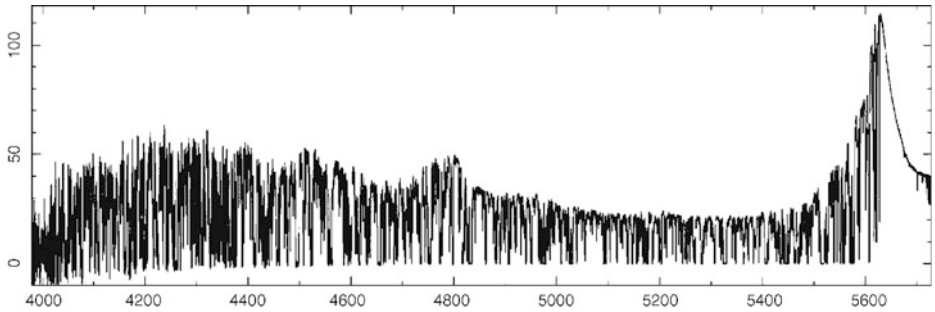


Fig. 5.55 Keck spectrum of the Lyman-α forest towards QSO $1422+231$, a QSO at $z = 3.62$. As an aside, this is a quadruply-imaged lensed QSO; it is strongly magnified by the gravitational lensing effect, so that this source is one of the brightest high-redshift QSOs—which eases obtaining high-quality spectra. The wavelength resolution is about 7 km/s. On the blue side of the Lyα emission line, a large variety of narrow absorption lines of neutral hydrogen in the intergalactic medium is visible. The statistical analysis of these lines provides information on the gas distribution in the Universe (see Sect. 8.5). Source: M. Rauch 1998, *The Lyman Alpha Forest in the Spectra of QSOs*, ARA&A 36, 267, Fig. 1, p. 268. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 36 ©1998 by Annual Reviews www.annualreviews.org

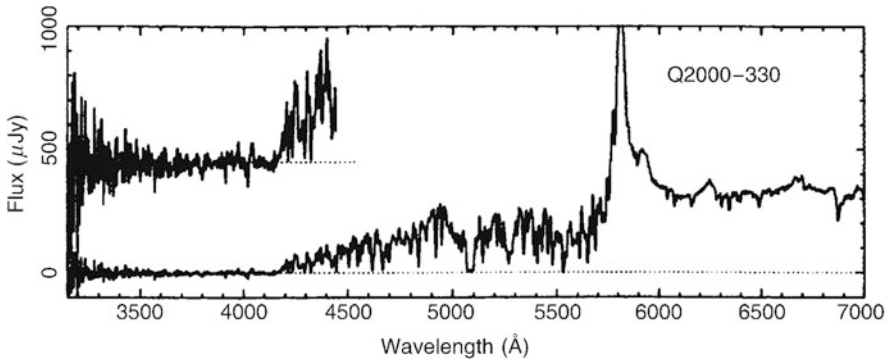


Fig. 5.56 A Lyman-limit system along the line-of-sight towards the QSO 2000−330 is absorbing virtually all radiation at wavelengths $\lambda \leq 912$ Å in the rest-frame of the absorber, here redshifted to about 4150 Å. Source: W.L.W. Sargent et al. 1989, *A survey of Lyman-limit absorption in the spectra of 59 high-redshift QSOs*, ApJS 69, 703, p. 706, Fig. 1. ©AAS. Reproduced with permission

- *Broad absorption lines:* For about 15 % of the QSOs, very broad absorption lines are found in the spectrum at redshifts slightly below $z_{\rm em}$ (Fig. 5.57). The lines show a profile which is typical for sources with outflowing material, as seen, for instance, in stars with stellar winds. However, in contrast to the latter, the Doppler width of the lines in the *broad absorption line* (BAL) QSOs is a significant fraction of the speed of light.

Interpretation. The metal systems with a redshift significantly smaller than $z_{\rm em}$ originate either in overdense regions in intergalactic space or they are associated with galaxies (or more specifically, galaxy halos) located along the line-of-sight. In fact, MgII systems always seem to be correlated with a galaxy at the same redshift as the absorbing gas. From the statistics of the angular separations of these associated galaxies to the QSO sight-line and from their redshifts, we obtain a characteristic extent of the gaseous halos of such galaxies of $\sim 25h^{-1}$ kpc. For CIV systems, the extent seems to be even larger, $\sim 40h^{-1}$ kpc.

The Lyα forest is caused by the diffuse intergalactic distribution of gas. In Sect. 8.5, we will discuss models of the Lyα forest and its relevance for cosmology more thoroughly (see also Fig. 5.58).

Broad absorption lines originate from material in the AGN itself, as follows immediately from their redshift and their enormous width. Since the redshift of the broad absorption lines is slightly lower than that of the corresponding emission lines, the absorbing gas must be moving towards us. The idea is that this is material flowing out at a very high velocity. BAL-QSOs (broad absorption line QSOs) are virtually always radio-quiet. The role of BAL-QSOs in the AGN family is unclear. A plausible interpretation is that the BAL property also depends on the orientation of the QSO.

and pressure broadening. m is the mass of the atom, which defines, together with the temperature T of the gas, the Maxwellian velocity distribution. If the natural line width is small compared to the thermal width, the Doppler profile dominates in the center of the line, that is for frequencies close to ν_0. The line profile is then well approximated by a Gaussian. In the wings of the line, the Lorentz profile dominates. For the wings of the line, where $\phi(\nu)$ is small, to become observable the optical depth needs to be high. This is the case in damped Lyα systems.

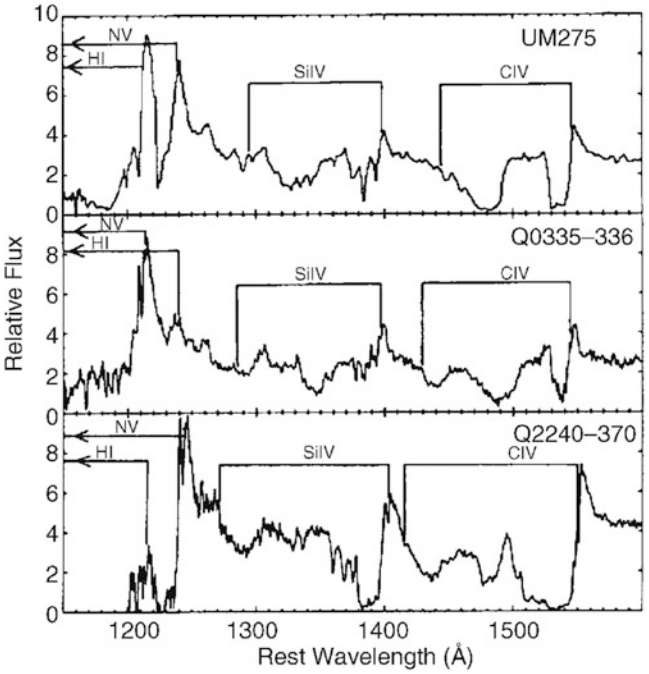


Fig. 5.57 Spectra of three BAL-QSOs, QSOs with broad absorption lines. On the blue side of every strong emission line very broad absorption is visible, such as can be caused by outflowing material. The wavelength range over which the absorption by a given line occurs is indicated by the *bars*. Such line shapes, with much lower width (of course) are also found in the spectra of stars with strong stellar winds. Source: D.A. Turnshek 1988, *BAL QSOs - Observations, models and implications for narrow absorption line systems*, in: QSO absorption lines: Probing the universe; Proceedings of the QSO Absorption Line Meeting, Baltimore, MD, Cambridge University Press, 1988, p. 17

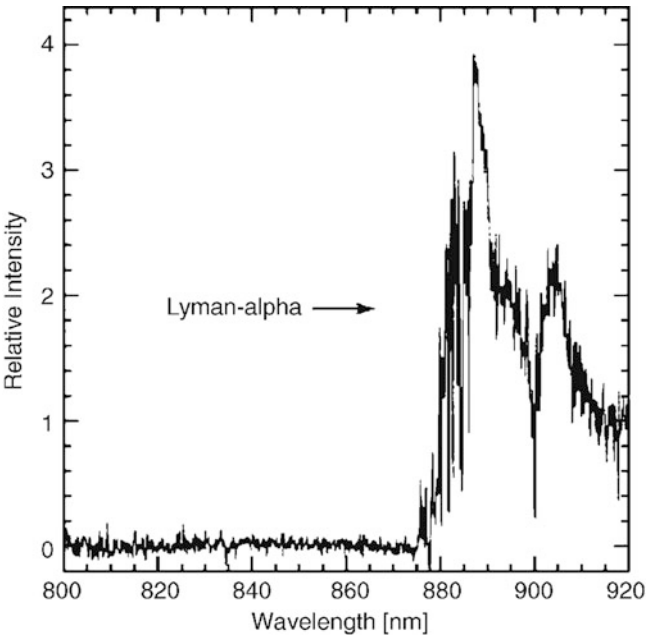


Fig. 5.58 A VLT spectrum of the QSO SDSS 1030 + 0524 at $z = 6.28$, currently one of the highest known QSO redshifts. The blue side of the Lyα emission line and the adjacent continuum are almost completely devoured by the dense Lyα forest. Credit: Laura Pentericci and Hans-Walter Rix; Max-Planck Institut für Astronomie, ESO

In this case, any QSO would be a BAL if observed from the direction into which the absorbing material streams out.

Discussion. Most absorption lines in QSO spectra are not physically related to the AGN phenomenon. Rather, they provide us with an opportunity to probe the matter along the line-of-sight to the QSO. The Lyα forest will be discussed in relation to this aspect in Sect. 8.5. Furthermore, absorption line spectroscopy of QSOs carried out with UV satellites has proven the existence of very hot gas in the halo of our Milky Way. Such UV spectroscopy provides one of the very few opportunities to analyze the intergalactic medium if its temperature is of the order of $\sim 10^6$ K—gas at this temperature is very difficult to detect since it emits in the extreme UV which is unobservable from our location inside the Milky Way, and since almost all atoms are fully ionized and therefore cause no absorption. Only absorption lines from very highly ionized metals (such as the five times ionized oxygen) can still be observed. Since the majority of the baryons should be found in this hot gas phase today, this test is of great interest for cosmology.

5.8 Problems

5.1. The spectral index of synchrotron radiation. Assume that a synchrotron source contains a population of relativistic electrons, with a power-law energy distribution, $N(E)\,\mathrm{d}E = A\,E^{-s}\,\mathrm{d}E$, where A is a constant of proportionality. The synchrotron emissivity of a single electron is a function $f(\nu/\nu_\mathrm{c})$ that depends only on the ratio of emitted frequency and the frequency ν_c given by (5.3).

1. Check that this last statement is compatible with the emitted power (5.5).
2. Calculate the shape of the synchrotron spectrum from the electron population.

5.2. Energy loss of electrons. The energy loss of an electron due to synchrotron radiation is given by (5.9).

1. Show that the energy loss due to synchrotron emission can be written as

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{4}{3}\sigma_\mathrm{T} c \gamma^2 U_B \ ,$$

   where σ_T is the Thomson cross section, and $U_\mathrm{B} = B^2/(8\pi)$ is the energy density in the magnetic field.
2. If a low-energy photon with frequency ν is Compton scattered by a relativistic electron of Lorentz factor γ, its frequency after scattering is on average $\nu' = (4/3)\gamma^2\nu$.

Show that the energy loss of an electron due to inverse Compton scattering is

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{4}{3}\sigma_{\mathrm{T}} c \gamma^2 U_\gamma \ ,$$

where U_γ is the energy density in the radiation field. Hence, the relative efficiency of relativistic electron cooling due to synchrotron emission and Compton scattering depends on the ratio of magnetic to photon energy density.

5.3. Spectrum of an optically thick accretion disk. Consider an optically thick accretion disk with a temperature profile given by (5.13). Neglecting any boundary effect (i.e., the fact that a real accretion disk extends only over a finite range in radii), show that the emitted spectrum is a power law, $L_\nu \propto \nu^{1/3}$. Comment: The true spectrum of an accretion disk deviates from this simple power law, mainly due to the existence of an inner boundary.

5.4. Mass growth of a black hole. Suppose that the black hole at some initial time $t = 0$ has mass $M_\bullet(0)$, and then accretes at constant efficiency ϵ as fixed Eddington ratio L/L_{edd}.

1. Show that its mass after some time t has grown to

$$M_\bullet(t) = M_\bullet(0) \exp\left(\frac{1-\epsilon}{\epsilon}\frac{L}{L_{\mathrm{edd}}}\frac{t}{t_{\mathrm{gr}}}\right) \ , \qquad (5.47)$$

where $t_{\mathrm{gr}} = M_\bullet c^2/L_{\mathrm{edd}} \approx 5\times10^8$ yr, independent of $M_\bullet$.

2. Suppose the initial (seed) mass is $M_\bullet(0) = 10M_\odot$. If the efficiency is $\epsilon = 0.1$, and the accretion occurs with Eddington luminosity, what is the black hole mass after 10^9 yr?

5.5. Properties of the BLR. Assume that the BLR is a spherical shell with characteristic radius r and thickness $\delta r \approx r$. Furthermore, assume that it consists of N_{c} clouds of radius r_{c} and electron number density n_{e}.

1. What is the covering factor of the BLR clouds as seen from the continuum source, i.e., which fraction of lines-of-sight from the center of the BLR intersect a cloud, in terms of the model parameters?
2. Calculate the filling factor, i.e., the volume fraction of the BLR that is filled with clouds.
3. Assume that the covering factor is 0.1, and that the filling factor is 10^{-6}. For a BLR radius of $r = 10^{16}$ cm and $n_{\mathrm{e}} = 10^{10}\,\mathrm{cm}^{-3}$, determine r_{c} and N_c. What is the total mass of the gas in the clouds in the BLR?

Comment: Given the uncertainty with which quantities like the covering factor can be determined, it is legitimate to neglect factors of order unity in the calculation.

5.6. Relative luminosity of AGN and host galaxy. Assume that the SMBH mass in an AGN host galaxy is 10^{-3} times the stellar mass of its spheroidal component, as found for nearby galaxies. Furthermore, assume that the spheroidal component contains a fraction f_{sph} of the total stellar mass of the galaxy. Let the AGN radiate with an Eddington ratio L/L_{edd}, and assume that 10% of the radiation comes out in the optical waveband.

1. Calculate the ratio of the optical AGN luminosity and the stellar luminosity, as a function of Eddington ratio, mass-to-light ratio of the stellar population, and the spheroidal fraction f_{sph}.
2. Discuss your result in terms of the detectability of the AGN, assuming $L/L_{\mathrm{edd}} \sim 0.1$.

5.7. Tidal disruption of a star. Show that a star of mass M_* and radius R_* can be disrupted by a SMBH only if the black hole mass $M_\bullet$ is not too large. Calculate this limiting mass for a Solar-like star.

Clusters and groups of galaxies

6

Galaxies are not uniformly distributed in space, but instead show a tendency to gather together in *galaxy groups* and *clusters of galaxies*. This effect can be clearly recognized in the projection of bright galaxies on the sky (see Figs. 6.1 and 6.2). The Milky Way itself is a member of a group, called the Local Group (Sect. 6.1), which implies that we are living in a locally overdense region of the Universe.

The transition between groups and clusters of galaxies is smooth. Historically, the distinction was made on the basis of the number of their member galaxies. Roughly speaking, an accumulation of galaxies is called a group if it consists of $N \lesssim 50$ members within a sphere of diameter $D \lesssim 1.5h^{-1}$ Mpc. Clusters have $N \gtrsim 50$ members and diameters $D \gtrsim 1.5h^{-1}$ Mpc. A formal definition of a cluster is presented further below. An example of a group and a cluster of galaxies is displayed in Fig. 6.3.

Clusters of galaxies are very massive: typical values are $M \gtrsim 3 \times 10^{14} M_\odot$ for massive clusters, whereas for groups $M \sim 3 \times 10^{13} M_\odot$ is characteristic, with the total mass range of groups and clusters extending over $10^{12} M_\odot \lesssim M \lesssim$ few$\times 10^{15} M_\odot$.

Originally, clusters of galaxies were characterized as such by the observed spatial concentration of galaxies. Today we know that, although the galaxies determine the optical appearance of a cluster, the mass contained in galaxies contributes only a small fraction to the total mass of a cluster. Through advances in X-ray astronomy, it was discovered that galaxy clusters are intense sources of X-ray radiation which is emitted by a hot gas ($T \sim 3 \times 10^7$ K) located between the galaxies. This intergalactic gas (*intracluster medium*, ICM) contains more baryons than the stars seen in the member galaxies. From the dynamics of galaxies, from the properties of the intracluster gas, and from the gravitational lens effect we deduce the existence of dark matter in galaxy clusters, dominating the cluster mass like it does for galaxies.

Clusters of galaxies play a very important role in observational cosmology. They are the most massive bound and relaxed (i.e., in a state of approximate dynamical equilibrium) cosmic structures, and therefore mark the most prominent density peaks of the large-scale structure in the Universe. For that reason, their cosmological evolution is directly related to the growth of cosmic structures, as will be discussed in Chaps. 7 and 8. Due to their high galaxy number density, clusters and groups are also ideal laboratories for studying interactions between galaxies and their effect on the galaxy population. For instance, the fact that elliptical galaxies are preferentially found in clusters indicates the

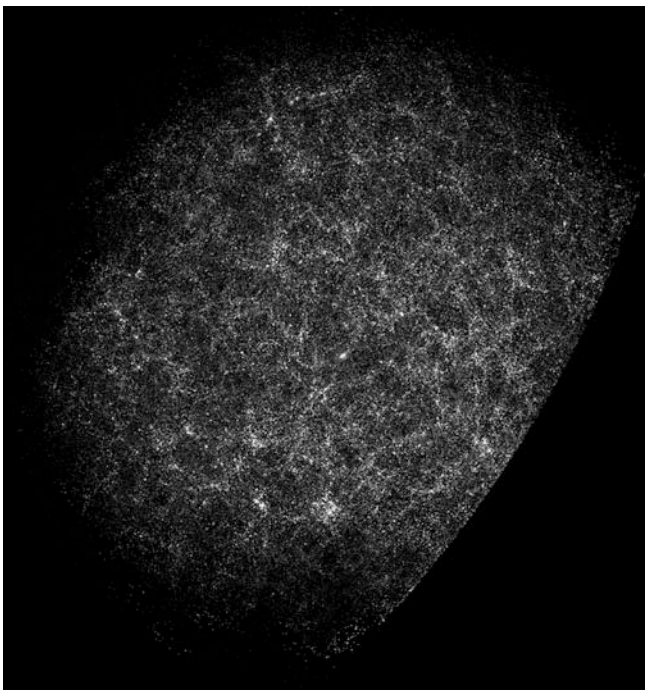


Fig. 6.1 The distribution of galaxies in the Northern sky, as compiled in the Lick catalog. This catalog contains the galaxy number counts for 'pixels' of $10' \times 10'$ each. It is clearly seen that the distribution of galaxies on the sphere is far from being homogeneous. Instead it is distinctly structured. For an all-sky map of bright galaxies, as observed at near-IR wavelengths, see Fig. 1.52. Source: Webpage E.J. Groth, Princeton University; adapted from M. Seldner et al. 1977, *New reduction of the Lick catalog of galaxies*, AJ 82, 249

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_6,

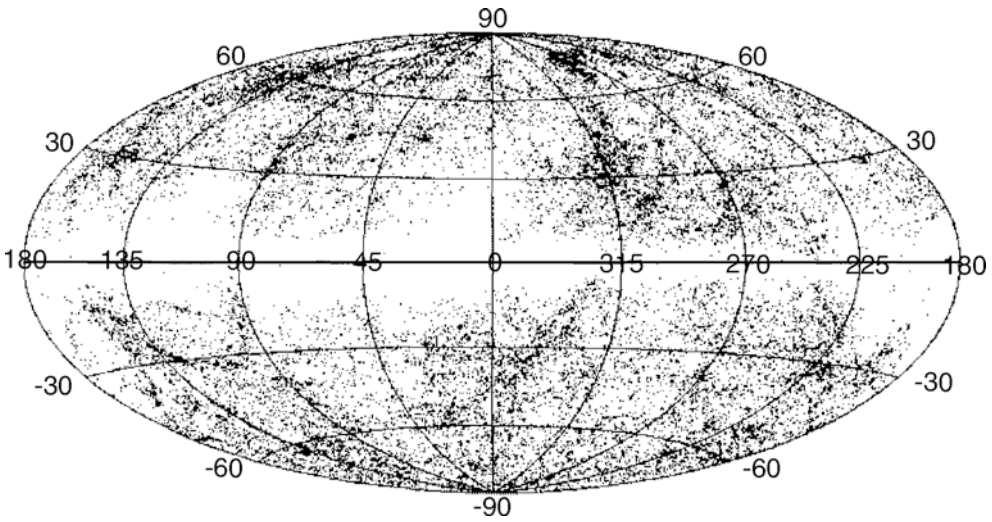


Fig. 6.2 The distribution of all galaxies brighter than $B < 14.5$ on the sphere, plotted in Galactic coordinates. The Zone of Avoidance is clearly seen as the region near the Galactic plane. Source: N.A. Sharp 1986, *The whole-sky distribution of galaxies*, PASP 98, 740, p. 753, Fig. 14

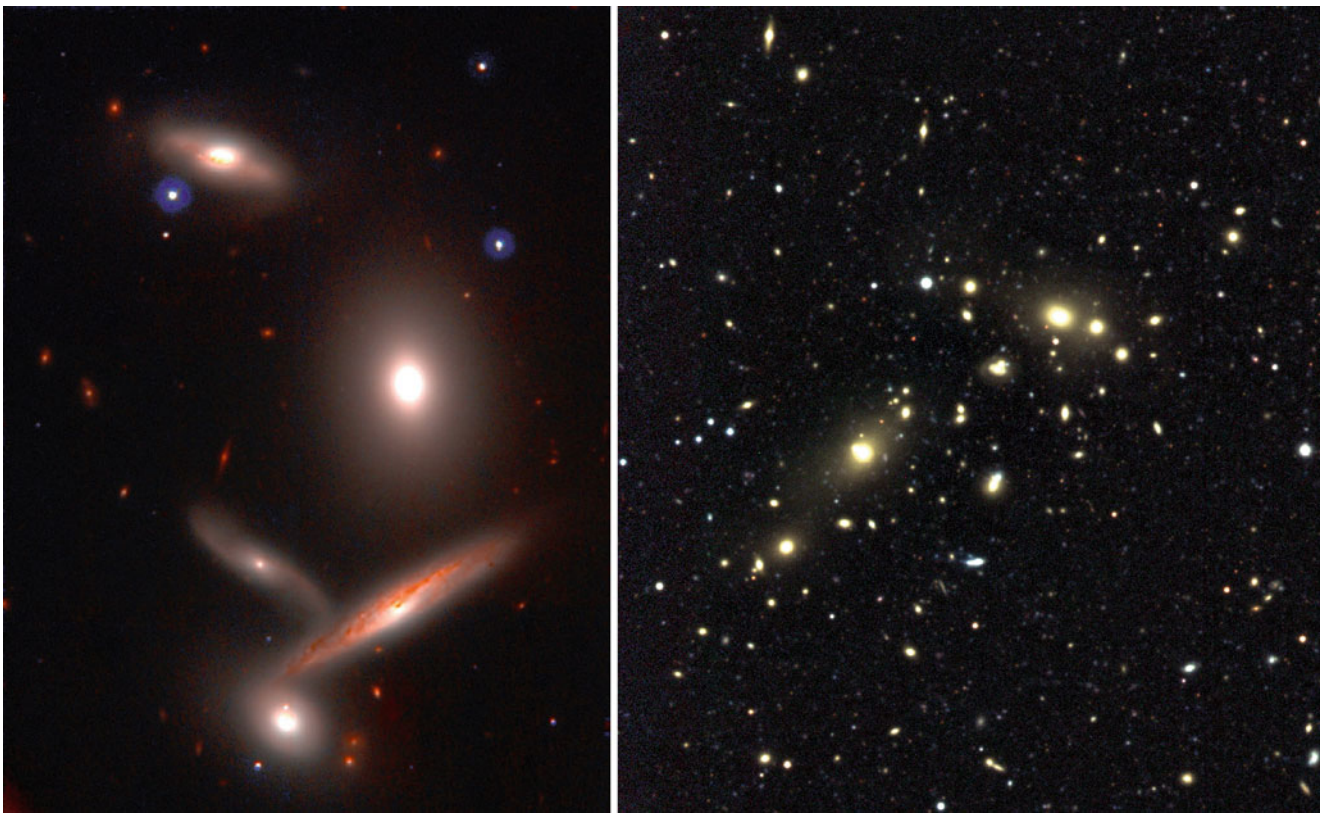


Fig. 6.3 The *left panel* shows HCG40, a compact group of galaxies, observed with the Subaru telescope on Mauna-Kea. The *right panel* displays the cluster of galaxies Cl 0053−37, observed with the WFI at the ESO/MPG 2.2-m telescope. Credits: *Left*: Copyright @ Subaru Telescope, NAOJ. All rights reserved. *Right*: M. Schirmer, European Southern Observatory

impact of the local galaxy density on the morphology and evolution of galaxies.

Outline of this chapter. We will start be discussing the nearest association of galaxies, namely the Local Group, of which the Milky Way is a member. In Sect. 6.2, we describe the identification of galaxy clusters with optical methods, and some of the resulting cluster and group catalogs. The spatial distribution of galaxies in clusters and their dynamics will be studied in Sect. 6.3. We will show that the relative motion of galaxies in clusters implies a much higher cluster mass than can be accounted for by the stars seen in the member galaxies. Whereas not all stars are bound in individual cluster galaxies, but some are distributed throughout the cluster,

forming the intracluster light component, this additional stellar component only constitutes a $\sim 20\,\%$ contribution to the overall stellar mass budget.

The space between the galaxies is filled by a hot gas, detected by its X-ray emission and its impact on the spectrum of the observed cosmic microwave background radiation seen in the direction of clusters. We study this intracluster medium in Sect. 6.4; in particular we show how the properties of the gas can be used for mass determination of clusters. These mass estimates are in good agreement with those obtained from the dynamics of galaxies, reinforcing the conclusion the clusters contain more mass than directly observed, even if the mass of the hot gas is taken into account. The central galaxies of many clusters contain an AGN, whose energy output has a distinct impact on the properties of the hot gas.

In Sect. 6.5 we show that there exist tight relations between the temperature of the hot intracluster gas, its X-ray and optical/near-IR luminosities, the galaxy velocity dispersion and the cluster mass. These scaling relations, analogous to the scaling relations of galaxies, indicate that clusters of the same mass have rather similar properties.

Clusters of galaxies can act as gravitational lenses, giving rise to spectacular imaging phenomena. Those will be discussed in Sect. 6.6, together with a method which allows one to obtain maps of the total matter distribution in clusters. In particular, gravitational lensing yields a third, fully independent method for determining cluster masses. We will find that more than $80\,\%$ of the cluster mass is made of dark matter, only $\sim 3\,\%$ of stars, and some $15\,\%$ of the baryons in the intracluster medium.

The dense environment of groups and clusters may affect the evolution of their member galaxies; we shall therefore discuss the galaxy population of clusters in Sect. 6.7; more generally, we will describe the properties of galaxies in relation to the density of their environment. Finally, we discuss in Sect. 6.8 some evolutionary aspects of the cluster population.

6.1 The Local Group

The Milky Way is a member of the *Local Group*. Within a distance of ~ 1 Mpc around our Galaxy, about 35 galaxies were known at the turn of the Millennium; these 'classical' Local Group members are listed in Table 6.1, and a sketch of their spatial distribution is given in Fig. 6.4. With the Sloan Digital Sky Survey (SDSS; see Sect. 1.4), about 20 additional very faint galaxies in the Local Group have been found. Most of them cannot be detected solely as overdensity of stars on the sky, because their density contrast is too low. However, by filtering the star catalog according to stellar colors and magnitudes, which together allow for the selection of stars from an old population at similar distances, spatial overdensities can be identified. We will return to them in Sect. 7.8.

6.1.1 Phenomenology

The Milky Way (MW), M31 (Andromeda; see Fig. 6.5), and M33 (Fig. 6.6) are the three spiral galaxies in the Local Group, and they are also its most luminous members. The Andromeda galaxy is located at a distance of 770 kpc from us, M33 at about 850 kpc. The Local Group member next in luminosity is the Large Magellanic Cloud (LMC, see Fig. 6.7), which is orbiting around the Milky Way, together with the Small Magellanic Cloud (SMC), at a distance of $\sim$ 50 kpc ($\sim$ 60 kpc, respectively, for the SMC). Both are satellite galaxies of the Milky Way and belong to the class of irregular galaxies (like about 11 other Local Group members). The other members of the Local Group are dwarf galaxies, which are very small and faint (see Fig. 6.8 for three examples). Because of their low luminosity and their low surface brightness, many of the known members of the Local Group were detected only fairly recently. For example, the Antlia galaxy, a dwarf spheroidal galaxy, was found in 1997. Its luminosity is about 10^4 times smaller than that of the Milky Way.

Many of the dwarf galaxies are grouped around the Galaxy or around M31; these are known as *satellite galaxies*. Distributed around the Milky Way are the LMC, the SMC, and about 20 dwarf galaxies, several of them in the so-called *Magellanic Stream* (see Fig. 2.19), a long, extended band of neutral hydrogen which was stripped from the Magellanic Clouds about 2×10^8 yr ago by tidal interactions with the Milky Way. The Magellanic Stream contains about $3 \times 10^8 M_\odot$ of neutral hydrogen.

The spatial distribution of satellite galaxies around the Milky Way shows a pronounced peculiarity, in that the 11 closest satellites form a highly flattened system. These satellites appear to lie essentially in a plane which is oriented perpendicular to the Galactic plane, concentrated along the minor axis of the disk. The satellites around M31 also seem to be distributed in an anisotropic way around their host. In fact, satellite galaxies around spirals seem to be preferentially located near the short axes of the projected light distribution, which has been termed the Holmberg effect, although the statistical significance of this alignment has been questioned, in particular in recent years. We will come back to this issue in Sect. 7.8.

In fact, the Local Group is not a group of galaxies in the sense of this chapter; its spatial extent is too large for a group of this mass, and it is not dynamically relaxed. The bimodal distribution of galaxies in the Local Group seen in Fig. 6.4 instead suggests that two small galaxy groups—one centered

Table 6.1 'Classical' members of the Local Group

Galaxy	Type	M_B	RA/dec.	ℓ,b	D(kpc)	v_r (km/s)
Milky Way	Sbc I-II	−20.0	1830 − 30	0, 0	8	0
LMC	Ir III-IV	−18.5	0524 − 60	280, −33	50	270
SMC	Ir IV-V	−17.1	0051 − 73	303, −44	63	163
Sgr I	dSph?		1856 − 30	6, −14	20	140
Fornax	dE0	−12.0	0237 − 34	237,−65	138	55
Sculptor Dwarf	dSph	−9.8	0057 − 33	286, −84	88	110
Leo I	dSph	−11.9	1005 + 12	226, +49	790	168
Leo II	dSph	−10.1	1110 + 22	220, +67	205	90
Ursa Minor	dSph	−8.9	1508 + 67	105, +45	69	−209
Draco	dSph	−9.4	1719 + 58	86, +35	79	−281
Carina	dSph	−9.4	0640 − 50	260, −22	94	229
Sextans	dSph	−9.5	1010 − 01	243, +42	86	230
M31	Sb I-II	−21.2	0040 + 41	121, −22	770	−297
M32 = NGC 221	dE2	−16.5	0039 + 40	121, −22	730	−200
M110 = NGC 205	dE5p	−16.4	0037 + 41	121, −21	730	−239
NGC 185	dE3p	−15.6	0036 + 48	121, −14	620	−202
NGC 147	dE5	−15.1	0030 + 48	120, −14	755	−193
And I	dSph	−11.8	0043 + 37	122, −25	790	–
And II	dSph	−11.8	0113 + 33	129, −29	680	–
And III	dSph	−10.2	0032 + 36	119, −26	760	–
Cas = And VII	dSph		2326 + 50	109, −09	690	–
Peg = DDO 216	dIr/dSph	−12.9	2328 + 14	94, −43	760	–
Peg II = And VI	dSph	−11.3	2351 + 24	106, −36	775	–
LGS 3	dIr/dSph	−9.8	0101 + 21	126, −41	620	−277
M33	Sc II-III	−18.9	0131 + 30	134, −31	850	−179
NGC 6822	dIr IV-V	−16.0	1942 − 15	025, −18	500	−57
IC 1613	dIr V	−15.3	0102 + 01	130, −60	715	−234
Sagittarius	dIr V	−12.0	1927 − 17	21, +16	1060	−79
WLM	dIr IV-V	−14.4	2359 − 15	76, −74	945	−116
IC 10	dIr IV	−16.0	0017 + 59	119, −03	660	−344
DDO 210, Aqr	dIr/dSph	−10.9	2044 − 13	34, −31	950	−137
Phoenix Dwarf	dIr/dSph	−9.8	0149 − 44	272, 68	405	56
Tucana	dSph	−9.6	2241 − 64	323, −48	870	–
Leo A = DDO 69	dIr V	−11.7	0959 + 30	196, 52	800	–
Cetus Dwarf	dSph	−10.1	0026 − 11	101, −72	775	–

Listed are the name of the galaxy, its morphological type, the absolute B-band magnitude, its position on the sphere in both right ascension/declination and in Galactic coordinates, its distance from the Sun, and its radial velocity. A sketch of the spatial configuration is displayed in Fig. 6.4

on M31, the other one centered on the Milky Way—are in a process of merging.

6.1.2 Mass estimate

We will present a simple mass estimate of the Local Group, from which we will find that it is considerably more massive than one would conclude from the observed luminosity of the associated galaxies.

M31 is one of the very few galaxies with a blueshifted spectrum. Hence, Andromeda and the Milky Way are approaching each other, at a relative velocity of $v \approx 120\,\mathrm{km/s}$. This value results from the velocity of M31 relative to the Sun of $v \approx 300\,\mathrm{km/s}$, and from the motion of the Sun around the Galactic center. Together with the distance to M31 of $D \sim 770\,\mathrm{kpc}$, we conclude that both galaxies will collide on a time-scale of $\sim 6 \times 10^9\,\mathrm{yr}$, if we disregard the transverse component of the relative velocity. From measurements of the proper motion of M31, one finds that its transverse velocity is small—thus a collision with the Milky Way will almost certainly occur.

The luminosity of the Local Group is dominated by the Milky Way and by M31, which together produce about 90 % of the total luminosity. If the mass density follows the light distribution, the dynamics of the Local Group should also

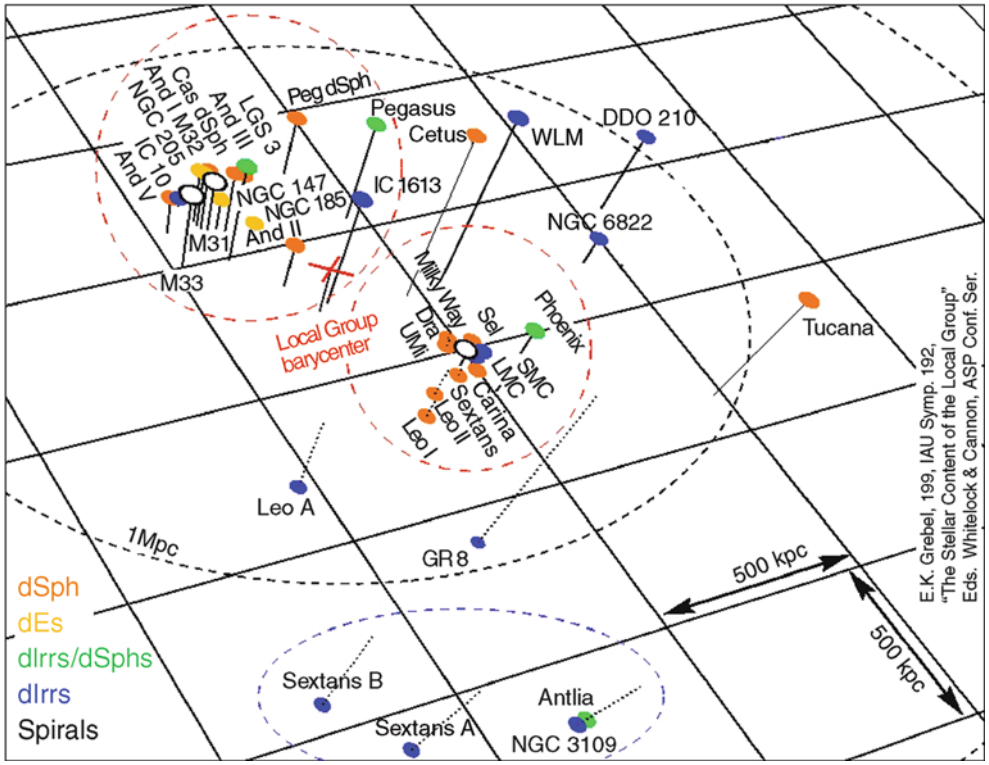


Fig. 6.4 Schematic distribution of galaxies in the Local Group, with the Milky Way at the center of the figure Shown are only the 'classical' Local Group members which were known before 2000; most of the newly found galaxies in the Local Group are ultra-faint dwarfs. Credit: E. Grebel, Astronomical Institute, University of Basel, Switzerland

be dominated by these two galaxies. Therefore, one can try to estimate the mass of the two galaxies from their relative motion, and with this also the mass of the Local Group.

In the early phases of the Universe, the Galaxy and M31 were close together and both took part in the Hubble expansion. By their mutual gravitational attraction, their relative motion was decelerated until it came to a halt—at a time $t_{\max}$ at which the two galaxies had their maximum separation $r_{\max}$ from each other. From this time on, they have been moving towards each other. The relative velocity $v(t)$ and the separation $r(t)$ follow from the conservation of energy,

$$\frac{v^2}{2} = \frac{GM}{r} - C \ , \tag{6.1}$$

where M is the sum of the masses of the Milky Way and M31, and C is an integration constant, related to the total energy of the M31/MW-system. This constant can be determined by considering (6.1) at the time of maximum separation, when $r = r_{\max}$ and $v = 0$. With this,

$$C = \frac{GM}{r_{\max}}$$

follows immediately. Since $v = \mathrm{d}r/\mathrm{d}t$, (6.1) is a differential equation for $r(t)$,

$$\frac{1}{2}\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 = GM\left(\frac{1}{r} - \frac{1}{r_{\max}}\right) \ .$$

It can be solved using the initial condition $r = 0$ at $t = 0$. For our purpose, an approximate consideration is sufficient. Solving the equation for $\mathrm{d}t$ we obtain, by integration, a relation between $r_{\max}$ and $t_{\max}$,

$$t_{\max} = \int_0^{t_{\max}} \mathrm{d}t = \int_0^{r_{\max}} \frac{\mathrm{d}r}{\sqrt{2GM}\sqrt{1/r - 1/r_{\max}}}$$
$$= \frac{\pi\, r_{\max}^{3/2}}{2\sqrt{2GM}} \ . \tag{6.2}$$

Since the differential equation is symmetric with respect to changing $v \to -v$, the collision will happen at $2t_{\max}$. Estimating the time from today to the collision, by assuming the relative velocity to be constant during this time, then yields $r(t_0)/v(t_0) = D/v = 770\,\mathrm{kpc}/(120\,\mathrm{km/s})$, and one obtains $2t_{\max} \approx t_0 + D/v$, or

$$t_{\max} \approx \frac{t_0}{2} + \frac{D}{2v} \approx 10^{10}\,\mathrm{yr} \ , \tag{6.3}$$

where $t_0 \approx 14 \times 10^9\,\mathrm{yr}$ is the current age of the Universe. Hence, together with (6.2) this yields

$$\frac{v^2}{2} = \frac{GM}{r} - \frac{GM}{r_{\max}} = \frac{GM}{r} - \frac{1}{2}\left(\frac{\pi GM}{t_{\max}}\right)^{2/3} \ . \tag{6.4}$$

Now by inserting the values $r(t_0) = D$ and $v = v(t_0)$, we obtain the mass M,

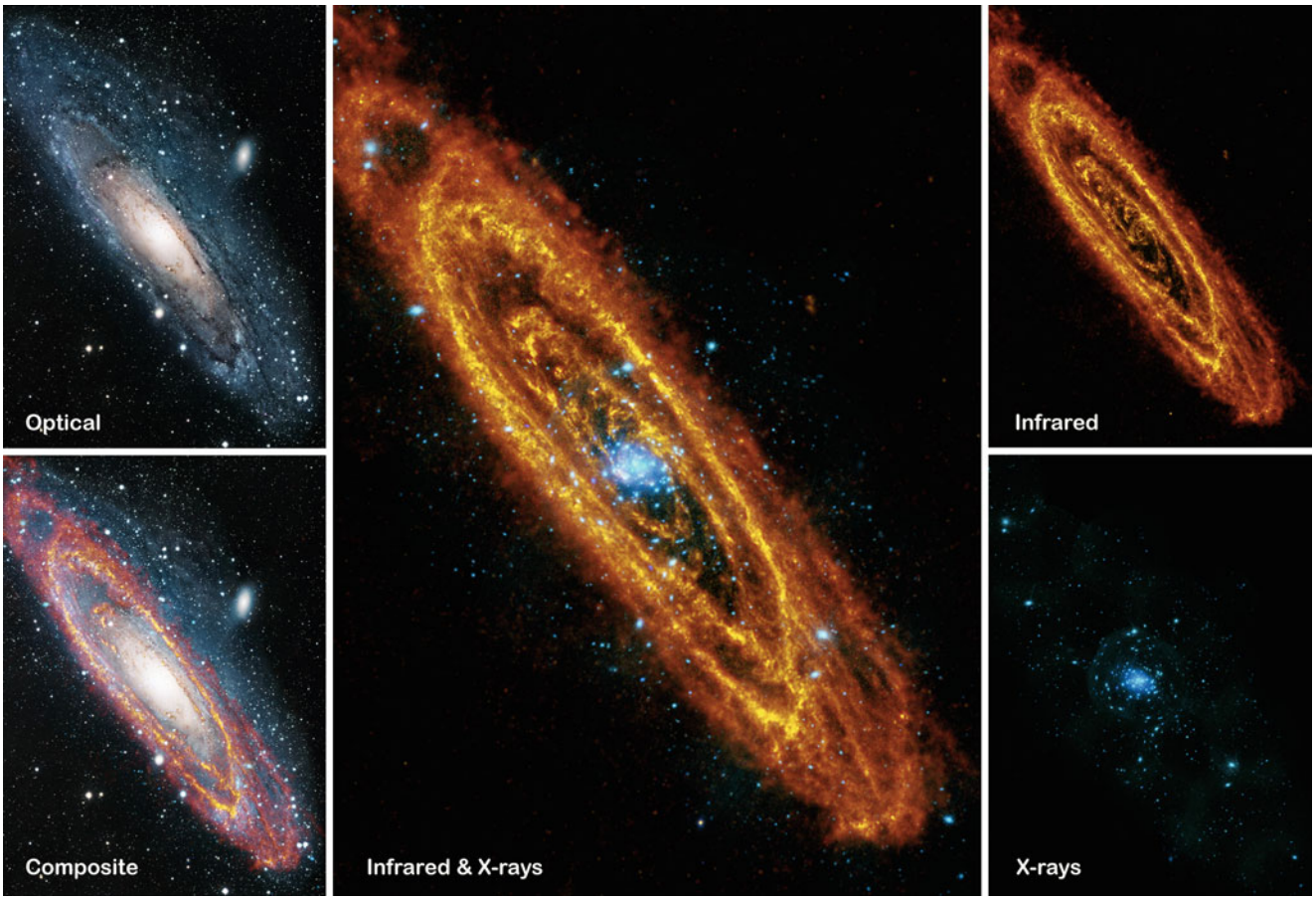


Fig. 6.5 M31, the Andromeda galaxy, seen in different wavelengths. All images show a region of $1.5° \times 2°$. Compared to the optical image (*top left*) which shows the stellar distribution of the galaxy, the far-infrared emission shown on the *top right* displays predominantly the dust component of the interstellar medium. Heated by young massive stars, the dust re-radiates the absorbed energy at long wavelengths, here shown with a 250 μm exposure taken with the Herschel Observatory. The X-ray image (*bottom right*), taken with XMM-Newton, mainly displays the distribution of X-ray binaries and supernova remnants; in particular the former are much more concentrated towards the central parts of the galaxy. The two composite images in the *center* and *bottom left* compare the distributions of the various components of M31. Credit & Copyright: infrared: ESA/Herschel/PACS/SPIRE/J. Fritz, U. Gent; X-ray: ESA/XMM-Newton/EPIC/W. Pietsch, MPE; optical: R. Gendler

$$M \sim 3 \times 10^{12} M_\odot \ . \tag{6.5}$$

This mass is much larger than the mass of the two galaxies as observed in stars and gas. The mass estimate yields a mass-to-light ratio for the Local Group of $M/L \sim 70\, M_\odot/L_\odot$, much larger than that of any known stellar population. This is therefore another indication of the presence of dark matter because we can see only about 5 % of the estimated mass in the Milky Way and Andromeda. Another mass estimate follows from the kinematics of the Magellanic Stream, which also yields $M/L \gtrsim 80 M_\odot/L_\odot$.

6.1.3 Other components of the Local Group

Tidal streams. One of the most interesting galaxies in the Local Group is the *Sagittarius dwarf galaxy* which was only discovered in 1994. Since it is located in the direction of the Galactic bulge, it is barely visible on optical images, if at all, as an overdensity of stars. Furthermore, it has a very low surface brightness. It was discovered in an analysis of stellar kinematics in the direction of the bulge, in which a coherent group of stars was found with a velocity distinctly different from that of bulge stars. In addition, the stars belonging to this overdensity have a much lower metallicity, reflected in their colors. The Sagittarius dwarf galaxy is located close to the Galactic plane, at a distance of about 16 kpc from the Galactic center and nearly in the direct extension of our line-of-sight to the GC. This proximity implies that it must be experiencing strong tidal gravitational forces on its orbit around the Milky Way; over the course of time, these will have the effect that the Sagittarius dwarf galaxy will be slowly disrupted. In fact, in recent years a relatively narrow band of stars was found around the Milky Way. These stars are located along the orbit of the Sagittarius galaxy (see Fig. 2.18). Their chemical composition supports the

Fig. 6.6 Multi-band (u,g,r) composite image of the Triangulum Galaxy (M33), a Local Group spiral with an estimated distance of $\sim$ 850 kpc. With its visible diameter of $\sim$ 15 kpc, it is the third largest galaxy of the Local Group; its stellar mass is about 1/10 that of the Milky Way. Observations of water masers in this galaxy enabled the measurement of its proper motion, indicating that it is heading towards M31. The bluish emission is due to regions of active star formation; indeed, the bright region near the top-right of the images is one of the most luminous HII regions known. This is the first image taken with the wide-field camera of the new 2-m Fraunhofer telescope on the Wendelstein Observatory and covers 30′ on a side. Credit: Wendelstein Observatory, Universitätssternwarte der Ludwig-Maximilians-Universität München

interpretation that they are stars stripped from the Sagittarius dwarf galaxy by tidal forces. In addition, globular clusters were identified which presumably once belonged to the Sagittarius dwarf galaxy, but which were also removed from it by tidal forces and are now part of the globular cluster population in the Galactic halo. Indeed, more tidal streams have been discovered recently, both in the Milky Way and in Andromeda, as well as in other neighboring galaxies.

The neighborhood of the Local Group. The Local Group is indeed a concentration of galaxies: while it contains more than 50 members within $\sim$ 1 Mpc, the next neighboring galaxies are found only in the Sculptor Group, which contains about six members[1] and is located at a distance of $D \sim 1.8$ Mpc. The next galaxy group after this is the M81-group of $\sim$ 8 galaxies at $D \sim 3.1$ Mpc, the two most prominent galaxies of which are displayed in Fig. 6.9.

[1]Of course, the numbers quoted here are those of currently known galaxies. Dwarf galaxies like Sagittarius would be very difficult to detect at the distances of these groups.

The other nearby associations of galaxies within 10 Mpc from us shall also be mentioned: the Centaurus group with 17 members and $D \sim 3.5$ Mpc, the M101-group with 5 members and $D \sim 7.7$ Mpc, the M66- and M96-group with together 10 members located at $D \sim 9.4$ Mpc, and the NGC 1023-group with 6 members at $D = 9.6$ Mpc.

Most galaxies are members of a group. Many more dwarf galaxies exist than luminous galaxies, and dwarf galaxies are located preferentially in the vicinity of larger galaxies. Some members of the Local Group are so under-luminous that they would hardly be observable outside the Local Group.

One large concentration of galaxies was already known in the eighteenth century (W. Herschel)—the *Virgo cluster*. Its galaxies extend over a region of about $10° \times 10°$ in the sky, and its distance is $D \sim 16$ Mpc. The Virgo cluster consists of about 250 large galaxies and more than 2000 smaller ones; the central galaxy of the cluster is the radio galaxy M87 (Fig. 1.11). In the classification scheme of galaxy clusters, Virgo is considered an irregular cluster. The closest regular massive galaxy cluster is the *Coma cluster* (see Fig. 1.17), at a distance of about $D \sim 90$ Mpc.

6.2 Optical cluster searches

6.2.1 The Abell catalog

George Abell compiled a catalog of galaxy clusters, published in 1958, in which he identified regions in the sky that show an overdensity of galaxies. This identification was performed by eye on photoplates from the *Palomar Observatory Sky Survey* (POSS), a photographic atlas of the Northern ($\delta > -30°$) sky (see Sect. 1.4). He omitted the Galactic disk region because the observation of galaxies is considerably more problematic there, due to extinction and the high stellar density (see also Fig. 6.2).

Abell's criteria and his catalog. The criteria Abell applied for the identification of clusters refer to an overdensity of galaxies within a specified solid angle. According to these criteria, a cluster contains $\geq$50 galaxies in a magnitude interval $m_3 \leq m \leq m_3 + 2$, where m_3 is the apparent magnitude of the third brightest galaxy in the cluster.[2] These galaxies must be located within a circle of angular radius

$$\theta_A = \frac{1'.7}{z}\,, \tag{6.6}$$

[2]The reason for choosing the third brightest galaxy is that the luminosity of the brightest galaxy may vary considerably among clusters. Even more important is the fact that there is a finite probability for the brightest galaxy in a sky region under consideration to not belong to the cluster, but to be located at some smaller distance from us.

Fig. 6.7 An image of the Large Magellanic Cloud (LMC), taken with the CTIO 4-m telescope. Credit & Copyright: AURA/NOAO/NSF

where z is the estimated redshift. The latter is estimated by the assumption that the luminosity of the tenth brightest galaxy in a cluster is the same for all clusters. A calibration of this distance estimate is performed on clusters of known redshift. θ_{A} is called the *Abell radius* of a cluster, and corresponds to a physical radius of $R_{\mathrm{A}} \approx 1.5h^{-1}\,\mathrm{Mpc}$.

The so-determined redshift should be within the range $0.02 \leq z \leq 0.2$ for the selection of Abell clusters. The lower limit is chosen such that a cluster can be found on a single POSS photoplate ($\sim 6° \times 6°$) and does not extend over several plates, which would make the search more difficult, e.g., because the photographic sensitivity may differ for individual plates. The upper redshift bound is chosen due to the sensitivity limit of the photoplates.

The Abell catalog contains 1682 clusters which all fulfill the above criteria. In addition, it lists 1030 clusters that were found in the search, but which do not fulfill all of the criteria (most of these contain between 30 and 49 galaxies). An extension of the catalog to the Southern sky was published by Abell, Corwin & Olowin in 1989. This ACO catalog contains 4076 clusters, including the members of the original catalog. Another important catalog of galaxy clusters is the Zwicky catalog (1961–1968), which contains more clusters, but which is considered less reliable, since the applied selection criteria resulted in more spurious cluster candidates than is the case for the Abell catalog.

Problems in the optical search for clusters. The selection of galaxy clusters from an overdensity of galaxies on the sphere is not without problems, in particular if these catalogs are to be used for statistical purposes. An ideal catalog ought to fulfill two criteria: first it should be complete, in the sense that all objects which fulfill the selection criteria are contained in the catalog. Second it should be pure (often also called 'reliable'), i.e., it should not contain any objects that do not belong in the catalog because they do not fulfill the criteria (so-called false positives). The Abell catalog is neither complete, nor is it pure. We will briefly discuss why completeness and reliability cannot be expected in a catalog compiled in this way.

A galaxy cluster is a three-dimensional object, whereas galaxy counts on images are necessarily based on the projection of galaxy positions onto the sky. Therefore, projection effects are inevitable. Random overdensities on the sphere caused by line-of-sight projection may easily be classified as clusters. The reverse effect is likewise possible: due to fluctuations in the number density of foreground galaxies, a cluster at high redshift may be classified as an insignificant fluctuation—and thus remain undiscovered.

Of course, not all members of a cluster classified as such are in fact galaxies in the cluster, as here projection effects also play an important role. Furthermore, the redshift estimate is relatively coarse. In the meantime, spectroscopic analyses have been performed for many of the Abell clusters, and it has been found that Abell's redshift estimates have an error of about 30 %—they are surprisingly accurate, considering the coarseness of his assumptions.

The Abell catalog is based on visual inspection of photographic plates. It is therefore partly subjective. Today, the Abell criteria can be applied to digitized images in an

Fig. 6.8 *Upper left*: NGC 6822, also known as Barnard's Galaxy, is one of the dwarf elliptical galaxies of the Local Group, located at a distance of 500 kpc from the Milky Way. This color composite image covers a region of 34′ on the side, and was taken with the WFI@ESO/MPG 2.2 m telescope on La Silla. The reddish nebulae in the image indicate regions of active star formation. *Upper right*: The Fornax dwarf spheroidal galaxy is a satellite of the Milky Way, at a distance of 140 kpc. The image size is about $17' \times 13'$, and was extracted from the Digitized Sky Survey II. *Bottom*: The Antlia dwarf galaxy lies at a distance of 1.3 Mpc, at the edge of the Local Group. This color-composite HST images covers $3'.2 \times 1'.5$. Credit: *Top left*: ESO; *Top right*: ESO/Digitized Sky Survey 2; *Bottom*: ESA/NASA

objective manner, using automated searches. From these, it was found that the results are not much different. The visual search thus was performed with great care and has to be recognized as a great accomplishment. For this reason, and in spite of the potential problems discussed above, the Abell and the ACO catalogs are still frequently used.

The clusters in the catalog are ordered by right ascension and are numbered. For example, Abell 851 is the 851st entry in the catalog, also denoted as A851. With a redshift of $z = 0.41$, A851 is the most distant Abell cluster.

Abell classes. The Abell and ACO catalogs divide clusters into so-called richness and distance classes. Table 6.2 lists the criteria for the richness classes, while Table 6.3 lists those for the distance classes, with the number of clusters in each class referring to the original Abell catalog (i.e., without the ACO extension).

There are six *richness classes*, denoted from 0 to 5, according to the number of cluster member galaxies. Richness class 0 contains between 30 and 49 members and therefore does not belong to the cluster catalog proper. One can see from Table 6.2 that the number of clusters rapidly decreases with increasing richness class, so only very few clusters exist with a very large number of cluster galaxies. As a reminder, the region of the sky from where the Abell clusters were detected is about half of the total sphere. Thus, only a few very rich clusters do indeed exist (at redshift $\lesssim 0.2$). The only cluster with richness class 5 is A 665.

The subdivision into six *distance classes* is based on the apparent magnitude of the tenth brightest galaxy, in accordance with the redshift estimate for the cluster. Hence, the distance class provides a coarse measure of the distance.

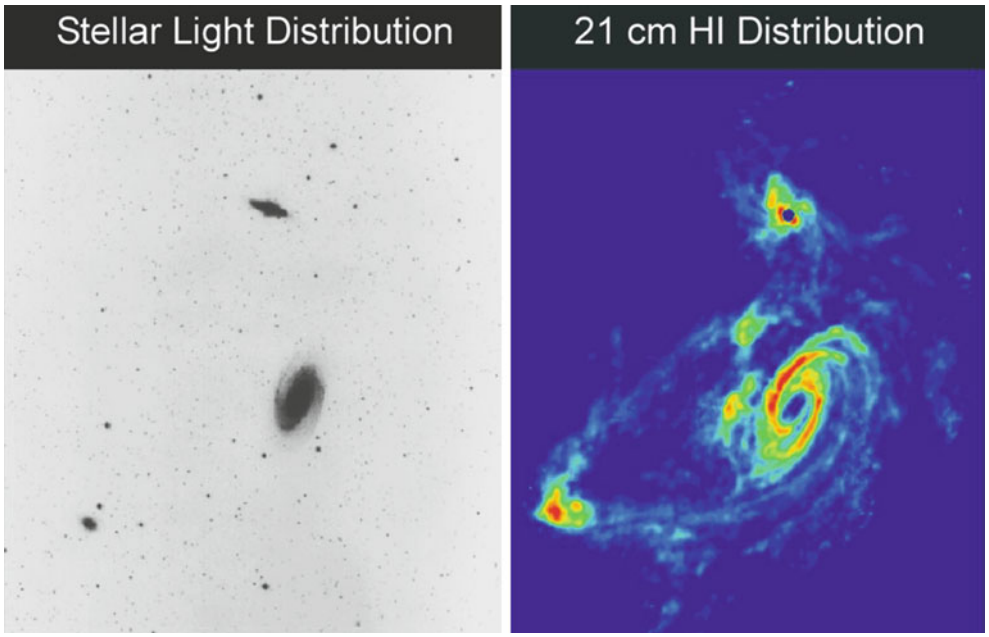


Fig. 6.9 *The left panel* shows an optical image of the galaxies M81 (*bottom*) and M82 (*top*), two members of the M81-group, about 3.1 Mpc away (see also Fig. 1.3 for a detailed view of M82). These two galaxies are moving around each other, and the gravitational interaction taking place, as clearly seen in the distribution of atomic hydrogen (*right panel*) which has been stripped off the galaxies due to gravitational interactions, may be the reason for the violent star formation in M82. M82 is an archetypical starburst galaxy. Credit: Image courtesy of National Radio Astronomy Observatory/AUI

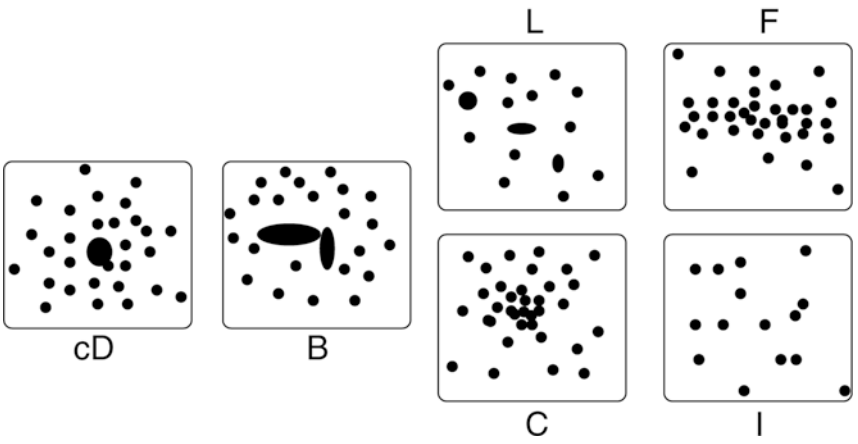


Fig. 6.10 Rough morphological classification of clusters by Rood & Sastry: cDs are those which are dominated by a central cD galaxy, Bs contain a pair of bright galaxies in the center. Ls are clusters with a nearly linear alignment of the dominant galaxies, Cs have a single core of galaxies, Fs are clusters with an oblate galaxy distribution, and Is are clusters with an irregular distribution. This classification has the more regular clusters at the *left* and irregular clusters at the *right*

6.2.2 Morphological classification of clusters

Clusters are also classified by the morphology of their galaxy distribution. Several classifications are used, one of which is displayed in Fig. 6.10. Since this is a description of the visual impression of the galaxy distribution, the exact class of a cluster is not of great interest. However, a rough classification can provide an idea of the state of a cluster, i.e., whether it is currently in dynamical equilibrium or whether it has been heavily disturbed by a merger process with another cluster. Therefore, one distinguishes in particular between regular and irregular clusters, and also those which are intermediate; the transition between classes is of course continuous. Regular clusters are 'compact' whereas, in contrast, irregular clusters are 'open' (Zwicky's classification criteria).

This morphological classification indeed points at physical differences between clusters, as correlations between morphology and other properties of galaxy clusters show. For example, it is found that regular clusters are completely dominated by early-type galaxies, whereas irregular clusters have a fraction of spirals nearly as large as in the general distribution of field galaxies. Very often, regular clusters are dominated by a cD galaxy at the center, and their central galaxy density is very high. In contrast, irregular clusters are significantly less dense in the center. Irregular clusters often show strong substructure, which is rarely found in regular clusters. Furthermore, regular clusters have a high richness, whereas irregular clusters have fewer cluster members. To summarize, regular clusters can be said to be in a relaxed state, whereas irregular clusters are still in the process of evolution.

6.2.3 Galaxy groups

Accumulations of galaxies that do not satisfy Abell's criteria are in most cases galaxy groups. Hence, groups are the continuation of clusters towards fewer member galaxies and are therefore presumably of lower mass, lower velocity dispersion, and smaller extent. The distinction between groups and clusters is at least partially arbitrary. It was defined by Abell mainly to be not too heavily affected by projection effects in the identification of clusters. Groups are of course more difficult to detect, since the overdensity criterion for them is more sensitive to projection effects by foreground and background galaxies than for clusters.

Table 6.2 Definition of Abell's richness classes

Richness class R	N	Number in Abell's catalog
(0)	(30–49)	($\geq$ 1000)
1	50–79	1224
2	80–129	383
3	130–199	68
4	200–299	6
5	$\geq$ 300	1

N is the number of cluster galaxies with magnitudes between m_3 and $m_3 + 2$ inside the Abell radius (6.6), where m_3 is the brightness of the third brightest cluster galaxy

Table 6.3 Definition of Abell's distance classes

Distance class	m_{10}	Estimated average redshift	Number in Abell's catalog with $R \geq 1$
1	13.3–14.0	0.0283	9
2	14.1–14.8	0.0400	2
3	14.9–15.6	0.0577	33
4	15.7–16.4	0.0787	60
5	16.5–17.2	0.131	657
6	17.3–18.0	0.198	921

m_{10} is the magnitude of the tenth brightest cluster galaxy

A special class of groups are the *compact groups*, assemblies of (in most cases, few) galaxies with very small projected separations. The best known examples for compact groups are Stephan's Quintet and Seyfert's Sextet (see Fig. 6.11). In 1982, a catalog of 100 compact groups (Hickson Compact Groups, HCGs) was published, where a group consists of four or more bright members. These were also selected on POSS photoplates, again solely by an overdensity criterion. The median redshift of the HCGs is about $z = 0.03$. Further examples of optical images of HCGs are given in Figs. 6.3 and 1.20.

Follow-up spectroscopic studies of the HCGs have verified that 92 of them have at least three galaxies with conforming redshifts, defined such that the corresponding recession velocities lie within 1000 km/s of the median velocity of group members. Of course, the similarity in redshift does not necessarily imply that these groups form a gravitationally bound and relaxed system. For instance, the galaxies could be tracers of an overdense structure which we happen to view from a direction where the galaxies are projected near each other on the sky. However, more than 40 % of the galaxies in HCGs show evidence of interactions, indicating that these galaxies have near neighbors in three-dimensional space. Furthermore, about three quarters of HCGs with four or more member galaxies show extended X-ray emission, most likely coming from intra-group hot gas, providing additional evidence for the presence of a common gravitational potential well (see Sect. 6.4). Compared to clusters, the intergalactic gas in groups has a lower temperature and, possibly, lower metallicity.

6.2.4 Modern optical cluster catalogs

The subjectivity of selecting overdensities on images by eye can of course be overcome by using digital (or digitized) astronomical images and employing algorithms to apply criteria to the data which define an overdensity, or a cluster, respectively. This approach solves one of the aforementioned problems in optical cluster searches. The other problem—namely projection effects—can be overcome if an additional distance measure for potential member galaxies can be applied.[3] There are two ways how such a distance indicator can be obtained: one either uses large spectroscopic catalogs of galaxies, such as the SDSS, or, as will be discussed next, one can employ the colors of early-type galaxies.

Color-magnitude diagram. We mentioned before that a large fraction of galaxies in clusters are early-type galaxies. Furthermore, we saw in Sect. 3.6 that early-type galaxies have rather uniform colors. Indeed, plotting the color of cluster galaxies versus their magnitude, one finds a very well-defined, nearly horizontal sequence (Fig. 6.12). This red cluster sequence (RCS) is populated by the early-type galaxies in the cluster.

The scatter of early-type galaxies around this sequence is very small, which suggests that all early-type galaxies in a cluster have nearly the same color, only weakly depending on luminosity. The small slope seen in Fig. 6.12 is mainly due to the fact that more massive ellipticals have a somewhat higher metallicity, rendering the stellar emission slightly redder. Even more surprising is the fact that the color-magnitude diagrams of different clusters at the same redshift define a very similar red cluster sequence: early-type cluster galaxies with the same redshift and luminosity have virtually the same color. Comparing the red sequences of clusters at different redshifts, one finds that the sequence of cluster galaxies is redder the higher the redshift is. This effect is caused by the redshift of the galaxies, which shift their spectral energy distribution towards longer wavelengths. Hence, by

[3] There are also other methods that have been used to construct cluster catalogs, which run under name of 'matched filter' techniques. They assume that galaxies are not only a collection of galaxies, but that the galaxy overdensity has certain properties. For example, the number density of cluster galaxies is expected to have a particular radial density profile and their distribution in luminosity should approximately follow a Schechter-type luminosity function. Applying these criteria to galaxy overdensities leads to cleaner cluster selection than a pure overdensity criterion. As a drawback, these criteria are more likely to select regular clusters than irregular ones, due to the assumed density profile.

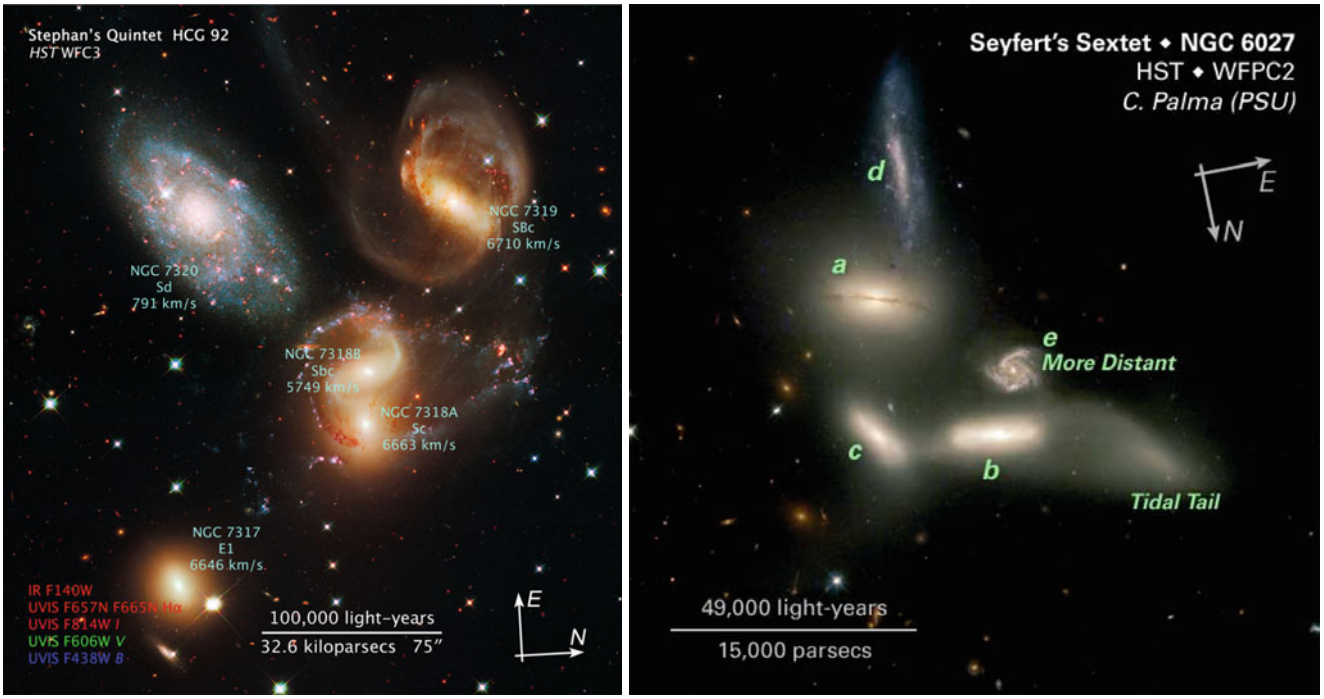


Fig. 6.11 *Left panel*: Stephan's Quintet, also known as Hickson Compact Group 92, is a very dense accumulation of galaxies with a diameter of about 80 kpc. The galaxy at the upper left (NGC 7320) is not a member of the group: its redshift indicates that it has a much smaller distance from us than the other four galaxies; in fact, it is close enough to us for HST being able to resolve individual stars. This galaxy has only $\sim 2\,\%$ of the luminosity of the other galaxies shown, i.e., it is an actively star-forming dwarf galaxy, as is also seen by its much bluer color compared to the other galaxies in the field. The remaining three spiral galaxies of the group show clear signs of interactions—distorted spiral arms and tidal tails. The strong interaction of the galaxy pair in the middle of the image gives rise to a strong burst of star formation. The elliptical galaxy at the bottom left appears to be less affected by galaxy interactions. The image is a color composite of optical and near-IR images, as well as a narrow band image at the Hα wavelength, all taken with the WFC3 instrument onboard HST. *Right panel*: Seyfert's Sextet, an apparent accumulation of six galaxies located very close together on the sphere. Only four of the galaxies (a–d) in fact belong to the group; the spiral galaxy (e) is located at significantly larger distance. Another object originally classified as a galaxy is no galaxy but instead a tidal tail that was ejected in tidal interactions of galaxies in the group. Credit: *Left*: NASA, ESA, and the Hubble SM4 ERO Team; *Right*: NASA, J. English (U. Manitoba), C. Palma, S. Hunsberger, S. Zonak, J. Charlton, S. Gallagher (PSU), and L. Frattare (STScI)

keeping the observed filter bands constant, the colors change as a function of redshift. In fact, the red cluster sequence is so precisely characterized that, from the color-magnitude diagram of a cluster alone, its redshift can be estimated with very high accuracy, provided the photometric calibration is sufficiently good. Furthermore, the accuracy of this estimated redshift strongly depends on the choice of the filters between which the color is measured. Since the most prominent spectral feature of early-type galaxies is the 4000 Å-break, the redshift is estimated best if this rest-frame wavelength, redshifted to $4000\,(1+z)$ Å, is well covered by the photometric bands employed.

This well-defined red cluster sequence is of crucial importance for our understanding of the evolution of galaxies. We know from Sect. 3.5 that the composition of a stellar population depends on the mass spectrum at its birth (the initial mass function, IMF) and on its age: the older a population is, the redder it becomes. The fact that cluster galaxies at the same redshift all have roughly the same color indicates that their stellar populations have very similar ages. However, the only age that is singled out is the age of the Universe itself. In fact, the color of cluster galaxies is compatible with their stellar populations being roughly the same age as the Universe at that particular redshift. This also provides an explanation for why the red cluster sequence is shifted towards *intrinsically* bluer colors at higher redshifts—there, the age of the Universe was smaller, and thus the stellar population was younger. This effect is of particular importance at high redshifts.

The RCS Survey. The cluster red sequence method was used in several multi-band imaging surveys for the detection of clusters. In fact, one of the large imaging surveys carried out with the CFHT was the RCS survey, with its main purpose to detect clusters out to large redshifts. It covered $100\,\mathrm{deg}^2$ in two filters, and yielded more than 1000 cluster and group candidates, out to redshifts larger than unity. As a follow-up, the RCS II survey covers $900\,\mathrm{deg}^2$ in three filters, and aims at detecting some 10^4 clusters out to $z \sim 1$.

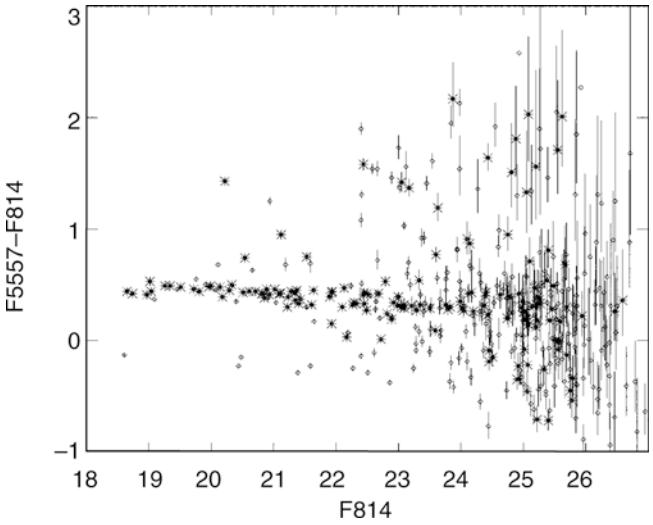


Fig. 6.12 Color-magnitude diagram of the cluster of galaxies Abell 2390, observed with the HST. *Star symbols* represent early-type galaxies, identified by their morphology, while *diamonds* denote other galaxies in the field. The red cluster sequence is clearly visible. Note that, due to projection effects, not all galaxies shown here are indeed cluster members; some of them are foreground or background galaxies. Source: M. Gladders & H. Yee 2000, *A New Method For Galaxy Cluster Detection. I. The Algorithm*, AJ 120, 2148, p. 2150, Fig. 1. ©AAS. Reproduced with permission

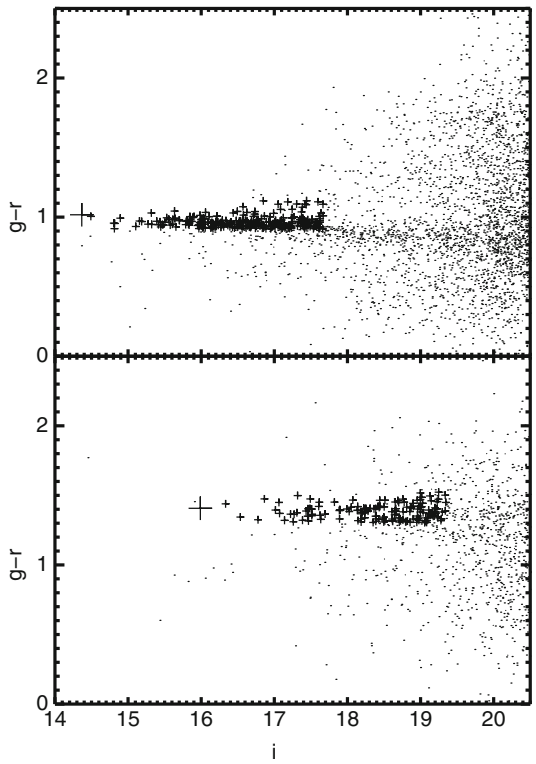


Fig. 6.13 Example of two clusters found by the maxBCG method. Shown are the color-magnitude relations for galaxies in the field of Abell 2142 at $z = 0.092$ (*top*) and Abell 1682 at $z = 0.23$ (*bottom*). In both cases, all galaxies within $2h^{-1}$ Mpc of the BCG are plotted as *small dots*. The BCG itself is denoted by a *big cross*, being the most luminous of the cluster members, whereas the smaller crosses show galaxies with $L \geq 0.4L^*$ whose colors lie within $\pm 2\sigma$ of the red sequence, which is 0.05 and 0.06 for the $g - r$ and $r - i$, respectively. If they lie closer that the estimated R_{200} from the BCG, they are considered to be cluster members. We note that the red sequence has almost zero slope in the color shown here. Although these two clusters were known before, their rediscovery provides one of the tests of the method. Source: B.P. Koester et al. 2007, *MaxBCG: A Red-Sequence Galaxy Cluster Finder*, ApJ 660, 221, p. 224, Fig. 1. ©AAS. Reproduced with permission

The maxBCG catalog. The large sky coverage of the SDSS, and the very homogeneous photometry of the five-band imaging data, makes this survey a prime resource for optical cluster finders. Whereas the SDSS is rather shallow, compared to the RCS surveys, and thus cannot find clusters at high redshifts, it enables the most complete cluster searches in the more local Universe. Therefore, several cluster catalogs have been constructed from the SDSS, one of which we want to briefly describe here.

This maxBCG cluster catalog is based on a search algorithm which makes use of three properties of massive clusters. The first is the already mentioned red cluster sequence, i.e., the homogeneous (and redshift dependent) color of early-type galaxies in clusters. Second, most massive clusters are found to have a dominant central galaxy, the brightest cluster galaxy (BCG), whose luminosity can be several times larger than the second brightest cluster member. The third property relates to the radial density profile of galaxies, which, in a first approximation, decreases roughly as $1/\theta$ from the center to the outside.

Thus, the algorithm searches for overdensities of galaxies with similar color, corresponding to the color of the red sequence within a specified redshift interval, where the brightest of the galaxies is located near the center of the overdensity, and where the radial decline of the galaxy number density is compatible with a $1/\theta$-law. More specifically, the maxBCG method searches for concentrations of luminous red galaxies in the redshift interval $0.1 \lesssim z \lesssim 0.3$, whose colors agree to within $\pm 2\sigma$ of the width of the red sequence in color and with the brightest of these galaxies near the center (Fig. 6.13). The choice of this redshift interval is motivated by the fact that the $g - r$ color of red galaxies is a simple function of redshift, as the strong 4000 Å-break of early-type galaxies moves through the g-filter in this redshift interval. The color of the overdense population yields an indication of the redshift, which then allows one to obtain the galaxy luminosity from the observed flux. Only red galaxies more luminous than $0.4L^*$ are taken into account. Given the depth of the SDSS, a red galaxy with $L \geq 0.4L^*$ at $z = 0.3$ can be detected—hence, this provides a volume-limited survey for such galaxies. Furthermore, the redshift estimate is used to obtain the physical projected radius R from the observed angular separation.

To characterize the cluster candidate, the number of red sequence galaxies with $L \geq 0.4L^*$ within $1h^{-1}$ Mpc of the BCG candidate, N_g, is calculated. For reasons that will become clear when we discuss the formation of dark matter halos in Sect. 7.5.1 (see also Problem 6.1), one defines the 'extent' of a cluster to be the radius inside of which the mean density is 200 times larger than the critical density of the Universe at this redshift. This radius is denoted by r_{200}, and the mass of the cluster within r_{200} is then denoted as $M_{200} = (4\pi/3)200\rho_{cr}(z)r_{200}^3$, often also called the virial mass of the cluster. From earlier cluster studies, it was found that there is a close relation between r_{200} (or M_{200}) and the number of galaxies within $1h^{-1}$ Mpc, roughly following $r_{200} \propto N_g$. With this estimate of the virial radius, the number of red sequence members within projected radius $R = r_{200}$ is measured and denoted by $N_{g,200}$, which is then called the richness of the cluster.

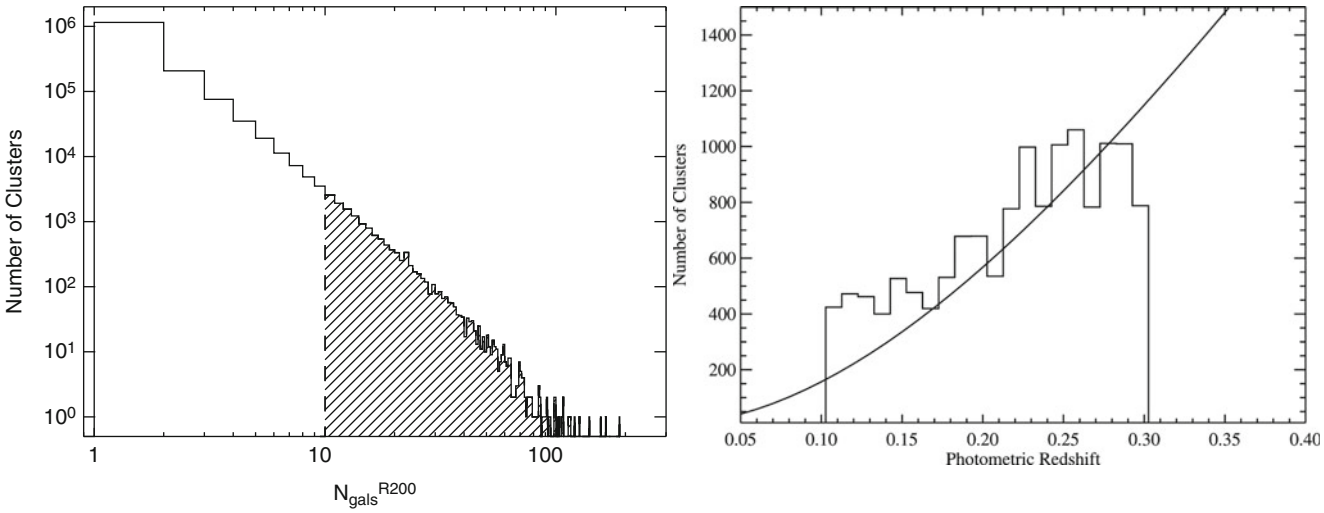


Fig. 6.14 *Left*: Histogram of the number of clusters found by the maxBCG method, as a function of cluster richness $N_{g,200}$. The maxBCG catalog consists of 13 823 clusters with $N_{g,200} \geq 10$, shown as hatched region in the histogram. The *right panel* shows the distribution of estimated redshifts, whereas the *solid curve* is the expected redshift distribution for a volume-limited survey with the sky area of the SDSS at a given mean number density of $2.3 \times 10^{-5} h^3\,\mathrm{Mpc}^{-3}$. Source: B.P. Koester et al. 2007, *A MaxBCG Catalog of 13,823 Galaxy Clusters from the Sloan Digital Sky Survey*, ApJ 660, 239, p. 243, 244, Figs. 3, 4. ©AAS. Reproduced with permission

These criteria have yielded a catalog of 13 823 clusters with $N_{g,200} \geq 10$ in the 7500 deg^2 of the SDSS. Their distribution in richness and redshift is shown in Fig. 6.14. This maxBCG catalog is one of the largest cluster catalogs available up to now and has been widely used.

The quality of the catalog can be assessed in a number of ways. Since the SDSS also has a large spectroscopic component, the spectroscopic redshifts for more than 5000 of the BCGs are known; they can be compared to the redshift estimated from the color of the red-sequence cluster members. The difference between the two redshifts has a very narrow distribution with a width of $\sigma_z \sim 0.01$—that is, the estimated cluster redshifts are very accurate.

A good catalog should be pure and complete. As mentioned before, purity measures the fraction of objects included in the catalog which are not real clusters, whereas completeness quantifies the number of real clusters which were missed by the selection algorithm. These two quantities can be estimated from simulations, in which mock cluster catalogs are generated and analyzed with the same detection algorithm as the real data. Based on such simulations, one concludes that the maxBCG catalog is $\sim$ 90 % pure and $\sim$ 85 % complete, for clusters with masses $\geq 10^{14} M_\odot$, corresponding to $N_{g,200} \approx 10$.

The available spectroscopy of the SDSS can also be used to determine the velocity dispersion σ_v for many of the clusters. The left panel of Fig. 6.15 shows a strong correlation between σ_v and richness, well fit by a power law of the form

$$\sigma_v = 500\,\mathrm{km/s} \left(\frac{N_{g,200}}{10} \right)^{0.31} . \qquad (6.7)$$

This strong correlation also shows that cluster richness is a good indicator of cluster mass, since σ_v is expected to be tightly related to the mass, according to the virial theorem; we will return to this aspect soon. The comoving number density of clusters as a function of redshift is shown in the right panel of Fig. 6.15, for different richness bins. The comoving space density is roughly constant for all bins except for the richest, which suggests that the maxBCG catalog does not suffer from serious redshift-dependent incompleteness. The slight decline with increasing redshift in the richest bin is actually expected from structure formation in the Universe, as we will discuss in Sect. 7.5.2.

For candidates with $N_{g,200} < 10$, i.e., mass below $\sim 10^{14}\,M_\odot$, the purity and completeness decrease; hence, whereas a large fraction of these candidates are probably clusters or groups at lower mass, projection effects will play an increasingly important role for decreasing richness. To reliably find groups, the selection criteria need to be sharpened, which can be done using spectroscopic redshifts. With those, one can search for galaxy overdensities on the sky which have the same redshifts within a few times the expected velocity dispersion in groups, i.e., the same radial velocity within $\sim$ 1000 km/s. This provides a much stricter redshift constraint than is possible with the red sequence method and thus substantially reduces projection effects. Such group catalogs were constructed from the Two-degree field Galaxy Redshift Survey (see Sect. 8.1.2) and the SDSS as well. The velocity dispersion in groups is significantly smaller than that in clusters; typical values for groups with only a few members are $\sigma_v \sim 300$ km/s (see Fig. 6.15).

6.3 Light distribution and cluster dynamics

6.3.1 Spatial distribution of galaxies

Most regular clusters show a centrally condensed number density distribution of cluster galaxies, i.e., the galaxy den-

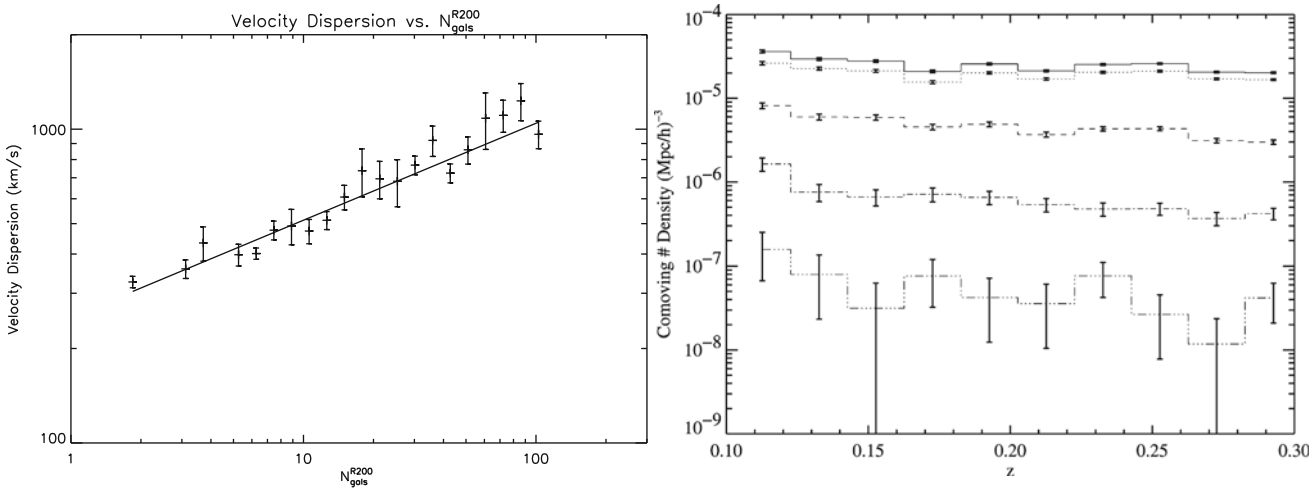


Fig. 6.15 *Left*: The velocity dispersion of maxBCG clusters, as a function of richness $N_{\mathrm{g},200}$. Note that the figure extends to richness as small as $N_{\mathrm{g},200} = 2$. At the threshold of the cluster catalog, $N_{\mathrm{g},200} = 10$, the characteristic velocity dispersion is $\sim 500\,\mathrm{km/s}$. *Right*: The comoving number density of clusters in the maxBCG catalog, as a function of redshift, for four different redshift bins (*from top to bottom*: $10 \le N_{\mathrm{g},200} < 20$, $20 \le N_{\mathrm{g},200} < 43$, $43 \le N_{\mathrm{g},200} < 91$, $91 \le N_{\mathrm{g},200} < 189$). Source: B.P. Koester et al. 2007, *A MaxBCG Catalog of 13,823 Galaxy Clusters from the Sloan Digital Sky Survey*, ApJ 660, 239, p. 251, Figs. 12, 13. ©AAS. Reproduced with permission

sity increases strongly towards the center. If the cluster is not very elongated, this density distribution can be assumed, to a first approximation, as being spherically symmetric. Only the projected density distribution $N(R)$ is observable. This is related to the three-dimensional number density $n(r)$ through

$$N(R) = \int_{-\infty}^{\infty} \mathrm{d}r_3 \, n\left(\sqrt{R^2 + r_3^2}\right) = 2\int_R^{\infty} \frac{\mathrm{d}r \; r\, n(r)}{\sqrt{r^2 - R^2}} \,, \tag{6.8}$$

where in the second step a simple transformation of the integration variable from the line-of-sight coordinate r_3 to the three-dimensional radius $r = \sqrt{R^2 + r_3^2}$ was made.

Of course, no function $N(R)$ can be observed, but only points (the positions of the galaxies) that are distributed in a certain way. If the number density of galaxies is sufficiently large, $N(R)$ is obtained by smoothing the point distribution. Alternatively, one considers parametrized forms of $N(R)$ and fits the parameters to the observed galaxy positions. In most cases, the second approach is taken because its results are more robust. A parametrized distribution needs to contain at least five parameters to be able to describe at least the basic characteristics of a cluster. Two of these parameters describe the position of the cluster center on the sky. One parameter is used to describe the amplitude of the density, for which, e.g., the central density $N_0 = N(0)$ may be used. A fourth parameter is a characteristic scale of a cluster, often taken to be the core radius r_{c}, defined such that at $R = r_{\mathrm{c}}$, the projected density has decreased to half the central value, $N(r_{\mathrm{c}}) = N_0/2$. Finally, one parameter is needed to describe 'where the cluster ends'; the Abell radius is a first approximation for such a parameter.[4]

Parametrized cluster models can be divided into those which are physically motivated, and those which are of a purely mathematical nature. One example for the latter is the family of Sérsic profiles which is not derived from dynamical models. Next, we will consider a class of distributions that are based on a dynamical model.

Isothermal distributions. These models are based on the assumption that the velocity distribution of the massive particles (this may be both galaxies in the cluster or dark matter particles) of a cluster is locally described by a Maxwell distribution, i.e., they are 'thermalized'. As shown from spectroscopic analyses of the distribution of the radial velocities of cluster galaxies, this is not a bad assumption. Assuming, in addition, that the mass profile of the cluster follows that of the galaxies (or vice versa), and that the temperature (or equivalently the velocity dispersion) of the distribution does not depend on the radius, so that one has an isothermal distribution of galaxies, then one obtains a one-parameter set of models, the so-called *isothermal spheres*. These can be described physically as follows.

In dynamical equilibrium, the pressure gradient must be equal to the gravitational acceleration, so that

$$\boxed{\frac{\mathrm{d}P}{\mathrm{d}r} = -\rho\,\frac{GM(r)}{r^2}} \,, \tag{6.9}$$

where $\rho(r)$ denotes the density of the distribution, e.g., the density of galaxies. By $\rho(r) = \langle m\rangle\, n(r)$, this mass density is related to the number density $n(r)$, where $\langle m\rangle$ is the average particle mass. $M(r) =$

[4] In principle, one might need more parameters for describing the radial profile of clusters, and thus the parameters mentioned are a bare minimum. However, as we shall see later, it appears that the mass profiles of clusters are all very similar. Cosmological simulations of structure evolution in the Universe predict that the density profiles of clusters can indeed be characterized by this minimum set.

$4\pi \int_0^r \mathrm{d}r'\, r'^2\, \rho(r')$ is the mass of the cluster enclosed within a radius r. By differentiation of (6.9), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^2}{\rho}\frac{\mathrm{d}P}{\mathrm{d}r}\right) + 4\pi G r^2 \rho = 0\,. \tag{6.10}$$

The relation between pressure and density is $P = nk_\mathrm{B}T$. On the other hand, the temperature is related to the velocity dispersion of the particles,

$$\frac{3}{2}k_\mathrm{B}T = \frac{\langle m\rangle}{2}\langle v^2\rangle\,, \tag{6.11}$$

where $\langle v^2\rangle$ is the mean squared velocity, i.e., the velocity dispersion, provided the average velocity vector is set to zero. The latter assumption means that the cluster does not rotate, or contract or expand. If T (or $\langle v^2\rangle$) is independent of r, then

$$\frac{\mathrm{d}P}{\mathrm{d}r} = \frac{k_\mathrm{B}T}{\langle m\rangle}\frac{\mathrm{d}\rho}{\mathrm{d}r} = \frac{\langle v^2\rangle}{3}\frac{\mathrm{d}\rho}{\mathrm{d}r} = \sigma_v^2\frac{\mathrm{d}\rho}{\mathrm{d}r}\,, \tag{6.12}$$

where σ_v^2 is the one-dimensional velocity dispersion, e.g., the velocity dispersion along the line-of-sight, which can be measured from the redshifts of the cluster galaxies. If the velocity distribution corresponds to an isotropic (Maxwell) distribution, the one-dimensional velocity dispersion is exactly 1/3 times the three-dimensional velocity dispersion, because of $\langle v^2\rangle = \sigma_x^2 + \sigma_y^2 + \sigma_z^2$, or

$$\sigma_v^2 = \frac{\langle v^2\rangle}{3}\,. \tag{6.13}$$

With (6.10), it then follows that

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{\sigma_v^2 r^2}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}r}\right) + 4\pi G r^2 \rho = 0\,. \tag{6.14}$$

Singular isothermal sphere. For general boundary conditions, the differential equation (6.14) for $\rho(r)$ cannot be solved analytically. However, one particular analytical solution of the differential equation exists: By substitution, we can easily show that

$$\boxed{\rho(r) = \frac{\sigma_v^2}{2\pi G r^2}} \tag{6.15}$$

solves (6.14). This density distribution is called *singular isothermal sphere*; we have encountered it before, in the discussion of gravitational lens models in Sect. 3.11.2. This distribution has a diverging density as $r \to 0$ and an infinite total mass $M(r) \propto r$. It is remarkable that this density distribution, $\rho \propto r^{-2}$, is just what is needed to explain the flat rotation curves of galaxies at large radii.

The divergence of the density towards the center may not appear reasonable, and thus one might search solutions of (6.14) with the more physical boundary conditions $\rho(0) = \rho_0$, the central density, and $(\mathrm{d}\rho/\mathrm{d}r)_{|r=0} = 0$, for the density profile to be flat at the center. Numerical solutions of (6.14) with these boundary conditions (thus, with a flat core) reveal that the central density and the core radius are related to each other by

$$\boxed{\rho_0 = \frac{9\sigma_v^2}{4\pi G r_\mathrm{c}^2}}\,. \tag{6.16}$$

Hence, these physical solutions of (6.14) avoid the infinite density of the singular isothermal sphere. However, these solutions also decrease outwards with $\rho \propto r^{-2}$, so they have a diverging mass as well. The origin of this mass divergence is easily understood because these isothermal distributions are based on the assumption that the velocity distribution is isothermal, thus Maxwellian with a spatially constant temperature. A Maxwell distribution has wings, hence it (formally) contains particles with arbitrarily high velocities. Since the distribution is assumed stationary, such particles must not escape, so their velocity must be lower than the escape velocity from the gravitational well of the cluster. But for a Maxwell distribution this is only achievable for an infinite total mass.

King models. To remove the problem of the diverging total mass, self-gravitating dynamical models with an upper cut-off in the velocity distribution of their constituent particles are introduced. These are called *King models* and cannot be expressed analytically. However, an analytical approximation exists for the central region of these mass profiles,

$$\boxed{\rho(r) = \rho_0\left[1 + \left(\frac{r}{r_\mathrm{c}}\right)^2\right]^{-3/2}}\,. \tag{6.17}$$

Using (6.8), we obtain from this the projected surface mass density

$$\boxed{\Sigma(R) = \Sigma_0\left[1 + \left(\frac{R}{r_\mathrm{c}}\right)^2\right]^{-1} \quad \text{with} \quad \Sigma_0 = 2\rho_0 r_\mathrm{c}}\,. \tag{6.18}$$

The analytical fit (6.17) of the King profile also has a diverging total mass, but this divergence is 'only' logarithmic.

These analytical models for the density distribution of galaxies in clusters are only approximations, of course, because the galaxy distribution in clusters is often heavily structured. Furthermore, these dynamical models are applicable to a galaxy distribution only if the galaxy number density follows the matter density. However, one finds that the distribution of galaxies in a cluster often depends on the galaxy type. The fraction of early-type galaxies (Es and S0s) is often largest near the center. Therefore, one should consider the possibility that the distribution of galaxies in a cluster may be different from that of the total matter. A typical value for the core radius is about $r_\mathrm{c} \sim 0.25h^{-1}$ Mpc.

6.3.2 Dynamical mass of clusters

The above argument relates the velocity distribution of cluster galaxies to the mass profile of the cluster, and from this we obtain physical models for the density distribution. This implies the possibility of deriving the mass, or the mass profile, respectively, of a cluster from the observed velocities of cluster galaxies. We will briefly present this method of mass determination here. For this, we consider the dynamical time-scale of clusters, defined as the time a typical galaxy needs to traverse the cluster once,

$$\boxed{t_{\rm cross} \sim \frac{R_{\rm A}}{\sigma_v} \sim 1.5 h^{-1} \times 10^9 \,{\rm yr}} \,, \tag{6.19}$$

where a (one-dimensional) velocity dispersion $\sigma_v \sim 1000\,{\rm km/s}$ was assumed. The dynamical time-scale is shorter than the age of the Universe. One therefore concludes that clusters of galaxies are gravitationally bound systems. If this were not the case they would dissolve on a timescale $t_{\rm cross}$. Since $t_{\rm cross} \ll t_0$ one assumes a *virial equilibrium*, hence that the virial theorem applies, so that in a time-average sense,

$$2E_{\rm kin} + E_{\rm pot} = 0 \,, \tag{6.20}$$

where

$$E_{\rm kin} = \frac{1}{2}\sum_i m_i\, v_i^2 \quad ; \quad E_{\rm pot} = -\frac{1}{2}\sum_{i\neq j} \frac{G m_i m_j}{r_{ij}} \tag{6.21}$$

are the kinetic and the potential energy of the cluster galaxies, m_i is the mass of the i-th galaxy, v_i is the absolute value of its velocity, and r_{ij} is the spatial separation between the i-th and the j-th galaxy. The factor $1/2$ in the definition of $E_{\rm pot}$ occurs since each pair of galaxies occurs twice in the sum.

In writing (6.21) we have assumed that the total mass of the cluster is the sum of all its member galaxies,

$$M := \sum_i m_i \,. \tag{6.22}$$

This assumption is not valid, since, as we will find below, most of the cluster mass is not contained in galaxies. However, if we assume that the total mass is distributed in the same way as the galaxies are, we can associate to each galaxy a 'representative' mass, so that the superposition of all these representative masses yields the total mass distribution of the cluster. The mass m_i used in the foregoing equations and below is meant to be this representative mass.

We further define the velocity dispersion, weighted by mass,

$$\langle v^2 \rangle := \frac{1}{M}\sum_i m_i\, v_i^2 \tag{6.23}$$

and the gravitational radius,

$$r_{\rm G} := 2M^2 \left(\sum_{i\neq j} \frac{m_i\, m_j}{r_{ij}} \right)^{-1} \,. \tag{6.24}$$

With this, we obtain

$$E_{\rm kin} = \frac{M}{2}\langle v^2 \rangle \quad ; \quad E_{\rm pot} = -\frac{G\,M^2}{r_{\rm G}} \tag{6.25}$$

for the kinetic and potential energy. Applying the virial theorem (6.20) yields the mass estimate

$$\boxed{M = \frac{r_{\rm G}\langle v^2 \rangle}{G}} \,. \tag{6.26}$$

Transition to projected quantities. The above derivation uses the three-dimensional separations r_i of the galaxies from the cluster center, which are, however, not observable. To be able to apply these equations to observations, they need to be transformed to projected separations. If the galaxy positions and the directions of their velocity vectors are uncorrelated, as it is the case, e.g., for an isotropic velocity distribution, then

$$\langle v^2 \rangle = 3\sigma_v^2, \;\; r_{\rm G} = \frac{\pi}{2} R_{\rm G} \;\; \text{with} \;\; R_{\rm G} = 2M^2 \left(\sum_{i\neq j} \frac{m_i m_j}{R_{ij}} \right)^{-1} \,, \tag{6.27}$$

where R_{ij} denotes the projected separation between the galaxies i and j. The parameters σ_v and $R_{\rm G}$ are direct observables; thus, the total mass of the cluster can be determined. One obtains

$$\boxed{M = \frac{3\pi R_{\rm G}\sigma_v^2}{2G} = 1.1 \times 10^{15} M_\odot \left(\frac{\sigma_v}{1000{\rm km/s}} \right)^2 \left(\frac{R_{\rm G}}{1{\rm Mpc}} \right)} \,. \tag{6.28}$$

We explicitly point out that this mass estimate no longer depends on the masses m_i of the individual galaxies—rather the galaxies are now test particles in the gravitational potential. With $\sigma_v \sim 1000\,{\rm km/s}$ and $R_{\rm G} \sim 1\,{\rm Mpc}$ as typical values for rich clusters of galaxies, one obtains a characteristic mass of $\sim 10^{15} M_\odot$ for rich clusters.

The 'missing mass' problem in clusters of galaxies. With the mass M and the number N of galaxies, one can now derive a characteristic mass $m = M/N$ for the luminous galaxies. This mass is found to be very high, $m \sim 10^{13} M_\odot$. Alternatively, M can be compared with the total optical luminosity of the cluster galaxies, $L_{\rm tot} \sim 10^{12}$-$10^{13}\, L_\odot$, and

hence the mass-to-light ratio can be calculated; typically

$$\boxed{\left(\frac{M}{L_{\rm tot}}\right) \sim 300\,h\left(\frac{M_\odot}{L_\odot}\right)}\,. \tag{6.29}$$

This value exceeds the M/L-ratio of early-type galaxies by at least a factor of 10. Realizing this discrepancy, Fritz Zwicky concluded as early as 1933, from an analysis of the Coma cluster, that clusters of galaxies must contain considerably more mass than is visible in galaxies—the dawn of the *missing mass problem*. As we will see further below, this discrepancy between the observed luminosity and estimated mass has by now been firmly established, since other methods for the mass determination of clusters also yield comparable values and indicate that a major fraction of the mass in galaxy clusters consists of (non-baryonic) dark matter. *The stars visible in galaxies contribute less than about 5 % to the total mass in clusters of galaxies.*

6.3.3 Additional remarks on cluster dynamics

Given the above line of argument, the question of course arises as to whether the application of the virial theorem is still justified if the main fraction of mass is not contained in galaxies. The derivation remains valid in this form as long as the spatial distribution of galaxies follows the total mass distribution. The dynamical mass determination can be affected by an anisotropic velocity distribution of the cluster galaxies and by the possibly non-spherical cluster mass distribution. In both cases, projection effects, which are dealt with relatively easily in the spherically-symmetric case, obviously become more complicated. This is also one of the reasons for the necessity to consider alternative methods of mass determination.

Two-body collisions of galaxies in clusters are of no importance dynamically, as is easily seen from the corresponding relaxation time-scale (3.3),

$$t_{\rm relax} = t_{\rm cross}\frac{N}{\ln N}\,,$$

which is much larger than the age of the Universe. The motion of galaxies is therefore governed by the collective gravitational potential of the cluster. The velocity dispersion is approximately the same for the different types of galaxies, and also only a weak tendency exists for a dependence of σ_v on galaxy luminosity, restricted to the brightest ones (see below in Sect. 6.7.2). From this, we conclude that the galaxies in a cluster are not "thermalized" because this would mean that they all have the same mean kinetic energy, implying $\sigma_v \propto m^{-1/2}$. Furthermore, the independence of σ_v of L reconfirms that collisions of galaxies with each other are not dynamically relevant; rather, the velocity distribution of galaxies is defined by collective processes during cluster formation.

Violent relaxation. One of the most important of the aforementioned processes is known as *violent relaxation*. This process very quickly establishes a virial equilibrium in the course of the gravitational collapse of a mass concentration. The reason for it are the small-scale density inhomogeneities within the collapsing matter distribution which generate, via Poisson's equation, corresponding fluctuations in the gravitational field. These then scatter the infalling particles and, by this, the density inhomogeneities are further amplified. The fluctuations of the gravitational field act on the matter like scattering centers. In addition, these field fluctuations change over time, yielding an effective exchange of energy between the particles. In a statistical average, all galaxies obtain the same velocity distribution by this process. As confirmed by numerical simulations, this process takes place on a time-scale of $t_{\rm cross}$, i.e., roughly as quickly as the collapse itself.

Dynamical friction. Another important process for the dynamics of galaxies in a cluster is *dynamical friction*. The simplest picture of dynamical friction is obtained by considering the following. If a massive particle of mass m moves through a statistically homogeneous distribution of massive particles, the gravitational force on this particle vanishes due to homogeneity. But since the particle itself has a mass, it will attract other massive particles and thus cause the distribution to become inhomogeneous. As the particle moves, the surrounding 'background' particles will react to its gravitational field and slowly start moving towards the direction of the particle trajectory. Due to the inertia of matter, the resulting density inhomogeneity will be such that an overdensity of mass will be established along the track of the particle, where the density will be higher on the side opposite to the direction of motion (thus, behind the particle) than in the forward direction (see Fig. 6.16). By this process, a gravitational field will form that causes an acceleration of the particle against the direction of motion, so that the particle will be slowed down. Because this 'polarization' of the medium is caused by the gravity of the particle, which is proportional to its mass, the deceleration will also be proportional to m. Furthermore, a fast-moving particle will cause less polarization in the medium than a slowly moving one because each mass element in the medium is experiencing the gravitational attraction of the particle for a shorter time, thus the medium becomes less polarized. In addition, the particle is on average farther away from the density accumulation on its backward track, and thus will experience a smaller acceleration if it is faster. Combining these arguments, one obtains for the dependence of this dynamical friction

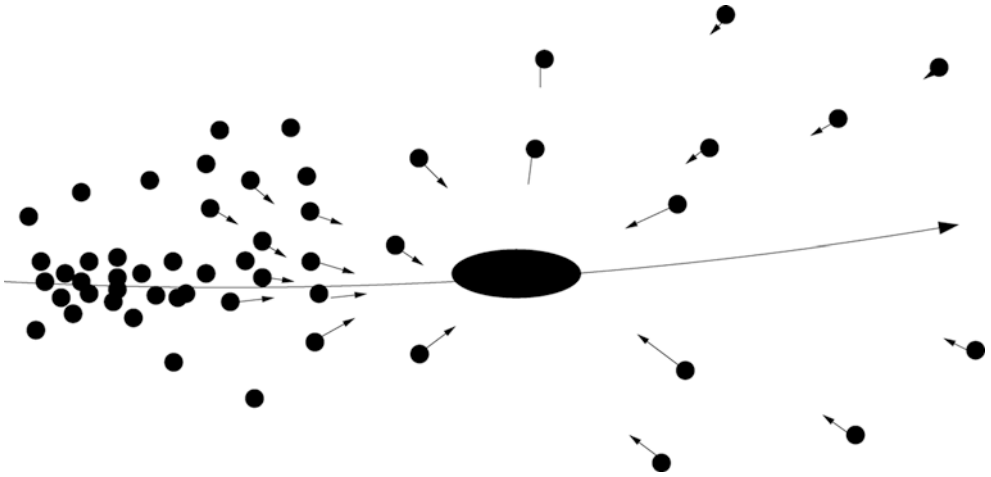


Fig. 6.16 The principle of dynamical friction. The gravitational field of a massive particle (here indicated by the *large symbol*) accelerates the surrounding matter towards its track. Through this, an overdensity establishes on the backward side of its orbit, the gravitational force of which decelerates the particle

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} \propto -\frac{m\,\rho\,\boldsymbol{v}}{|\boldsymbol{v}|^3}\ , \tag{6.30}$$

where ρ is the mass density in the medium. Applied to clusters of galaxies, this means that the most massive galaxies will experience the strongest dynamical friction, so that they are subject to a significant deceleration through which they move deeper into the potential well. The most massive cluster galaxies should therefore be concentrated around the cluster center, so that a spatial separation of galaxy populations with respect to their masses occurs (mass segregation). If dynamical friction acts over a sufficiently long time, the massive cluster galaxies in the center may merge into a single one. This is one possible explanation for the formation of cD galaxies. Furthermore, as the most massive (and thus presumably also the most luminous) galaxies are affected strongest by dynamical friction, and are thus the prime candidates for merging with the central galaxy, this may explain the observed gap of $\sim$ 2 mag between the brightest and second brightest cluster galaxy.

Dynamical friction also plays an important role in other dynamical processes in astrophysics. For example, the Magellanic Clouds experience dynamical friction on their orbit around the Milky Way and thereby lose kinetic energy. Consequently, their orbit will become smaller over the course of time and, in a distant future, these two satellite galaxies will merge with our Galaxy. In fact, dynamical friction is of vital importance in galaxy merger processes which occur in the evolution of the galaxy population, a subject we will return to in Chap. 10.

Compact groups have a lifetime which is much shorter than the age of the Universe. The dynamical time-scale is $t_{\mathrm{dyn}} \sim R/\sigma_v \sim 0.02\,H_0^{-1}$, thus small compared to $t_0 \sim H_0^{-1}$. By dynamical friction, galaxies in groups lose kinetic (orbital) energy and move closer to the dynamical center where interactions and mergers with other group galaxies take place, as also seen by the high fraction of member galaxies with morphological signs of interactions. Since the lifetime of compact groups is shorter than the age of the Universe, they must have formed not too long ago. If we do not happen to live in a special epoch of cosmic history, such groups must therefore still be forming today. From dynamical studies, one estimates that—as in clusters—the total mass of groups is significantly larger than the sum of the mass visible in galaxies; a typical mass-to-light ratio is $M/L \sim 50h$ (in Solar units), which is comparable to that of the Local Group.

6.3.4 Intergalactic stars in clusters of galaxies

The space between the galaxies in a cluster is filled with hot gas, as visible from X-ray observations. Besides this hot gas there are also stars in between the galaxies. This *intracluster light (ICL)* is the most recently discovered component in clusters. The detection of such an intergalactic stellar population comes as a surprise at first sight, because our understanding of star formation implies that they can only form in the dense centers of molecular clouds. Hence, one expects that stars cannot form in intergalactic space. This is not necessarily implied by the presence of intergalactic stars, however, since they can also be stripped from galaxies in the course of gravitational interactions between galaxies in the cluster, or the stripping of stars from the outer parts of galaxies in the tidal gravitational field (in the central region) of the cluster, and so form an intergalactic population. The fate of these stars is thus somewhat similar to that of the interstellar medium, which is metal-enriched by the processes of stellar evolution in galaxies before it is removed from these galaxies and becomes part of the intergalactic medium in clusters; otherwise, the substantial metallicity of the ICM could not be explained.[5] This interpretation is strengthened by the fact that a diffuse optical light component is also

[5] Of course, stars in galaxies are not subject to ram pressure stripping as is their gas.

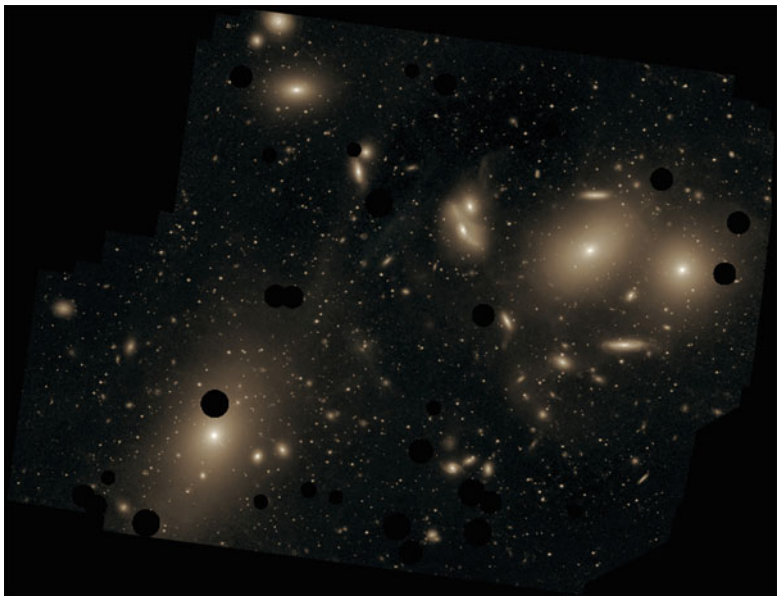


Fig. 6.17 This image shows the central part of the Virgo cluster, with its central galaxy M87 located in the *lower left corner*. The size of the image is about 1.5°; it shows the diffuse light in the cluster between the cluster galaxies. *Dark spots* indicate regions that were masked, e.g. because of bright foreground stars. The brightest parts before saturation have a surface brightness of $\mu_V \sim 26.5\,\mathrm{mag/arcsec^2}$, the faintest visible features have $\mu_V \sim 28.5\,\mathrm{mag/arcsec^2}$. Credit: Chris Mihos (Case Western Reserve University)/European Southern Observatory

seen in (compact) galaxy groups where the strength of tidal interactions is stronger than in clusters.

Observation of intracluster light. The observation of diffuse optical light in clusters of galaxies and, related to this, the detection of the intracluster stellar population, is extremely difficult. Although first indications were already found with photographic plate measurements, the surface brightness of this cluster component is so low that even with CCD detectors the observation is extraordinarily challenging. To quantify this, we note that the surface brightness of this diffuse light component is about $30\,\mathrm{mag\,arcsec^{-2}}$ at a distance of several hundred kpc from the cluster center. This value needs to be compared with the brightness of the night sky, which is about $21\,\mathrm{mag\,arcsec^{-2}}$ in the V-band. One therefore needs to correct for the effects of the night sky to better than a tenth of a percent for the intergalactic stellar component to become visible in a cluster. Furthermore, cluster galaxies and objects in the foreground and background need to be masked out in the images, in order to measure the radial profile of this diffuse component. This is possible only up to a certain limiting magnitude, of course, up to which individual objects can be identified. The existence of weaker sources has to be accounted for with statistical methods, which in turn use the luminosity function of galaxies. An example of this ICL is shown in Fig. 6.17.

The identification of this diffuse optical light as truly intergalactic origin is hampered by the fact that many clusters host a central cD galaxy. As we mentioned in Sect. 3.2.2 (see also Fig. 3.11), such galaxies have an extended brightness profile with a surface brightness substantially higher than the extrapolation of a de Vaucouleurs profile at large radii. Thus the natural question arises whether the diffuse component is just part of the cD envelope or a separate entity. If the ICL belongs to the central galaxy, it should be gravitationally bound to it; otherwise, it is a genuine intracluster component. This issue can be investigated by kinematical observations of individual stars in the ICL. One finds that the velocity dispersion of the stars in the ICL strongly increases away from the central galaxy, suggesting that they are unbound to it. One of the best individual tracers of the ICL are planetary nebulae which are formed in the final stages of Solar-mass stars. Since they emit a large fraction of their energy in a single emission line, they are ideal targets for spectroscopy. The kinematic study of intergalactic planetary nebulae show that they are not part of the central cluster galaxy. Related studies were carried out also with red giant stars and globular clusters. Also, Type Ia supernovae were found in clusters, but outside any cluster galaxy.

The diffuse light component was investigated in a statistical superposition of the images of several galaxy clusters. Statistical fluctuations in the sky background and uncertainties in the flatfield[6] determination are in this case averaged out. In this analysis an $r^{-1/4}$-law is found for the light distribution in the inner region of clusters, i.e., the (de Vaucouleurs) brightness profile of the central galaxy is measured (see Fig. 6.18). For radii larger than about $\sim 50\,\mathrm{kpc}$, the brightness profile exceeds the extrapolation of the de Vaucouleurs profile, and is detected out to very large distances from the cluster center.

[6]The flatfield of an image (or, more precisely, of the system consisting of telescope, filter, and detector) is defined as the image of a uniformly illuminated field, so that in the ideal case each pixel of the detector produces the same output signal. This is not the case in reality, however, as the sensitivity differs for individual pixels. For this reason, the flatfield measures the sensitivity distribution of the pixels, which is then accounted for in the image analysis.

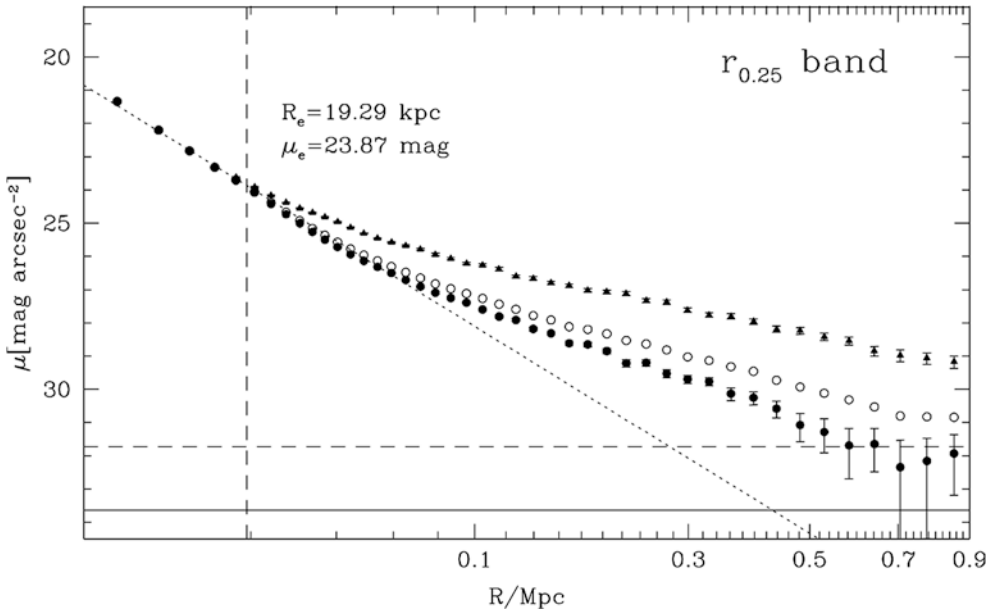


Fig. 6.18 The average light profile of 683 galaxy clusters from the maxBCG catalog (Sect. 6.2.4) with $0.2 \leq z \leq 0.3$. The *upper row of symbols* (*triangles*) show the total light profile, the *open circles* show the light profile after masking out identified galaxies (except the central BCG). Since these are detectable only up to some brightness limit, one can correct for this incompleteness by statistically subtracting fainter galaxies, based on an assumed luminosity function. This corrected brightness profile is shown as *filled circles* with error bars, which thus indicates the sum of the BCG light profile plus the intracluster light. In the inner part, this follows a de Vaucouleurs profile, indicated by the *dotted line*, but beyond ~ 60 kpc there is clear excess of light, detected out to almost 1 Mpc. Source: S. Zibetti et al. 2005, *Intergalactic stars in $z \sim 0.25$ galaxy clusters: systematic properties from stacking of Sloan Digital Sky Survey imaging data*, MNRAS 358, 949, p. 957, Fig. 5. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

The ICL is best studied in nearby groups and clusters; only for them can individual objects be detected to a low luminosity level. Furthermore, the cosmological surface brightness dimming $\propto (1+z)^{-4}$ (see problem 6.2) renders the detection of ICL increasingly more difficult when turning to higher redshift. Nevertheless, an ICL component was discovered out to redshift $z \sim 0.5$, perhaps even as far out as $z \sim 1$. The diffuse cluster component accounts for about 10 % of the total optical light in a cluster; in some clusters this fraction can be even higher. Therefore, models of galaxy evolution in clusters should provide an explanation for these observations.

6.4 Hot gas in galaxy clusters

One of the most important discoveries of the UHURU X-ray satellite, launched in 1970, was the detection of X-ray radiation from massive clusters of galaxies. With the later Einstein X-ray satellite and more recently ROSAT, X-ray emission was also detected from lower-mass clusters and groups. Three examples for the X-ray emission of galaxy clusters are displayed in Figs. 6.19, 6.20, and 6.21. Figure 6.19 shows the Coma cluster of galaxies, observed with two different X-ray observatories. Although Coma was considered to be a fully relaxed cluster, distinct substructure is visible in its X-ray radiation. The cluster RXJ 1347−1145 (Fig. 6.20) is the most X-ray luminous cluster in the ROSAT All-Sky Survey (Sect. 6.4.5). A large mass estimate of this cluster also follows from the analysis of the gravitationally lensed arcs (see Sect. 6.6) that are visible in Fig. 6.20; the cover of this book shows a more recent image of this cluster, taken with the ACS camera on-board HST, where a large number of arcs can be readily detected. Finally, Fig. 6.21 shows a superposition of the X-ray emission and an optical image of the cluster MS 1054−03, which is situated at $z = 0.83$ and was for many years the highest redshift cluster known.

6.4.1 General properties of the X-ray radiation

Clusters of galaxies are the brightest extragalactic X-ray sources besides AGNs. If an X-ray telescope is pointed away from the Galactic disk, about 85 % of the detected sources are AGNs, the remaining ~ 15 % are clusters. In contrast to AGNs, for which the X-ray emission is essentially point-like, the X-ray emission of clusters is extended. Their characteristic luminosity is $L_X \sim 10^{43}$ erg/s up to $\sim 10^{45}$ erg/s for the most massive systems. The fact that this X-ray emission is spatially extended implies that it does not originate from individual galaxies. The spatial region from which we can

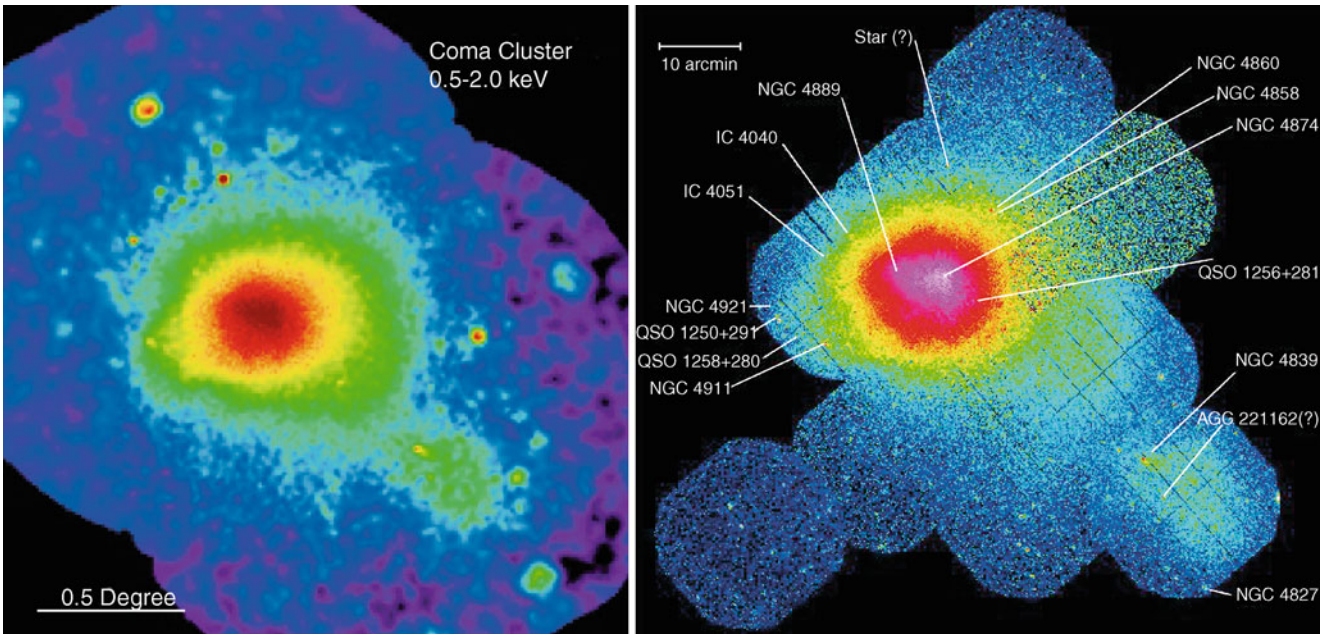


Fig. 6.19 X-ray images of the Coma cluster, taken with the ROSAT-PSPC (*left*) and XMM-EPIC (*right*). The image size in the *left panel* is $2.7° \times 2.5°$, much larger than the optical image shown in Fig. 1.17. A remarkable feature is the secondary maximum in the X-ray emission at the lower right of the cluster center which shows that even Coma, long considered to be a regular cluster, is not completely in an equilibrium state, but is dynamically evolving, presumably by the accretion of a galaxy group. Credit: *left*: S.L. Snowden, NASA, GSFC; *right*: U. Briel et al. 2001, *A mosaic of the Coma cluster of galaxies with XMM-Newton*, A&A 365, L60, p. L62, Fig. 1. ©ESO. Reproduced with permission

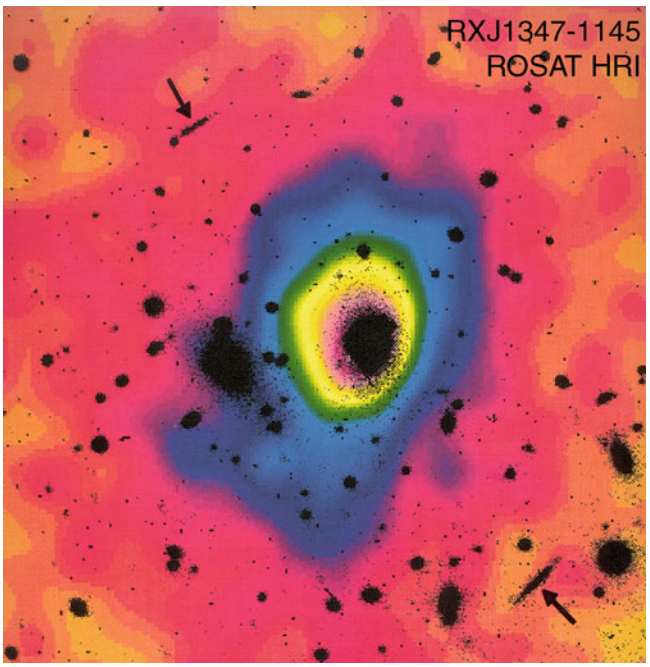


Fig. 6.20 RXJ 1347−1145 is one of the most luminous galaxy clusters in the X-ray domain. A color-coded ROSAT/HRI image of this cluster, which shows the distribution of the intergalactic gas, is superposed on an optical image of the cluster with redshift $z = 0.45$. The *two arrows* indicate giant arcs, images of background galaxies which are strongly distorted by the gravitational lens effect. Credit: Max-Planck-Institut für extraterrestrische Physik

detect this radiation can have a size of 1 Mpc or even larger. In accordance with the extended nature of the X-ray source, no variability of its X-ray flux has been detected.

Continuum radiation. The spectral energy distribution of the X-rays leads to the conclusion that the emission process is optically thin thermal bremsstrahlung (free-free radiation) from a hot gas which is collisionally ionized. This radiation is produced by the acceleration of electrons in the Coulomb field of protons and atomic nuclei. Since an accelerated electrically charged particle emits radiation, such scattering processes between electrons and protons in an ionized gas yields emission of photons. From the spectral properties of this radiation, the gas temperature in galaxy clusters can be determined, which is, for clusters with mass between $\sim 3 \times 10^{13} M_\odot$ and $\sim 10^{15} M_\odot$, in the range of 10^7 K to 10^8 K, or 1 to 10 keV, respectively.

The emissivity of bremsstrahlung is described by

$$\epsilon_\nu^{\mathrm{ff}} = \frac{32\pi Z^2 e^6 n_\mathrm{e} n_\mathrm{i}}{3 m_\mathrm{e} c^3} \sqrt{\frac{2\pi}{3 k_\mathrm{B} T m_\mathrm{e}}}\, \mathrm{e}^{-h_\mathrm{P}\nu/k_\mathrm{B}T}\, g_{\mathrm{ff}}(T,\nu)\ , \tag{6.31}$$

where e denotes the elementary charge, n_e and n_i the number density of electrons and ions, respectively, Ze the charge of the ions, and m_e the electron mass. The dimensionless func-

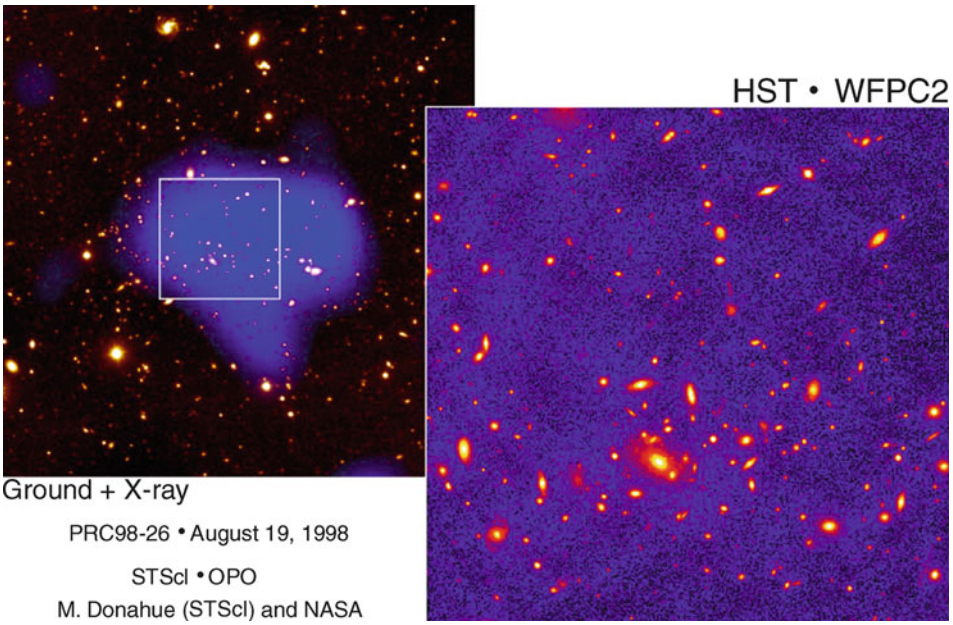


Fig. 6.21 The cluster of galaxies MS 1054−03 is, at $z = 0.83$, the highest-redshift cluster in the Einstein Medium Sensitivity Survey, which was compiled from observations with the Einstein satellite (see Sect. 6.4.5). *On the right*, an HST image of the cluster is shown, while *on the left* is an optical image, obtained with the 2.2-m telescope of the University of Hawaii, superposed (in *blue*) with the X-ray emission of the cluster measured with the ROSAT-HRI. Credit: Megan Donahue/STSCI, Isabella Gioia/Univ. of Hawaii and NASA

tion $g_{\rm ff}$ is called Gaunt-factor; it is a quantum mechanical correction factor of order 1, or, more precisely,

$$g_{\rm ff} \approx \frac{3}{\sqrt{\pi}} \ln\left(\frac{9k_{\rm B}T}{4h_{\rm P}\nu}\right) .$$

Hence, the spectrum described by (6.31) is flat for $h_{\rm P}\nu \ll k_{\rm B}T$, and exponentially decreasing for $h_{\rm P}\nu \gtrsim k_{\rm B}T$, as is displayed in Fig. 6.22.

The temperature of the gas in massive clusters is typically $T \sim 5 \times 10^7\,{\rm K}$, or $k_{\rm B}T \sim 5\,{\rm keV}$—X-ray astronomers usually specify temperatures and frequencies in keV (see Appendix C). For a thermal plasma with Solar abundances, the total bremsstrahlung emission is

$$\epsilon^{\rm ff} = \int_0^\infty {\rm d}\nu \epsilon_\nu^{\rm ff} \approx 3.0 \times 10^{-27} \sqrt{\frac{T}{1\,{\rm K}}} \left(\frac{n_{\rm e}}{1\,{\rm cm}^{-3}}\right)^2 {\rm erg\,cm^{-3}s^{-1}} . \tag{6.32}$$

Line emission. The assumption that the X-ray emission originates from a hot, diffuse gas (intracluster medium, ICM) was confirmed by the discovery of line emission in the X-ray spectrum of clusters. One of the most prominent lines in massive clusters is located at energies just below 7 keV: it is the Lyman-α ("Kα") line of 25-fold ionized iron (thus, of an iron nucleus with only a single electron). Slightly less ionized iron has a strong transition at somewhat lower energies of $E \sim 6.4\,{\rm keV}$. Later, other lines were also discovered in the X-ray spectrum of clusters. As a rule, the hotter the gas is, thus the more completely ionized it is, the weaker the line emission. The X-ray emission of clusters with relatively low temperatures, $k_{\rm B}T \lesssim 2\,{\rm keV}$, is sometimes dominated by line emission from highly ionized atoms (C, N, O, Ne, Mg, Si, S, Ar, Ca, and a strong line complex of iron at $E \sim 1\,{\rm keV}$—see Fig. 6.22). The emissivity of a thermal plasma with Solar abundance and temperatures in the range $10^5\,{\rm K} \lesssim T \lesssim 4 \times 10^7\,{\rm K}$ can roughly be approximated by

$$\epsilon \approx 6.2 \times 10^{-19} \left(\frac{T}{1\,{\rm K}}\right)^{-0.6} \left(\frac{n_{\rm e}}{1\,{\rm cm}^{-3}}\right)^2 {\rm erg\,cm^{-3}\,s^{-1}} . \tag{6.33}$$

Equation (6.33) accounts for free-free emission as well as line emission. Compared to (6.32), one finds a different dependence on temperature: while the total emissivity for bremsstrahlung is $\propto T^{1/2}$, it increases again towards lower temperatures where the line emission becomes more important (see also Fig. 10.3 for the temperature dependence of the emissivity of a gas). It should be noted in particular that the emissivity depends quadratically on the density of the plasma, since both bremsstrahlung and the collisional excitation responsible for line emission are two-body processes. Thus in order to estimate the mass of the hot gas from its X-ray luminosity, the spatial distribution of the gas needs to be known. For example, if the gas in a cluster is locally inhomogeneous, the value of $\langle n_{\rm e}^2 \rangle$ which determines the X-ray emissivity may deviate significantly from $\langle n_{\rm e} \rangle^2$. As we will see later, clusters of galaxies satisfy a number of scaling relations, and one relation between the gas mass and the X-ray luminosity is found empirically, from which the gas mass can be estimated. One finds that the mass in the intracluster gas is about five to ten times larger than the mass of the stars in the galaxies, where this ratio slightly increases with increasing cluster mass.

Morphology of the X-ray emission. From the morphology of their X-ray emission, one can roughly distinguish between

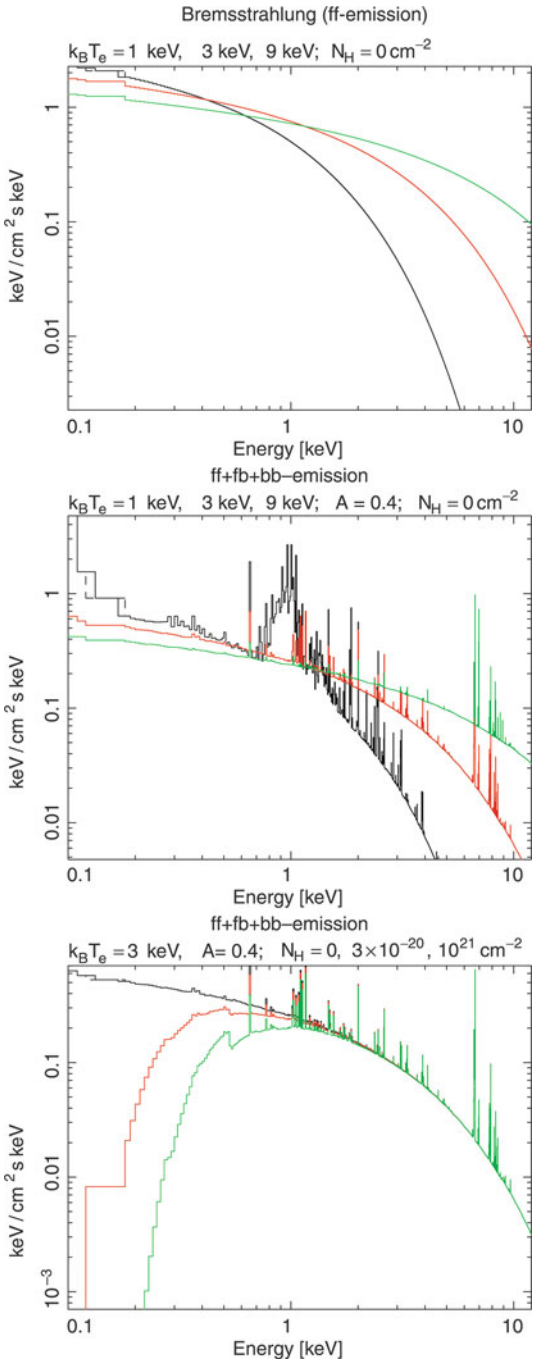


Fig. 6.22 X-ray emission of an optically thin hot plasma. In the *top panel*, the bremsstrahlung spectrum is shown, for three different gas temperatures; the radiation of hotter gas extends to higher photon energies, and above $E \sim k_B T$ the spectrum is exponentially cut off. In the central panel, atomic transitions and recombination radiation in the hot gas are also taken into account, where a metallicity of 40 % of the Solar value is assumed. These additional radiation mechanisms become more important towards lower T, as can be seen from the $T = 1$ keV curve. In the *bottom panel*, photo-electric absorption by the Galactic interstellar medium is included, with different column densities in hydrogen; for this gas, Solar metallicity is assumed. The Galactic absorption produces a cut-off in the spectrum towards lower energies. Credit: T. Reiprich, Argelander-Institut für Astronomie, Universität Bonn

regular and irregular clusters, as is also done in the classification of the galaxy distribution. In Fig. 6.23, X-ray surface brightness contours are superposed on optical images of four galaxy clusters or groups, covering a wide range of cluster mass and X-ray temperature. Regular clusters show a smooth brightness distribution, centered on the optical center of the cluster, and an outwardly decreasing surface brightness. In contrast, irregular clusters may have several brightness maxima, often centered on cluster galaxies or subgroups of cluster galaxies.

Typically, regular clusters have an X-ray luminosity L_X and temperature that smoothly increases with cluster mass. In contrast, irregular clusters at a given mass can be either hotter or cooler than regular clusters. Irregular clusters are the result of a recent merging event, and their temperature depends on the stage of the merging process. In the initial phases of the merger, the kinetic energy is not yet thermalized, and thus the gas remains at approximately the same temperature it had before the merging event. Later on, the gas is heated by shock fronts in which the kinetic energy is transformed into internal energy of the gas—i.e., heat. In this phase, the gas temperature can be higher than that of a regular cluster of the same mass. Finally, the cluster settles into an equilibrium state. Indications of past merger events can be seen in substructures of the X-ray emitting gas; even for the Coma cluster, which is frequently considered a typical example of a relaxed cluster, signs of previous merger events can be detected, as shown in Fig. 6.24.

The trend emerges that in clusters with a larger fraction of spiral galaxies, L_X and T are lower. Irregular clusters typically also have a lower central density of galaxies and gas compared to regular clusters. Clusters of galaxies with a dominant central galaxy often show a strong central peak in X-ray emission. The X-ray emission often deviates from axial symmetry, so that the assumption of clusters being roughly spherically symmetric is not well founded in these cases.

6.4.2 Models of the X-ray emission

Hydrostatic assumption. To draw conclusions about the properties of the intergalactic (intra-cluster) medium from the observed X-ray radiation and about the distribution of mass in the cluster, the gas distribution needs to be modeled. In fact, as we shall see below, assuming the geometry of the cluster (e.g., spherical symmetry), the radial dependence of the gas density can be directly reconstructed. For this, we first consider the speed of sound in the cluster gas,

$$c_s \approx \sqrt{\frac{P}{\rho_g}} = \sqrt{\frac{nk_BT}{\rho_g}} = \sqrt{\frac{k_BT}{\mu\, m_p}} \sim 1000\,\mathrm{km\,s^{-1}}\,,$$

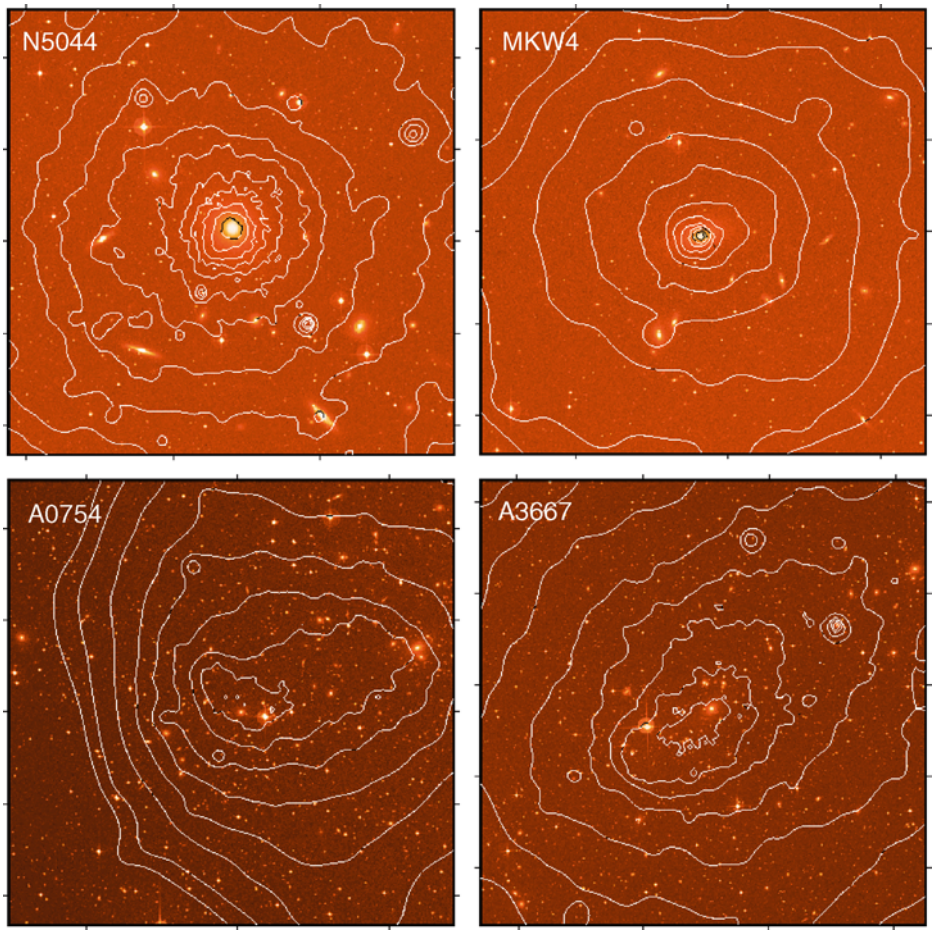


Fig. 6.23 Surface brightness contours of the X-ray emission for four different groups or clusters of galaxies. Each image is about 30′ on a side. *Upper left*: the galaxy group NGC 5044, at redshift $z = 0.009$, with an X-ray temperature of $T \approx 1.07\,\mathrm{keV}$ and a virial mass of $M_{200} \approx 0.32h^{-1} \times 10^{14} M_\odot$. *Upper right*: the group MKW4, at $z = 0.02$, with $T \approx 1.71\,\mathrm{keV}$ and $M_{200} \approx 0.5h^{-1} \times 10^{14} M_\odot$. *Lower left*: the cluster of galaxies A 0754, at $z = 0.053$, with $T \approx 9.5\,\mathrm{keV}$ and $M_{200} \approx 13.1h^{-1} \times 10^{14} M_\odot$. *Lower right*: the cluster of galaxies A 3667, at $z = 0.056$, with $T \approx 7.0\,\mathrm{keV}$ and $M_{200} \approx 5.6h^{-1} \times 10^{14} M_\odot$. The X-ray data were obtained by ROSAT, and the optical images were taken from the Digitized Sky Survey. These clusters are part of the HIFLUGCS survey, which we will discuss more thoroughly in Sect. 6.4.5. Credit: T. Reiprich, Argelander-Institut für Astronomie, Universität Bonn

where P denotes the gas pressure, ρ_g the gas density, and n the number density of gas particles. Then, the *average molecular mass* is defined as the average mass of a gas particle in units of the proton mass,

$$\mu := \frac{\langle m \rangle}{m_\mathrm{p}} \,, \tag{6.34}$$

so that $\rho_\mathrm{g} = n \langle m \rangle = n\mu m_\mathrm{p}$. For a gas of fully ionized hydrogen, one gets $\mu = 1/2$ because in this case one has one proton and one electron per ∼proton mass. The cluster gas also contains helium and heavier elements; accounting for them yields $\mu \sim 0.60$. The sound-crossing time for the cluster is

$$t_\mathrm{sc} \approx \frac{2R_\mathrm{A}}{c_\mathrm{s}} \sim 7 \times 10^8\,\mathrm{yr} \,,$$

and is thus, for a cluster with $T \sim 10^8\,\mathrm{K}$, significantly shorter than the lifetime of the cluster, which can be approximated roughly by the age of the Universe. Since the sound-crossing time defines the time-scale on which deviations from the pressure equilibrium are evened out, the gas can be in hydrostatic equilibrium, provided that the last major merging event happened longer ago than the sound-crossing time. In this case, the equation

$$\boxed{\nabla P = -\rho_\mathrm{g} \nabla \Phi} \tag{6.35}$$

applies, with Φ denoting the gravitational potential. Equation (6.35) describes how the gravitational force is balanced by the pressure force. In the spherically symmetric case in which all quantities depend only on the radius r, we obtain

$$\frac{1}{\rho_\mathrm{g}} \frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{\mathrm{d}\Phi}{\mathrm{d}r} = -\frac{GM(r)}{r^2} \,, \tag{6.36}$$

where $M(r)$ is the mass enclosed within radius r. Here, $M(r)$ is the total enclosed mass, i.e., not just the gas mass, because the potential Φ is determined by the total mass. Note the similarity of this equation with (6.9), except that there P described the dynamical pressure of the galaxies or dark matter particles. By inserting $P = nk_\mathrm{B}T = \rho_\mathrm{g} k_\mathrm{B} T/(\mu m_\mathrm{p})$ into (6.36), we obtain

$$\boxed{M(r) = -\frac{k_\mathrm{B} T r^2}{G\mu m_\mathrm{p}} \left(\frac{\mathrm{d} \ln \rho_\mathrm{g}}{\mathrm{d}r} + \frac{\mathrm{d} \ln T}{\mathrm{d}r} \right)} \,. \tag{6.37}$$

Fig. 6.24 In *white*, on optical image of the Coma cluster is shown. Superposed on this SDSS image is the X-ray emission, shown in *pink*, where the smooth component of the X-ray image was subtracted to highlight the filamentary structure of the hot gas. These filaments are most likely due to a past merger events, when smaller groups fall into the main cluster; their gas was stripped by ram-pressure during infall, leaving trails of gas. The sidelength of the image is 23′, corresponding to about 600 kpc. Credit: X-ray: NASA/CXC/MPE/J. Sanders et al., Optical: SDSS

This equation is of central importance for the X-ray astronomy of galaxy clusters because it shows that we can derive the mass profile $M(r)$ from the radial profiles of ρ_g and T. Thus, if one can measure the density and temperature profiles, the mass of the cluster, and hence the total density, can be determined as a function of radius.

However, these measurements are not without difficulties. $\rho_\mathrm{g}(r)$ and $T(r)$ need to be determined from the X-ray luminosity and the spectral temperature, using the bremsstrahlung emissivity (6.31). Obviously, they can be observed only in projection in the form of the surface brightness

$$I_\nu(R) = 2\int_R^\infty \mathrm{d}r\ \frac{\epsilon_\nu(r)\, r}{\sqrt{r^2 - R^2}}\ , \tag{6.38}$$

from which the emissivity, and thus density and temperature, need to be derived by de-projection, i.e., the inversion of (6.38) to obtain $\epsilon_\nu(r)$ in terms of $I_\nu(R)$. Furthermore, the angular and energy resolution of X-ray telescopes prior to XMM-Newton and Chandra were not high enough to measure both $\rho_\mathrm{g}(r)$ and $T(r)$ with sufficient accuracy, except for the nearest clusters. For this reason, the mass determination was often performed by employing additional, simplifying assumptions.

Isothermal gas distribution. From the radial profile of $I(R)$, $\epsilon(r)$ can be derived by inversion of (6.38). Since the spectral bremsstrahlung emissivity depends only weakly on T for $h_\mathrm{P}\nu \ll k_\mathrm{B}T$—see (6.31)—the radial profile of the gas density ρ_g can be derived from $\epsilon(r)$. The X-ray satellite ROSAT was sensitive to radiation of $0.1\,\mathrm{keV} \lesssim E \lesssim 2.4\,\mathrm{keV}$, so that the X-ray photons detected by it are typically from the regime where $h_\mathrm{P}\nu \ll k_\mathrm{B}T$.

Assuming that the gas temperature is spatially constant, $T(r) = T_\mathrm{g}$, (6.37) simplifies, and the mass profile of the cluster can be determined from the density profile of the gas.

The β-model. A commonly used method consists of fitting the X-ray data by a so-called β-model. This model is based on the assumption that the density profile of the total matter (dark and luminous) is described by an isothermal distribution, i.e., it is assumed that the temperature of the gas is independent of radius, and at the same time that the mass distribution in the cluster is described by the isothermal model that was discussed in Sect. 6.3.1. With (6.9) and (6.12), we then obtain for the total density $\rho(r)$

$$\frac{\mathrm{d}\ln\rho}{\mathrm{d}r} = -\frac{1}{\sigma_v^2}\,\frac{GM(r)}{r^2}\ . \tag{6.39}$$

On the other hand, in the isothermal case (6.37) reduces to

$$\frac{\mathrm{d}\ln\rho_\mathrm{g}}{\mathrm{d}r} = -\frac{\mu m_\mathrm{p}}{k_\mathrm{B}T_\mathrm{g}}\,\frac{GM(r)}{r^2}\ . \tag{6.40}$$

The comparison of (6.39) and (6.40) then shows that $\mathrm{d}\ln\rho_\mathrm{g}/\mathrm{d}r \propto \mathrm{d}\ln\rho/\mathrm{d}r$, or

$$\rho_\mathrm{g}(r) \propto [\rho(r)]^\beta \quad \text{with} \quad \boxed{\beta := \frac{\mu m_\mathrm{p}\sigma_v^2}{k_\mathrm{B}T_\mathrm{g}}} \tag{6.41}$$

must apply; thus the gas density follows the total density to some power. Here, the index β depends on the ratio of the dynamical temperature, measured by σ_v, and the gas temperature. Now, using the King approximation for an isothermal mass distribution—see (6.17)—as a model for the mass distribution, we obtain

$$\boxed{\rho_\mathrm{g}(r) = \rho_{\mathrm{g}0}\left[1 + \left(\frac{r}{r_\mathrm{c}}\right)^2\right]^{-3\beta/2}}\ , \tag{6.42}$$

where $\rho_{\mathrm{g}0}$ is the central gas density. The brightness profile of the X-ray emission in this model is then, according to (6.38),

$$\boxed{I(R) \propto \left[1 + \left(\frac{R}{r_\mathrm{c}}\right)^2\right]^{-3\beta+1/2}}\ . \tag{6.43}$$

The X-ray emission of many clusters is well described by this profile,[7] yielding values for r_c of 0.1 to $0.3h^{-1}$ Mpc and a value for the index $\beta = \beta_{\rm fit} \approx 0.65$. Alternatively, β can be measured, with the definition given in (6.41), from the gas temperature T_g and the velocity dispersion of the galaxies σ_v, which yields typical values of $\beta = \beta_{\rm spec} \approx 1$. Such a value would also be expected if the mass and gas distributions were both isothermal. In this case, they should have the same temperature, which was presumably determined by the formation of the cluster.

The β-discrepancy. The fact that the two values for β determined above differ from each other (the so-called β-discrepancy) is almost certainly due to the fact that the β-model is too simple. We can see, e.g., from Fig. 6.25 that the gas distribution, at least in the inner part of clusters, does not follow a smooth distribution, nor is its temperature constant. The latter is also reflected by the fact that the measured values for $\beta_{\rm fit}$ often depend on the angular range over which the brightness profile is fitted: the larger this range, the larger $\beta_{\rm fit}$ becomes, and thus the smaller the discrepancy. This behavior can be understood if the central region of the clusters have a lower temperature than at larger radii. Furthermore, temperature measurements of clusters are often not very accurate because it is the emission-weighted temperature which is measured, which is, due to the quadratic dependence of the emissivity on ρ_g, dominated by the regions with the highest gas density. The fact that the innermost regions of clusters where the gas density is highest tend to have a temperature below the bulk temperature of the cluster (see Fig. 6.26) may lead to an underestimation of 'the' cluster temperature. In addition, the near independence of the spectral form of $\epsilon_\nu^{\rm ff}$ from T for $h_{\rm P}\nu \ll k_{\rm B}T$ renders the measurement of T difficult. Chandra and XMM-Newton can measure the X-ray emission at energies of up to $E \lesssim 10\,$keV, which resulted in considerably improved temperature determinations.

Such investigations have revealed that the temperature behavior shown in Fig. 6.26 is typical for many clusters: The temperature decreases towards the center and towards the edge, while it is rather constant over a larger range at intermediate radii. Many clusters are found, however, in which the temperature distribution is by no means radially symmetric, but shows distinct substructure. Finally, as another possible explanation for the β-discrepancy, it should be mentioned that the velocity distribution of those galaxies from which σ_v is measured may be anisotropic.

[7] We point out that the pair of (6.42) and (6.43) is valid independently of the validity of the assumptions from which (6.42) was obtained. If the observed X-ray emission is very well described by (6.43), the gas density profile (6.42) can be obtained from it, independently of the validity of the assumptions made before.

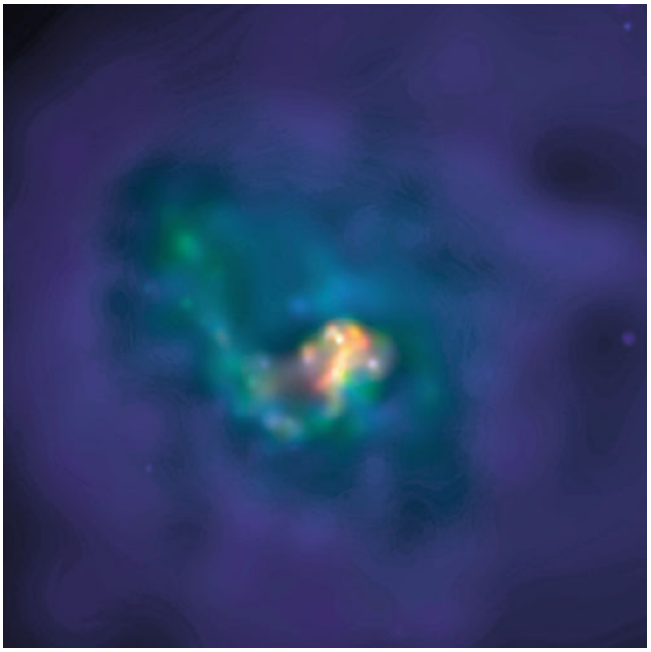


Fig. 6.25 Chandra image of the Centaurus cluster; the size of the field is $3' \times 3'$. Owing to the excellent angular resolution of the Chandra satellite, the complexity of the morphology in the X-ray emission of clusters can be analyzed. *Colors* indicate photon energies, from low to high in *red*, *yellow*, *green*, and *blue*. The gas in the center of the cluster is significantly cooler than that at larger radii. Credit: NASA/IoA/J. Sanders & A. Fabian

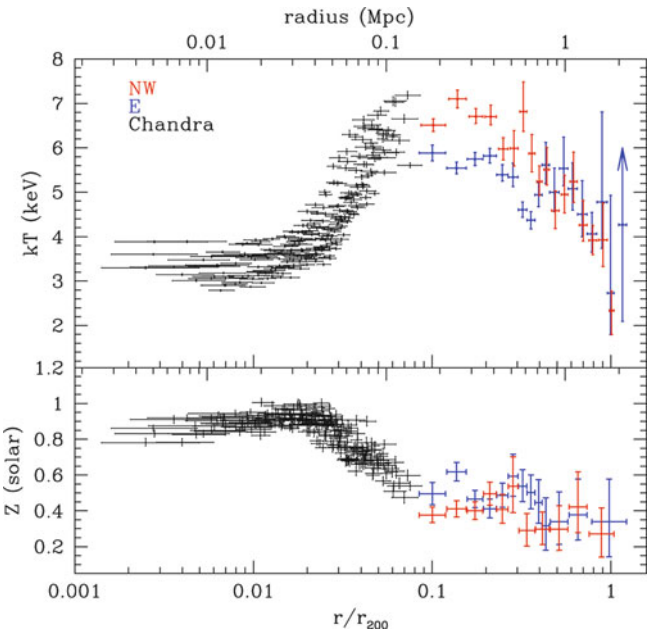


Fig. 6.26 Temperature (*top*) and metallicity (*bottom*) profile of the Perseus cluster, measured along two different directions from the cluster center outwards. *Black symbols* show measurements in the inner region, obtained with Chandra. The *colored points* are obtained from Suzaku data, with each color corresponding to one radial direction. Source: A. Simionescu et al. 2011, *Baryons at the Edge of the X-ray-Brightest Galaxy Cluster*, Science 331, 1576, Fig. 2. Reprinted with permission from AAAS

Besides all the uncertainty as to the validity of the β-model, we also need to mention that numerical simulations of galaxy clusters, which take dark matter and gas into account, have repeatedly come to the conclusion that the mass determination of clusters, utilizing the β-model, should achieve an accuracy of better than $\sim 20\,\%$, although different gas dynamical simulations have arrived at distinctly different results.

Dark matter in clusters from X-ray observations. Based on measurements of their X-ray emission, a mass estimate can be performed for galaxy clusters. It is found, in agreement with the dynamical method, that clusters contain much more mass than is visible in galaxies. The total mass of the intergalactic medium is clearly too low to account for the missing mass; its gas mass is only $\sim 15\,\%$ of the total mass of a cluster.

Only $\sim 3\,\%$ of the mass in clusters of galaxies is due to stars in galaxies, and about 15% is contributed by the intergalactic gas, whereas the remaining $\sim 80\,\%$ consists of dark matter which therefore dominates the mass of the clusters.

6.4.3 Cooling "flows"

Expected consequences of gas cooling. In examining the intergalactic medium, we assumed hydrostatic equilibrium, but we disregarded the fact that the gas cools by its emission and by that loses internal energy. For this reason, once established, a hydrostatic equilibrium in principle cannot be maintained over arbitrarily long times. To decide whether this gas cooling is important for the dynamics of the system, the cooling time-scale needs to be considered. This cooling time turns out to be very long,

$$t_{\rm cool} := \frac{u}{\epsilon^{\rm ff}} \approx 8.5 \times 10^{10}\,{\rm yr} \left(\frac{n_{\rm e}}{10^{-3}\,{\rm cm}^{-3}}\right)^{-1} \left(\frac{T_{\rm g}}{10^8\,{\rm K}}\right)^{1/2} , \tag{6.44}$$

where $u = (3/2) n k_{\rm B} T_{\rm g}$ is the energy density of the gas and $n_{\rm e}$ the electron density. Hence, the cooling time is longer than the Hubble time nearly everywhere in the cluster, which allows a hydro*static* equilibrium to be established to a very good approximation. In the centers of clusters, however, the density may be sufficiently large to yield $t_{\rm cool} \lesssim t_0 \sim H_0^{-1}$. Here, the gas can cool quite efficiently, by which its pressure decreases. This then implies that, at least close to the center, the hydrostatic equilibrium can no longer be maintained. To re-establish pressure equilibrium, gas needs to flow inwards and is thus compressed. Hence, an inward-directed mass flow should establish itself. The corresponding density increase will further accelerate the cooling process. Since, in addition, the emissivity (6.33) of a relatively cool gas increases with decreasing temperature, this process should then very quickly lead to a strong compression and cooling of the gas in the centers of dense clusters. It is a process which, once started, will accelerate and quickly lead to the cooling down to very low temperatures. In parallel to this increase in density, the X-ray emission should strongly increase, because $\epsilon^{\rm ff} \propto n_{\rm e}^2$. As a result of this process, a radial density and temperature distribution should be established with a nearly unchanged pressure distribution. In Fig. 6.25, the cooler gas in the center of the Centaurus cluster is clearly visible.

Some predictions of this so-called *cooling flow* model have indeed been verified observationally. In the centers of many massive clusters, one observes a sharp central peak in the surface brightness $I(R)$. However, we need to stress that, as yet, no inwards *flows* have been measured. Such a measurement would be very difficult, though, due to the small expected velocities. The amount of cooling gas can be considerable. We can estimate the mass rate $\dot{M}$ at which the gas should cool and flow inwards due to this cooling. The internal energy U of the gas is related to its mass M by $U = M\,u/\rho$, with u as given above. The loss of this energy due to cooling is the luminosity, $L = \dot{U}$, so that

$$\dot{M} \sim \frac{L\,\rho}{u} \sim \frac{L \mu m_{\rm p}}{k_{\rm B} T_{\rm g}} . \tag{6.45}$$

For some massive clusters, this estimate lead to cooling rates of tens or even hundreds of Solar masses per year. However, after spectroscopic observations by XMM-Newton became available, we have learned that these very high cooling rates implied by the models were *significantly* overestimated.

The fate of the cooling gas. The gas cooling in this way will accumulate in the center of the cluster, but despite the expected high mass of cold gas, no clear evidence has been found for it. In clusters harboring a cD galaxy, the cooled gas may, over a Hubble time, contribute a considerable fraction of the mass of this galaxy. Hence, the question arises whether cD galaxies may have formed by accretion in cooling flows. In this scenario, the gas would be transformed into stars in the cD galaxy. However, the star-formation rate in these central galaxies is much lower than the rate by which cluster gas cools, according to the 'old' cooling flow models sketched above.

The absence of massive cooling flows. The sensitivity and spectral resolution achieved with XMM-Newton have strongly modified our view of cooling flows. In the standard model of cooling flows, the gas cools from the cluster temperature down to temperatures significantly below 1 keV.

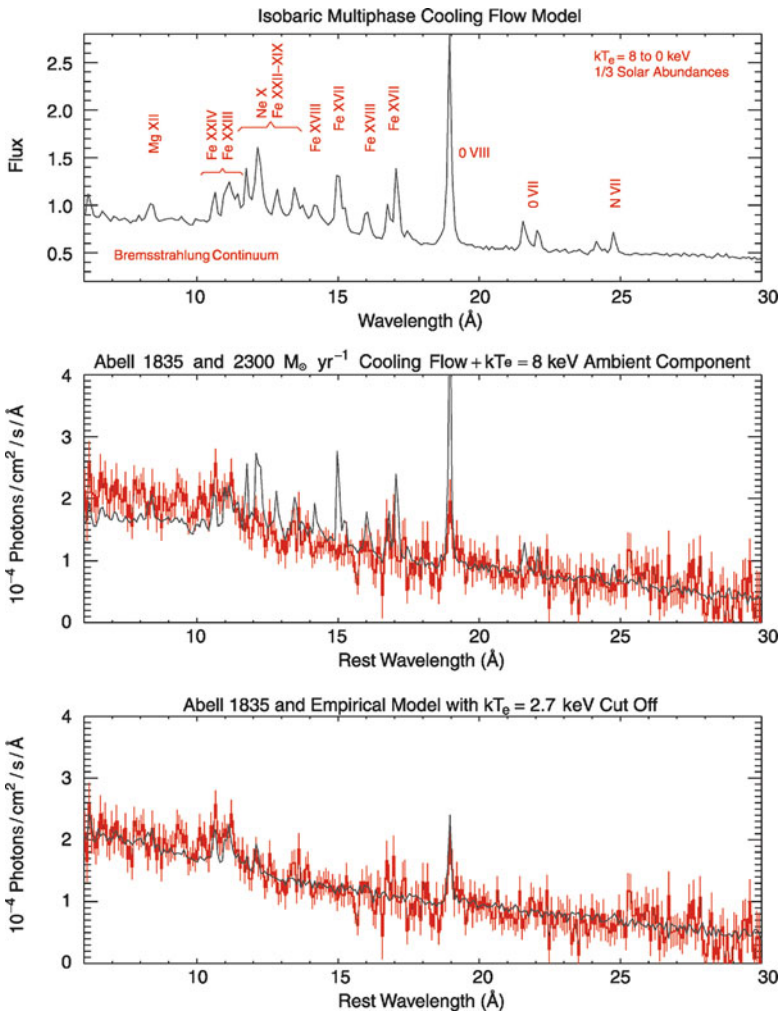


Fig. 6.27 In the *top panel*, a model spectrum of a cooling flow is shown, in which the gas cools down from 8 keV to $T_g = 0$. The strong lines of FeXVII can be seen. In the *central panel*, the spectrum of Abell 1835 is superposed on the model spectrum; clear discrepancies are visible, especially the absence of strong emission lines from FeXVII. If the gas is not allowed to cool down to temperatures below 3 keV (*bottom panel*), the agreement with observation improves visibly. Source: J.R. Peterson et al. 2003, *High Resolution X-ray Spectroscopic Constraints on Cooling-Flow Models*, astro-ph/0310008, Fig. 2. Reproduced by permission of the author

In this process many atomic lines are emitted, produced by various ionization stages, e.g., of iron, which strongly depend on temperature. Figure 6.27 (top panel) shows the expected spectrum of a cooling flow in which the gas cools down from the cluster temperature of $T_g \approx 8$ keV to essentially $T_g = 0$, where a chemical composition of 1/3 Solar abundance is assumed. In the central panel, this theoretical spectrum is compared with the spectrum of the cluster Abell 1835, where very distinct discrepancies become visible. In the bottom panel, the model was modified such that the gas cools down only to $T_g = 3$ keV; this model clearly matches the observed spectrum better.

Hence, cooler gas in the inner regions of clusters is directly detected spectroscopically. However, the temperature measurements from X-ray spectroscopy are significantly different from the prediction of the cooling flow model according to which drastic cooling should take place in the gas, because the process of compression and cooling will accelerate for ever decreasing T_g. Therefore, one expects to find gas at all temperatures lower than the temperature of the cluster. But this seems not to be the case: whereas the central temperature can be considerably smaller than that at larger radii (see, e.g., Fig. 6.26), no gas seems to be present at very small temperatures, although the cooling flow model predicts the existence of such gas. A minimum temperature seems to exist, below which the gas cannot cool, or the amount of gas that cools to $T_g \sim 0$ is considerably smaller than expected from the cooling flow model. This lower mass rate of gas that cools down completely would then also be compatible with the observed low star-formation rates in the central galaxies of clusters. In fact, a correlation between the cooling rate of gas as determined from XMM observations and the regions of star formation in clusters has been found (see Fig. 6.28). About 50 % of X-ray luminous clusters show an infrared excess of the BCG, indicating ongoing star formation; furthermore, whereas the BCG in clusters with long cooling time rarely show optical emission lines, most of those in cooling flow clusters do. All these are clear indications that a few percent of the cooling mass rate (6.45) as estimated from the cooling flow model indeed arrive at the central cluster galaxy.

There may be exceptional cases where a larger fraction of the cooling gas can reach the gravitational center of the cluster and cool down to low temperatures. Recently,

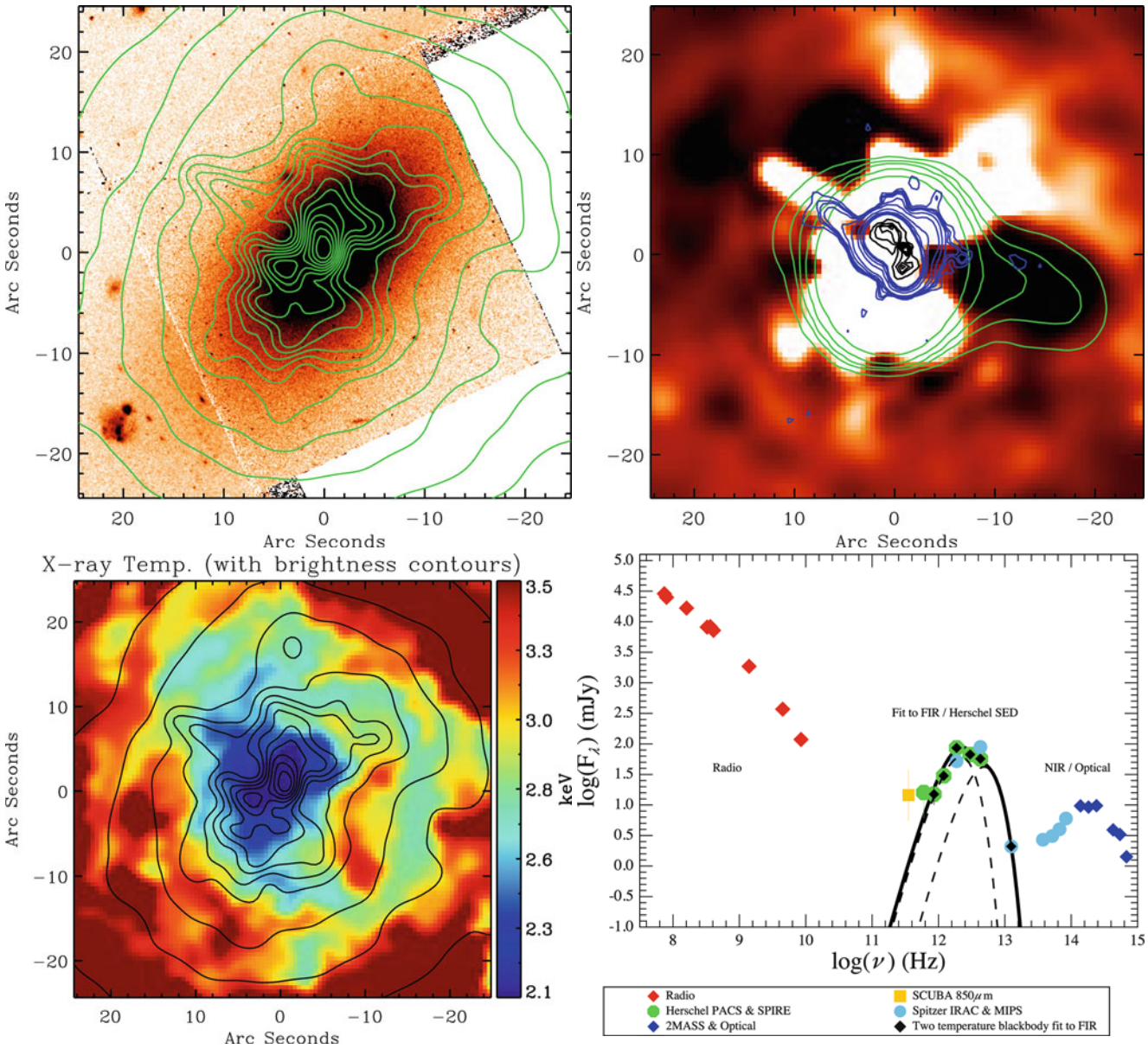


Fig. 6.28 The cool-core cluster Abell 2597 at $z = 0.082$. The *upper left panel* shows the Chandra X-ray map as green contours, superposed on an optical image of the brightest cluster galaxy (BCG) of this cluster. Note that the overall orientation of the X-ray emission follows that of the light in the BCG. *Upper right*: The color map is obtained from subtracting a smoothed version of the X-ray emission from the one shown in the *left panel*; this so-called 'unsharp mask' image highlights the small-scale brightness variations. Superposed on this are radio contours at three different frequencies, increasing *from green to blue to black*. Clearly, the BCG contains an active nucleus, and the radio jet apparently causes an X-ray cavity on the right of the galaxy center. The *lower left panel* shows the temperature map, as obtained from resolved X-ray spectroscopy, with the total X-ray emission superposed. The temperature in the inner part is markedly smaller than at larger radii, a clear sign of a cool core. Note also the small-scale structure in the temperature map. The spectral energy distribution of the core, shown at the *lower right*, yields a clear indication of ongoing star formation, seen from the far-IR radiation due to warm dust. Further indications for ongoing star formation is obtained from the UV-radiation, as well as emission lines of the object (not shown). Source: *Upper* and *lower right panels*: G.R. Tremblay et al. 2012, *Residual cooling and persistent star formation amid active galactic nucleus feedback in Abell 2597*, MNRAS 424, 1042, p. 1046, 1053, Figs. 1, 4. Lower left panel: G.R. Tremblay et al. 2012, *Multiphase signatures of active galactic nucleus feedback in Abell 2597*, MNRAS 424, 1026, p. 1036, Fig. 7. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

a cluster at $z \sim 0.6$ was discovered with an extremely large cooling mass rate $\dot{M} \sim 4000 M_\odot/\mathrm{yr}$ where the BCG shows signs of massive star formation, estimated to be $\sim 700 M_\odot/\mathrm{yr}$. Such cases are rare, however, and may be a transitional state in the cluster evolution.

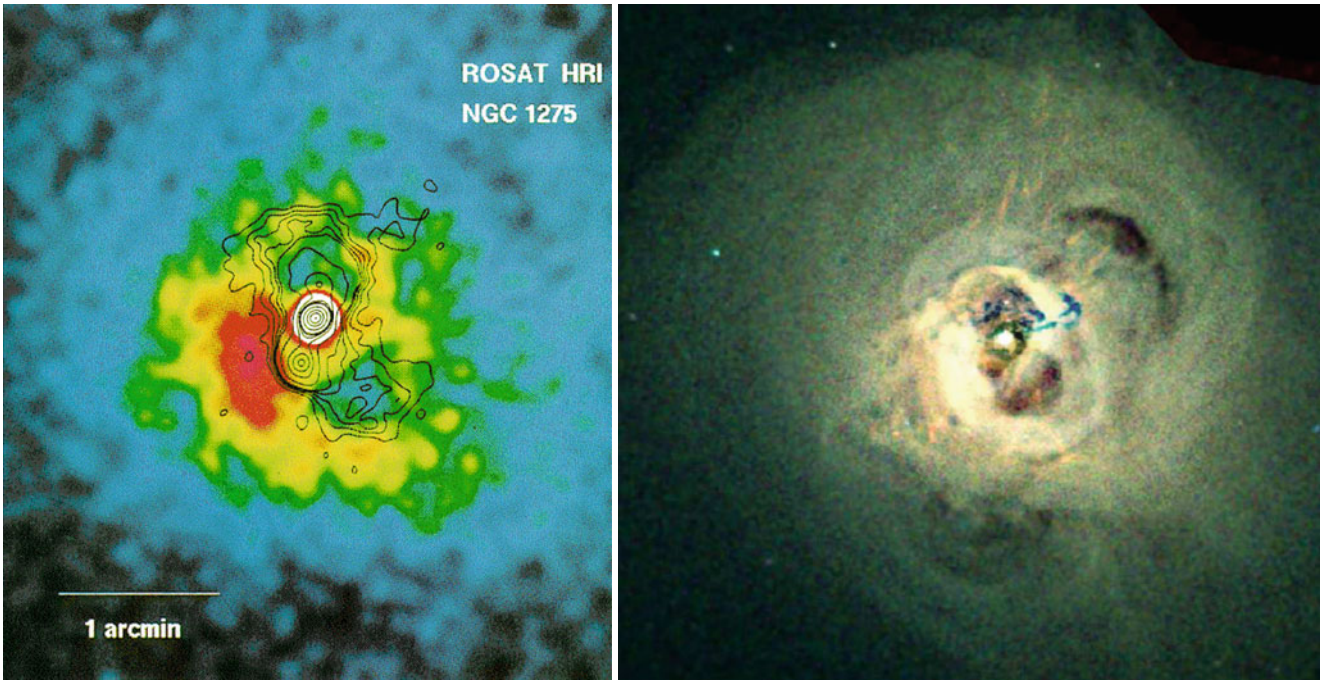


Fig. 6.29 The central region of the Perseus cluster. The *left panel* shows a color-coded X-ray image, taken with ROSAT, with radio contours superposed. The radio emission stems from the central galaxy of the cluster, the radio galaxy NGC 1275 (also called Perseus A). Clearly identifiable is the effect of the radio jets on the X-ray emission—at the location of the radio lobes the X-ray emission is strongly suppressed. The *right panel* shows a slightly larger region in X-rays, taken with the considerably better resolution of Chandra. The fine structure of the X-ray gas is much better recognized here. In addition to the two cavities overlapping with the radio emission, another pair of bubbles at larger distance from the center is seen. The high X-ray luminosity and small distance from us allows a very detailed analysis of this cluster. In particular, sound waves in the X-ray gas can be identified which probably are due to earlier activity of the central black hole. The blue filaments near the center indicates hard X-ray emission; the hardening of the observed spectrum is due to an infalling galaxy at high velocity, whose interstellar medium absorbs the low-energy X-ray photons. Credit: *Left*: H. Böhringer, MPE. *Right*: NASA/CXC/IoA/A. Fabian et al.

Cool-core clusters. Not all clusters show indications for a cool inner region with a strongly peaked X-ray emission. If clusters are far from an equilibrium state, for example due to a recent merger or infall of a group, the gas will be far from a quasi-steady state, and the foregoing consideration will not apply, not even approximately. Strong mixing of the gas by turbulent motions, or shock fronts which develop when the intracluster medium of two colliding or merging clusters intersect, will prevent the development of a cool, condensed central region. One thus distinguishes between *cool-core clusters* and non-cool-core clusters. The former ones are expected to be close to hydrostatic equilibrium, whereas the latter ones may deviate from it strongly.

What prevents massive cooling flows? One way to explain the clearly suppressed cooling rates in cooling flows is by noting that many clusters of galaxies harbor an active galaxy in their center. In most cases, this AGN is not a luminous quasar, but radio galaxies are the most common type of AGNs in the BCG of clusters, the activity of which, e.g., in the form of (radio-)jets, may affect the ICM. For instance, energy could be transferred from the jet to the ICM, by which the ICM is heated. This heating might then prevent the temperature from dropping to arbitrarily small values. This hypothesis is supported by the fact that many clusters are known in which the ICM is clearly affected by the central AGN—see Fig. 6.29 for one of the first examples where this effect was seen, and Fig. 6.30 where two active nuclei are detected. In the cluster Abell 2597 shown in Fig. 6.28 the interaction of the radio source with the intracluster gas is also clearly seen. Plasma from the jet seems to locally displace the X-ray emitting gas. By friction and mixing in the interface region between the jet and the ICM, the latter is certainly heated. It is unclear, though, whether this explanation is valid for every cluster, because not every cluster in which a very cool ICM is expected also contains an observed AGN. On the other hand, this is not necessarily an argument against the hypothesis of AGNs as heating sources, since AGNs often have a limited time of activity and may be switching on and off, depending on the accretion rate. Thus, the gas in a cluster may very well be heated by an AGN even if it is currently (at the time of observation) inactive. Evidence for the occurrence of this effect also in galaxy groups is shown in Fig. 6.31.

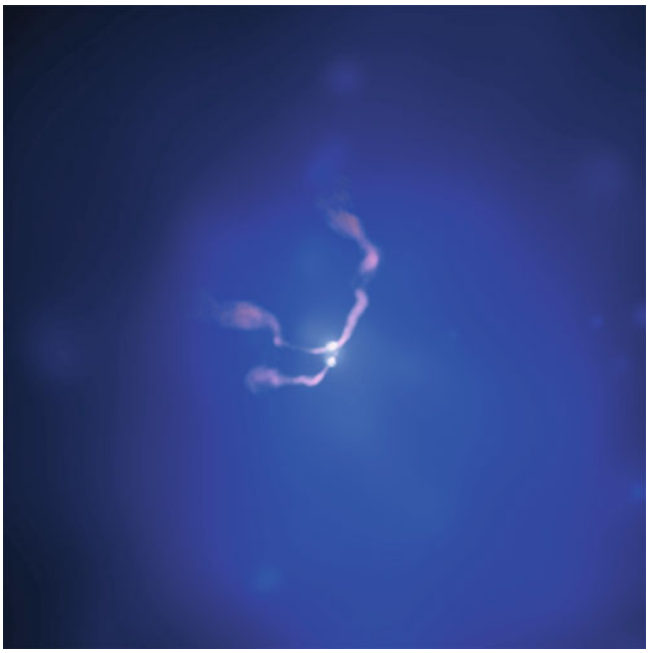


Fig. 6.30 A composite X-ray (*blue*, taken by Chandra) and radio (*pink*, VLA) image of the galaxy cluster Abell 400. The radio jets are launched by a binary supermassive black hole in the center of the galaxy NGC 1128, also known as 3C 75, a wide angle tailed radio source. The energy of the jets is partly transferred to the hot intracluster medium, which is thereby heated. The strong curvature of the jets is caused by the relative motion of the galaxy through the ICM, i.e., the jets are 'pushed back' by ram pressure. With a (projected) separation of about 8 kpc, this is one of the closest supermassive black hole binary system known. Credit: X-ray: NASA/CXC/AIfA/D. Hudson & T. Reiprich et al.; Radio: NRAO/VLA/NRL

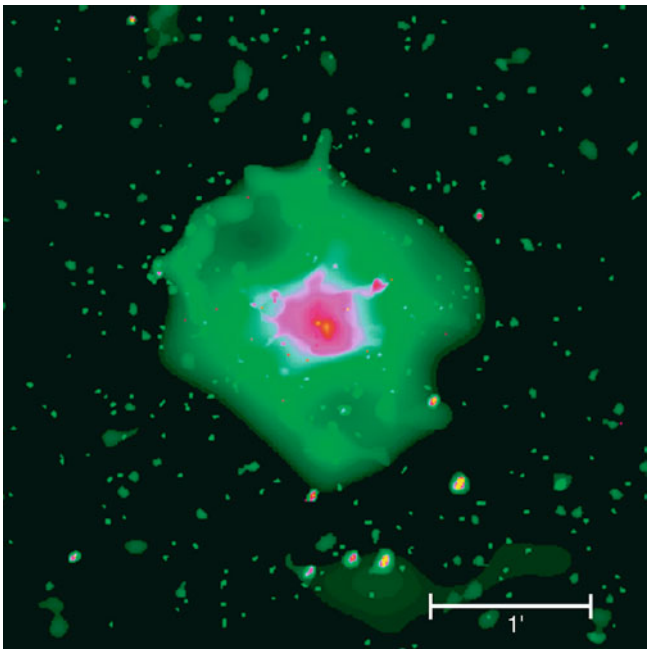


Fig. 6.31 Galaxy groups are also X-ray emitters, albeit weaker than clusters of galaxies. Moreover, the temperature of the ICM is lower than in clusters. This $4' \times 4'$ Chandra image shows HCG 62. Note the complexity of the X-ray emission and the two symmetrically aligned regions that seem to be virtually devoid of hot ICM—possibly holes blown free by jets from the central galaxy of this group (NGC 4761). Credit: NASA/CfA/J. Vrtilek et al.

Two more examples of the impact of a central AGN on the cluster gas are shown in Fig. 6.32. The energetics of this interaction can be enormous, as can be seen in the cluster MS 0735.6+7421 shown in the right panel of Fig. 6.32. The large size of the cavities in the X-ray emitting gas implies that a huge amount of energy was needed to push the gas away. From the cavity size and the gas density, one estimates that about $10^{12} M_\odot$ of gas has been displaced, requiring an energy of about 10^{62} erg. If this energy was generated by accretion onto the supermassive black hole located in the central galaxy of the cluster, with a mass-to-energy conversion of 10 %, the mass of the SMBH has grown by $\sim 6 \times 10^8 M_\odot$. Assuming that the gas was removed with about the sound speed of the ICM, this energy was released in the past $\sim 10^8$ yr, implying a mean luminosity of the central AGN of $\sim 3 \times 10^{46}$ erg/s of mechanical energy. Hence, by all accounts, this is a very energetic event which strongly impacts on the ICM of this cluster.

Feedback. The heating of the intracluster medium by a central radio source in clusters is the most visible example of *feedback*. We will encounter other examples lateron when we discuss the evolution of galaxies. As the gas cools and sinks towards the center, the central supermassive black hole can get fresh fuel and starts producing energy. The corresponding energy output in form of kinetic power in the radio jets or in radiation then heats the gas again, preventing efficient cooling and thus limits the mass accretion rate—and thus the fueling of the AGN. Hence, one might have a feedback loop, which does not need to have a stable equilibrium. The mass accretion rate may vary in time, as well as the AGN power output. Most likely, the feedback loop is somewhat more complicated than depicted here, but there is no doubt that AGN feedback is essential for understanding the gas in galaxy clusters.

Wide angle tail radio galaxies. Radio galaxies in clusters often have a different radio morphology than isolated ones. The radio jets necessarily interact with the intracluster gas and will be affected by it. Correspondingly, if the radio galaxy has a significant velocity relative to the ICM, the shape of the jets will get bend by ram pressure. A typical example for such *wide angle tail radio source* is shown in Fig. 6.33. Radio sources with such strong bended jets essentially only occur in clusters. Therefore, such sources can be used to search for clusters, and indeed these cluster searches have been successful—up to redshifts of order unity.

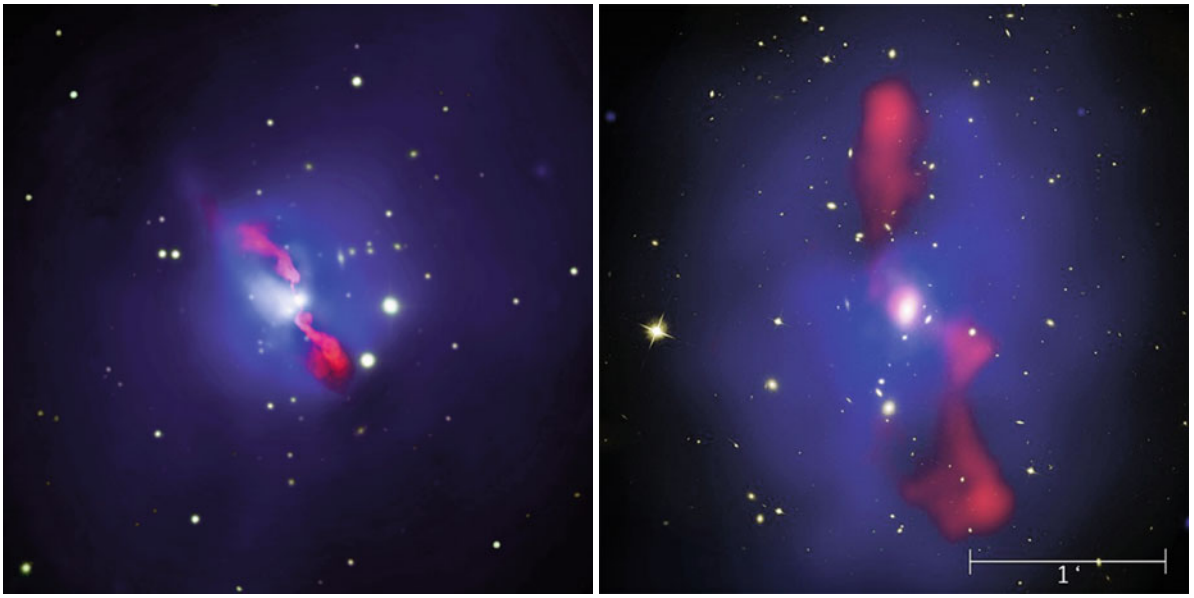


Fig. 6.32 The *left panel* shows the galaxy cluster Hydra A at redshift $z = 0.054$. This composite image is 4′.8 on the side and shows the X-ray emission (*blue*) and radio emission (*red*) superposed on the optical image. Clearly seen is the impact of the radio jets on the intracluster gas—at the location of the jets, the X-ray emission is strongly suppressed, whereas around the radio jets, overdensities are visible. Here, the gas, which has been pushed away from the jet, is accumulated. The *right panel* shows a composite image of the galaxy cluster MS 0735.6+7421 with redshift $z = 0.216$, showing the inner 700 kpc (corresponding to 200″) of the cluster. The Chandra X-ray image is shown in *blue*, and the radio emission as seen by the VLA shown in *red*, both superposed on an optical HST image of the cluster. The X-ray emission displays large cavities, located right at the position where the radio jets pass through the intracluster medium. This cluster contains presumably the most energetic interaction of the central AGN with the intracluster medium. Credit: *Left*: X-ray: NASA/CXC/U.Waterloo/C. Kirkpatrick et al.; Radio: NSF/NRAO/VLA; Optical: Canada-France-Hawaii-Telescope/DSS. *Right*: NASA, ESA, CXC, STScI, and B. McNamara (University of Waterloo), NRAO, and L. Birzan and team (Ohio University); journal article: B.R. McNamara et al. 2009, *An Energetic AGN Outburst Powered by a Rapidly Spinning Supermassive Black Hole or an Accreting Ultramassive Black Hole*, ApJ 698, 594, p. 595, Fig. 1 ©AAS. Reproduced with permission

The Bullet cluster. Clusters of galaxies are indeed excellent laboratories for hydrodynamical and plasma-physics processes on large scales. Shock fronts, for instance in merging clusters, cooling fronts (which are also called 'contact discontinuities' in hydrodynamics), and the propagation of sound waves can be observed in their intracluster medium. A particularly good example is the galaxy cluster 1E 0657−56 displayed in Fig. 6.34, also called the 'Bullet cluster'. To the right of the cluster center, strong and relatively compact X-ray emission (the 'bullet') is visible, while further to the right of it one sees an arc-shaped discontinuity in surface brightness. From the temperature distribution on both sides of the discontinuity one infers that it is a shock front—in fact, the shape of the gas distribution on the right resembles that of the air around a supersonic plane or bullet. The strength of the shock implies that the 'bullet' is moving at about $v \sim 3500\,\mathrm{km/s}$ through the intergalactic medium of the cluster. The interpretation of this observation is that we are witnessing the collision of two clusters, where one less massive cluster has passed, from left to right in Fig. 6.34, through a more massive one. The 'bullet' in this picture is understood to be gas from the central region of the less massive cluster, which is still rather

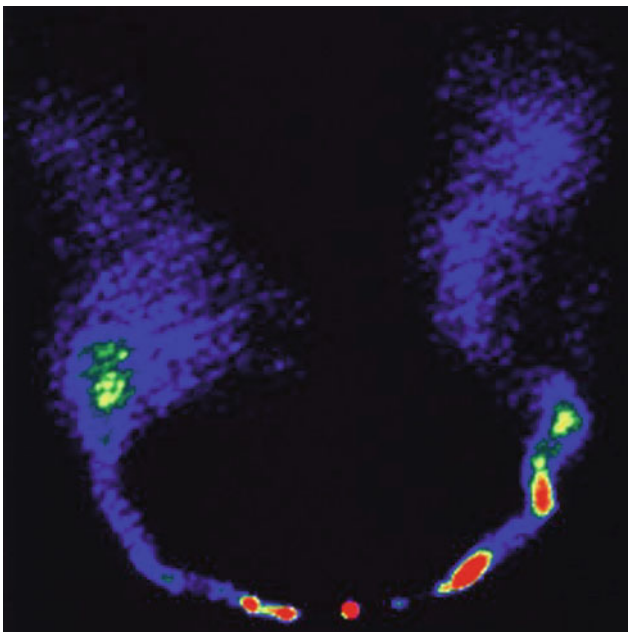


Fig. 6.33 NGC 1265, a radio galaxy in the Perseus cluster. Owing to its motion through the intracluster gas, with an estimated velocity of $\sim 2000\,\mathrm{km/s}$, its radio jets get bend by ram pressure. Credit: Image courtesy of NRAO/AUI and C. O'Dea & F. Owen

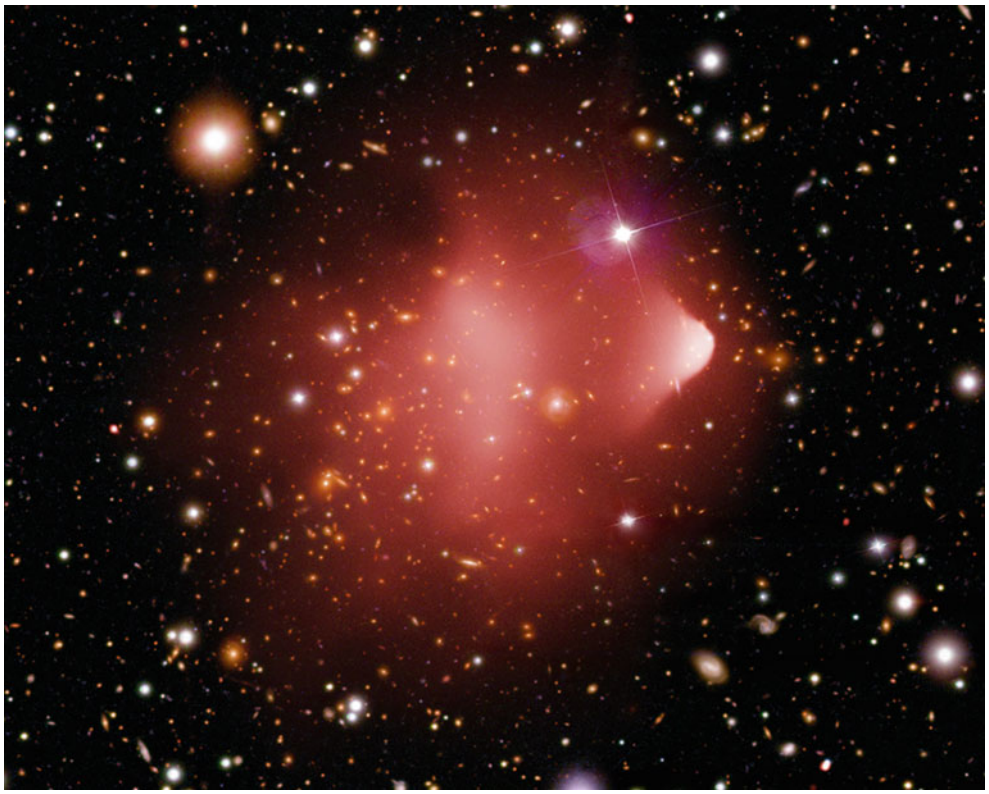


Fig. 6.34 The cluster of galaxies 1E 0657−56 is a perfect example of a merging cluster. The X-ray emission of this cluster as observed by Chandra, shown in *red*, is superposed on an optical HST image. The most remarkable feature in the X-ray map is the compact region to the right (westwards) of the cluster center (from which the cluster derives its name the 'Bullet cluster'), and the sharp transition in the surface brightness further at its right edge. An analysis of the brightness profile and of the X-ray temperature distribution shows that this must be a shock front moving at about 2.5 times the speed of sound, or $v \sim 3500\,\mathrm{km/s}$, through the gas. To the right of this shock front, a group of galaxies is visible. Credit: X-ray: NASA/CXC/CfA/M. Markevitch et al.; Optical: NASA/STScI; Magellan/U.Arizona/D. Clowe et al.

compact. This interpretation is impressively supported by the group of galaxies to the right of the shock front, which are probably the former member galaxies of the less massive cluster. As this cluster crosses through the more massive one, its galaxies and dark matter are moving collisionlessly, whereas the gas is decelerated by friction with the gas in the massive cluster: the galaxies and the dark matter are thus able to move faster through the cluster than the gas, which is lagging behind (and whose momentum transferred to the gas of the more massive cluster displaces the latter from its original location, centered on the corresponding galaxy distribution). We will see below that this interpretation is verified by a gravitational lens investigation of this double cluster.

6.4.4 The Sunyaev–Zeldovich effect

Electrons in the hot gas of the intracluster medium can scatter photons of the cosmic microwave background. The optical depth and thus the scattering probability for this Compton scattering is relatively low, but the effect is nevertheless observable and, in addition, is of great importance for the analysis of clusters, as we will now see.

Spectral signature. A photon moving through a cluster of galaxies towards us will change its direction through scattering and thus will not reach us. But since the cosmic background radiation is isotropic, for any CMB photon that is scattered out of the line-of-sight, another photon exists—statistically—that is scattered into it, so that the total number of photons reaching us is preserved. However, the energy of the photons changes slightly through scattering by the hot electrons, in a way that they have an (on average) higher frequency after scattering. Hence, by this inverse Compton scattering (Sect. 5.4.4), energy is on average transferred from the electrons to the photons, as can be seen in Fig. 6.35.

As a consequence, this scattering leads to a reduced number of photons at lower energies, relative to the Planck spectrum, which are shifted to higher energy. There is one photon energy where the intensity is unchanged, corresponding to a frequency of 218 GHz; below that frequency, the intensity is lower than that of the CMB, for higher frequencies, the intensity is increased. This effect is called the *Sunyaev–Zeldovich effect* (SZ-effect). It was predicted in 1970 and has now been observed in a large number of clusters. One example is presented in Fig. 6.36, where the frequency dependence of the effect is clearly seen.

The CMB spectrum, measured in the direction of a galaxy cluster, deviates from a Planck spectrum; the degree of this deviation depends on the temperature of the cluster gas and on its density, and is independent of the cluster redshift.

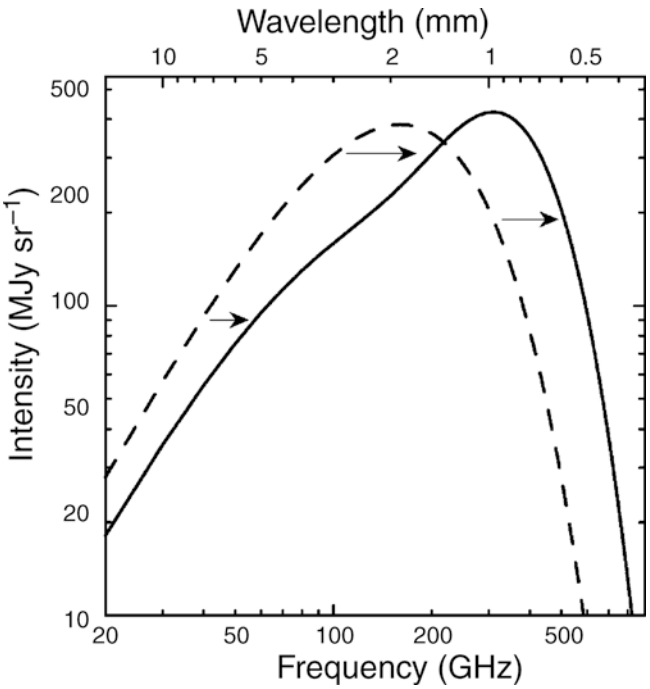


Fig. 6.35 The influence of the Sunyaev–Zeldovich effect on the cosmic background radiation. The *dashed curve* represents the Planck distribution of the unperturbed CMB spectrum, the *solid curve* shows the spectrum after the radiation has passed through a cloud of hot electrons. The magnitude of this effect, for clarity, has been very much exaggerated in this sketch. Source: J. Carlstrom et al. 2002, ARA&A 40, 643, Fig. 1, p. 646. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 40 ©2002 by Annual Reviews www.annualreviews.org

In the Rayleigh–Jeans domain of the CMB spectrum, at wavelengths larger than about 2 mm, the intensity of the CMB is decreased by the SZ-effect. For the change in specific intensity in the RJ part, one obtains

$$\boxed{\frac{\Delta I_\nu^{\mathrm{RJ}}}{I_\nu^{\mathrm{RJ}}} = -2y}\,, \tag{6.46}$$

where

$$\boxed{y = \int \mathrm{d}l\, \frac{k_\mathrm{B} T_\mathrm{g}}{m_\mathrm{e} c^2}\, \sigma_\mathrm{T}\, n_\mathrm{e}} \quad \text{with} \quad \sigma_\mathrm{T} = \frac{8\pi}{3}\left(\frac{e^2}{m_\mathrm{e} c^2}\right)^2 \tag{6.47}$$

is the *Compton-y parameter* and σ_T the Thomson cross section for electron scattering. Obviously, y is proportional to the optical depth with respect to Compton scattering, given as an integral over $n_\mathrm{e}\,\sigma_\mathrm{T}$ along the line-of-sight. Furthermore, y is proportional to the gas temperature, because that defines the average energy transfer per scattering event [see (5.35)]. Overall, y is proportional to the integral over the gas pressure $P = nk_\mathrm{B}T_\mathrm{g}$ along the line-of-sight through the cluster.

The fact that the SZ-signal $\Delta I_\nu / I_\nu$ is independent of cluster redshift allows the investigation of clusters at high redshifts, provided the SZ-signal is spatially resolved. As we will see below, the SZ-effect can also be used to detect clusters in the first place, and this selection is much less biased to low redshifts than for flux-limited optical or X-ray surveys. As an example, the right-hand panel of Fig. 6.37 shows a very massive, high-redshift cluster that was selected by an SZ-survey.

Observations of the SZ-effect provide another possibility for analyzing the gas in clusters. For instance, if one can spatially resolve the SZ-effect, which is possible today with interferometric methods (see Fig. 6.38), one obtains information about the spatial density and temperature distribution. Here it is of crucial importance that the dependence on temperature and gas density is different from that in X-ray emission. Because of the quadratic dependence of the X-ray emissivity on n_e, the X-ray luminosity depends not only on the total gas mass, but also on its spatial distribution. Small-scale clumps in the gas, for instance, would strongly affect the X-ray emission. In contrast, the SZ-effect is linear in gas density and therefore considerably less sensitive to small-scale inhomogeneities in the ICM.

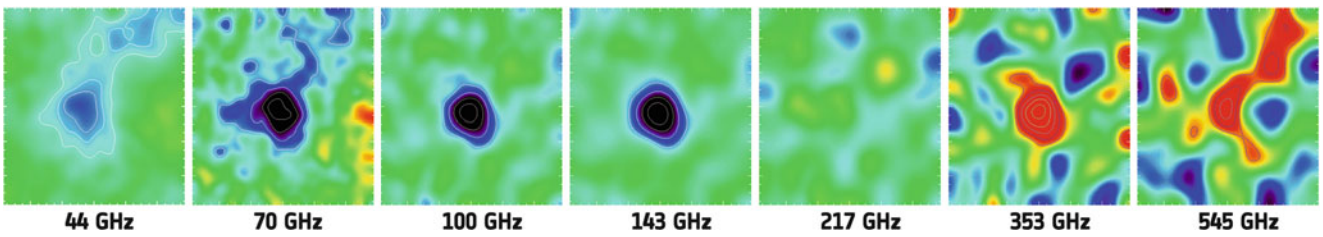


Fig. 6.36 Maps of the nearby cluster Abell 2319 in seven frequencies, obtained by the Planck satellite. These maps (with about 2° sidelength) clearly indicate the Sunyaev–Zeldovich effect caused by the hot gas in this cluster. At low frequencies, the SZ-effect causes a decrease of the surface brightness, whereas at high frequencies, the intensity is increased. The transition between these two regimes occurs at 218 GHz, as shown in Fig. 6.35; indeed, in the map of this frequency, no signal is seen. Credit and Copyright: ESA/LFI & HFI Consortia

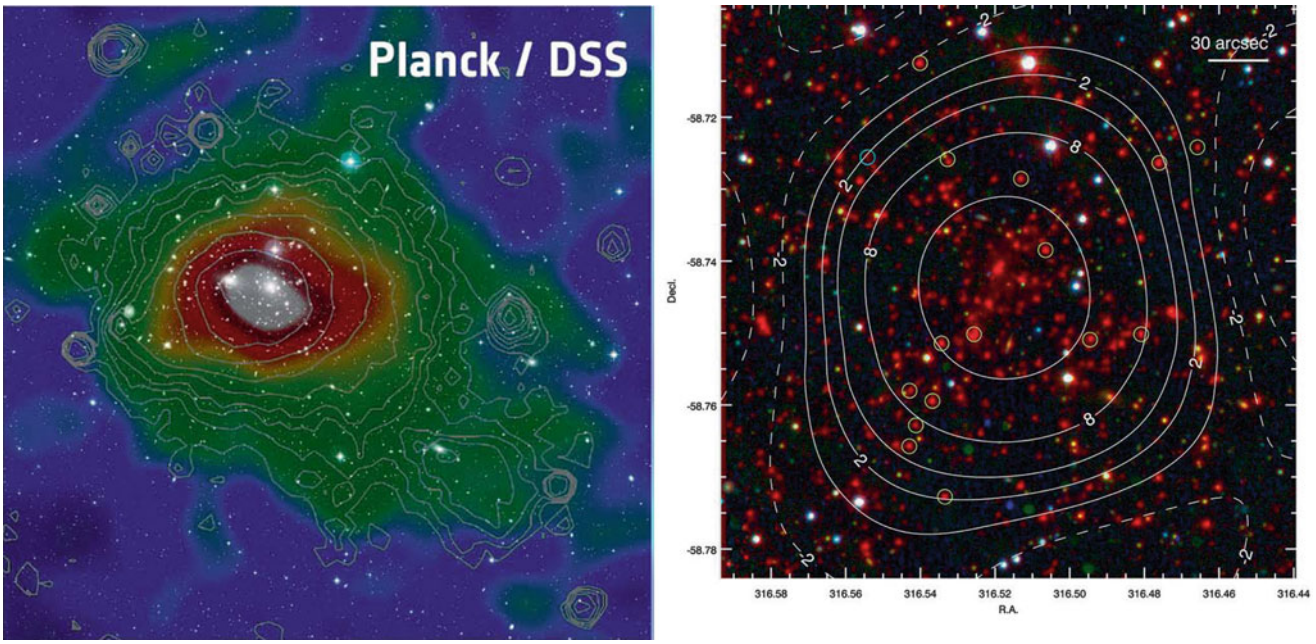


Fig. 6.37 *Left*: The Sunyaev–Zeldovich effect of the Coma cluster, as seen by the Planck satellite. The microwave temperature depletion is color coded, whereas the white contours display the X-ray emission of the cluster as measured by ROSAT. Both maps are superposed on the optical image from the Digitized Sky Survey 2. Note the very close correspondence between the SZ-signal and the X-ray emission. *Right*: The cluster SPT-CL J2106−5844 was detected by its SZ-signal, which is shown here as contour lines, superposed on a composite optical and mid-IR image, taken with the Magellan telescope and the Spitzer observatory, respectively. The cluster has a redshift of $z = 1.13$, and is (one of) the most massive clusters at redshift $z > 1$, with an estimated virial mass of $M_{200} \sim 1.3 \times 10^{15} M_\odot$. The cluster is also detected in X-rays, showing a very high temperature of $T \sim 11$ keV, and an X-ray luminosity of $L_X \sim 1.4 \times 10^{45}$ erg/s. The image is 4′8 on the side. Encircled galaxies have their cluster membership spectroscopically confirmed. Credit and Copyright: *Left*: Planck image: ESA/LFI & HFI Consortia; ROSAT image: Max-Planck-Institut für extraterrestrische Physik; DSS image: NASA, ESA, and the Digitized Sky Survey. Acknowledgment: Davide De Martin (ESA/Hubble). *Right*: R.J. Foley et al. 2011, *Discovery and Cosmological Implications of SPT-CL J2106−5844, the Most Massive Known Cluster at* $z > 1$, ApJ 731, 86, p. 3, Fig. 1. ©AAS. Reproduced with permission

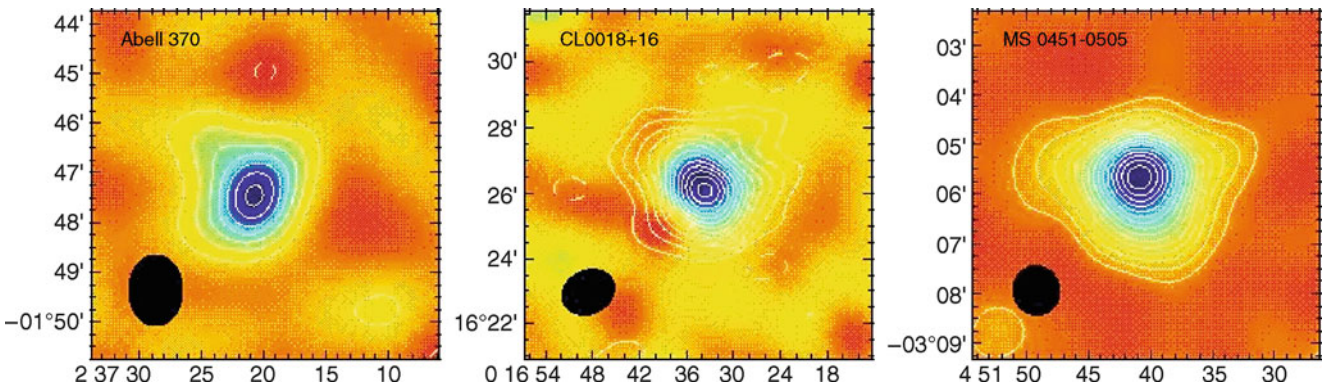


Fig. 6.38 Sunyaev–Zeldovich maps of three clusters of galaxies at $0.37 < z < 0.55$. Plotted is the temperature difference of the measured CMB relative to the average CMB temperature (or, at fixed frequency, the difference in radiation intensities). The black ellipse in each image specifies the instrument's beam size. For each of the clusters shown here, the spatial dependence of the SZ-effect is clearly visible. Since the SZ-effect is proportional to the electron density, the mass fraction of baryons in clusters can be measured if one additionally knows the total mass of the cluster from dynamical methods or from the X-ray temperature. The analysis of the clusters shown here yields for the mass fraction of the intergalactic gas $f_g \approx 0.08\, h^{-1}$. Source: L. Grego et al. 2001, *Galaxy Cluster Gas Mass Fractions from Sunyaev-Zeldovich Effect Measurements: Constraints on* Ω_m, ApJ 552, 2, p. 7, Fig. 1. ©AAS. Reproduced with permission

The integrated *y*-parameter. The amplitude of the spectral distortion caused by the SZ-effect is given by y, and SZ-maps essentially provide a map of y as a function of angular position in the cluster. When integrating y across the cluster, we get the total SZ-signal,

$$\int \mathrm{d}^2\theta\; y = \frac{1}{D_A^2} \int \mathrm{d}^2 R\; y \propto \frac{1}{D_A^2} \int \mathrm{d}V\; n_e\, T_g\;; \qquad (6.48)$$

hence, the integrated SZ-effect is proportional to the integrated SZ-parameter

$$Y = M_{\rm g}\, T_{\rm g}\ . \tag{6.49}$$

We see that Y is a measure of the product of gas mass and temperature. Both of these quantities can also be determined from X-ray observations, so that an independent estimate of Y can be obtained. In order to distinguish between the two, one usually denotes the result from SZ-observations by $Y_{\rm SZ}$, in contrast to $Y_{\rm X}$ when it is determined from X-ray studies. These two can be different in general, due to the quadratic dependence of the X-ray emission on local gas density or through the temperature variation inside clusters.

Kinetic SZ-effect. Beside the thermal SZ-effect just described, there is a related effect, called the kinetic SZ-effect. This is due to the fact that clusters may have a peculiar velocity. Suppose the peculiar velocity of a cluster is directed towards us, then the photons scattering in its intracluster gas and reaching us will be scattered by electrons which have an average velocity towards us. These scattered photons thus experience on average a blueshift, which is visible in the CMB spectrum in the direction of this cluster. The kinetic SZ-effect has a different spectral signature than the thermal SZ-effect and can thus in principle be distinguished from it. A robust measurement of the kinetic SZ-effect would allow a direct measurement of the line-of-sight component of the peculiar velocities of clusters. However, the expected amplitude of the kinetic SZ-effect is smaller than that of the thermal SZ-effect by a factor of ~ 10 and thus much harder to detect. First detections have recently been reported in the literature.

Distance determination. For a long time, the SZ-effect was mainly considered a tool for measuring distances to clusters of galaxies, and from this to determine the Hubble constant. We will now schematically show how the SZ-effect, in combination with the X-ray emission, allows us to determine the distance to a cluster. The change in the CMB intensity has the dependence

$$\frac{\left|\Delta I_\nu^{\rm RJ}\right|}{I_\nu^{\rm RJ}} \propto n_{\rm e}\, L\, T_{\rm g}\ ,$$

where L is the extent of the cluster along the line-of-sight. To obtain this relation, we replace the l-integration in (6.47) by a multiplication with L, which yields the correct functional dependence. On the other hand, the surface brightness of the X-ray radiation behaves as

$$I_{\rm X} \propto L n_{\rm e}^2\ .$$

Combining these two relations, we are now able to eliminate $n_{\rm e}$. Since $T_{\rm g}$ is measurable from the X-ray spectrum, the dependence

$$\frac{\left|\Delta I_\nu^{\rm RJ}\right|}{I_\nu^{\rm RJ}} \propto \sqrt{L\, I_{\rm X}}$$

remains. Now assuming that the cluster is spherical, its extent L along the line-of-sight equals its transverse extent $R = \theta D_{\rm A}$, where θ denotes its angular extent and $D_{\rm A}$ the angular-diameter distance (4.49) to the cluster. With this assumption, we obtain

$$D_{\rm A} = \frac{R}{\theta} \sim \frac{L}{\theta} \propto \left(\frac{\Delta I_\nu^{\rm RJ}}{I_\nu^{\rm RJ}}\right)^2 \frac{1}{I_{\rm X}}\ . \tag{6.50}$$

Hence, the angular-diameter distance can be determined from the measured SZ-effect, the X-ray temperature of the ICM, and the surface brightness in the X-ray domain. Of course, this method is more complicated in practice than sketched here, but it is applied to the distance determination of clusters. In particular, the assumption of the same extent of the cluster along the line-of-sight as its transverse size is not well justified for any individual cluster due to triaxiality, but one expects this assumption to be valid on average for a sample of clusters. Hence, the SZ-effect is another method of distance determination, independent of the redshift of the cluster, and therefore suitable for determining the Hubble constant.

Discussion. The natural question arises whether this method, in view of the assumptions it is based on, can compete with the determination of the Hubble constant via the distance ladder and Cepheids, as described in Sect. 3.9, or from the CMB, to be discussed in Sect. 8.7. The same question also needs to be asked for the determination of H_0 by means of the time delay in gravitational lens systems, which we discussed in Sect. 3.11.4. In both cases, the answer is the same: presumably neither of the two methods will provide a determination of the Hubble constant with an accuracy comparable to that achieved by the local methods and from the angular fluctuations in the CMB. Nevertheless, both methods are of great value for cosmology: first, the distance ladder has quite a number of rungs. If only one of these contains an as yet undetected severe systematic error, it could affect the resulting value for H_0. Second, the Hubble Key Project measured the expansion rate in the local Universe, typically within ~ 100 Mpc (the distance to the Coma cluster). As we will see later, the Universe contains inhomogeneities on these length scales. Thus, it may well be that we live in a slightly overdense or underdense region of the Universe, where the Hubble constant deviates from the global value. In contrast to this, both the SZ-effect and the lensing method measure the Hubble constant on truly cosmic scales, and both methods do so in only a single step—there is no distance ladder involved. For these reasons, these two methods are of considerable interest in additionally confirming our H_0 measurements. Another aspect adds to this, which must not be underestimated: even if the same or a similar value results from these measurements as the one from the Hubble Key Project, we still have learned an important fact, namely that the local Hubble constant agrees with the one measured on cosmological scales—as predicted by our cosmological model, which can thus be tested. Indeed, both methods have been applied to quite a number of lens systems and luminous clusters showing an SZ effect, respectively, and they yield values for H_0 which are compatible within the error bars with the value of H_0 obtained from the Hubble Key Project.

6.4.5 X-ray and SZ catalogs of clusters

As we have seen, projection effects may play a crucial role in the selection of galaxy clusters through searching for an overdensity of galaxies on the sphere using optical methods. A more reliable way of selecting clusters is by their X-ray emission, since the hot X-ray gas signifies a deep potential well, thus a real three-dimensional overdensity of matter, so that projection effects become virtually negligible. The X-ray emission is $\propto n_{\rm e}^2$, which again renders projection effects improbable. In addition, the X-ray emission, its temperature

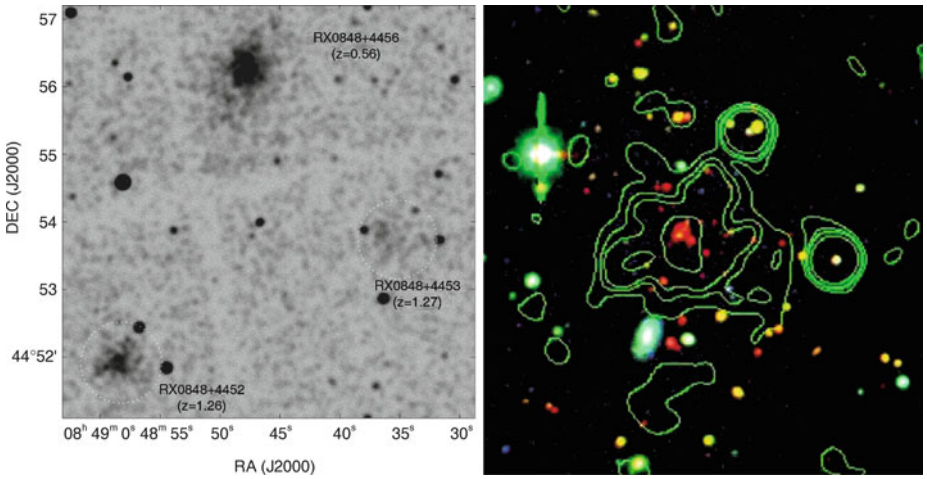


Fig. 6.39 *On the left*: Chandra image of a $6' \times 6'$-field with two clusters of galaxies at high redshift. *On the right*: a $2' \times 2'$-field centered on one of the clusters presented on the left (RX J0849+4452), in B, I, and K, overlaid with the X-ray brightness contours. Source: S.A. Stanford et al. 2001, *The Intracluster Medium in $z > 1$ Galaxy Clusters*, ApJ 552, 504, p. 505, 507, Figs. 1, 3. ©AAS. Reproduced with permission

in particular, seems to be a very good measure for the cluster mass, as we will discuss further below. Whereas the selection of clusters is not based on their temperature, but on the X-ray luminosity, we shall see that L_X is also a good indicator for the mass of a cluster (see Sect. 6.5).

The first cosmologically interesting X-ray catalog of galaxy clusters was the EMSS (Extended Medium Sensitivity Survey) catalog. It was constructed from archival images taken by the *Einstein observatory* which were scrutinized for X-ray sources other than the primary target in the field-of-view of the respective observation. These were compiled and then further investigated using optical methods, i.e., photometry and spectroscopy. The EMSS catalog contains 835 sources, most of them AGNs, but it also contains 104 clusters of galaxies. Among these are six clusters at redshift ≥ 0.5; the most distant is MS 1054−03 at $z = 0.83$ (see Fig. 6.21). Since the Einstein images all have different exposure times, the EMSS is not a strictly flux-limited catalog. But with the flux limit known for each exposure, the luminosity function of clusters can be derived from this.

The same method as was used to compile the EMSS was applied to ROSAT archival images by various groups, leading to several catalogs of X-ray-selected clusters. The selection criteria of the different catalogs vary. Since ROSAT was more sensitive than the Einstein observatory, these catalogs contain a larger number of clusters, and also ones at higher redshift (Fig. 6.39). Furthermore, ROSAT performed a survey of the full sky, the ROSAT All-Sky Survey (RASS). The RASS contains about 10^5 sources distributed over the whole sky. The identification of extended sources in the RASS (in contrast to non-extended sources—about five times more AGNs than clusters are expected) yielded a catalog of clusters which, owing to the relatively short exposure times in the RASS, contains the brightest clusters. The exposure time in the RASS is not uniform over the sky since the applied observing strategy led to particularly long exposures for the regions around the Northern and Southern ecliptic pole (see Fig. 6.40).

From the luminosity function of X-ray clusters, a mass function can be constructed, using the relation between L_X and the cluster mass that will be discussed in the following section. Furthermore, as we will explain in more detail in Sect. 8.2, this cluster mass function is an important probe for cosmological parameters.

More recently, the sensitivity and throughput of SZ telescopes and instruments became sufficiently large to not only study the SZ-effect of known clusters, but to survey the sky for SZ sources. In 2009 the first SZ-selected clusters were found, and at the time of writing, at least three telescopes are used for constructing cluster catalogs through SZ-selection. The South Pole Telescope (SPT), the Atacama Cosmology Telescope, and the Planck satellite, together have found more than 1000 clusters. The selection criteria for SZ clusters is quite different from those of X-ray (or optical) cluster samples, since we have seen that the strength of the SZ-signal is redshift-independent, as long as the cluster is spatially resolved. If, however, the angular resolution of the observations is not high enough to resolve the SZ-signal from a cluster, distant clusters are more difficult to find, since the signal gets diluted. On the other hand, the SZ-signal is weak, so that only the more massive clusters can be readily discovered. The different selection effects are clearly visible in the redshift and mass distribution of the resulting cluster catalogs, shown in Fig. 6.41. The fact that Planck finds the more massive clusters at redshifts $z \lesssim 0.5$ originates from its all-sky survey, in which the rare objects are found, whereas the SPT clusters in Fig. 6.41 were selected from a region of 720 deg^2.

6.4.6 Radio relics

In some galaxy clusters, one finds extended, diffuse radio sources at large cluster-centric radii, the so-called radio relics. These radio sources do not coincide with member galaxies in the cluster, and thus have an intracluster origin. Their shapes are often very elongated or irregular, as can

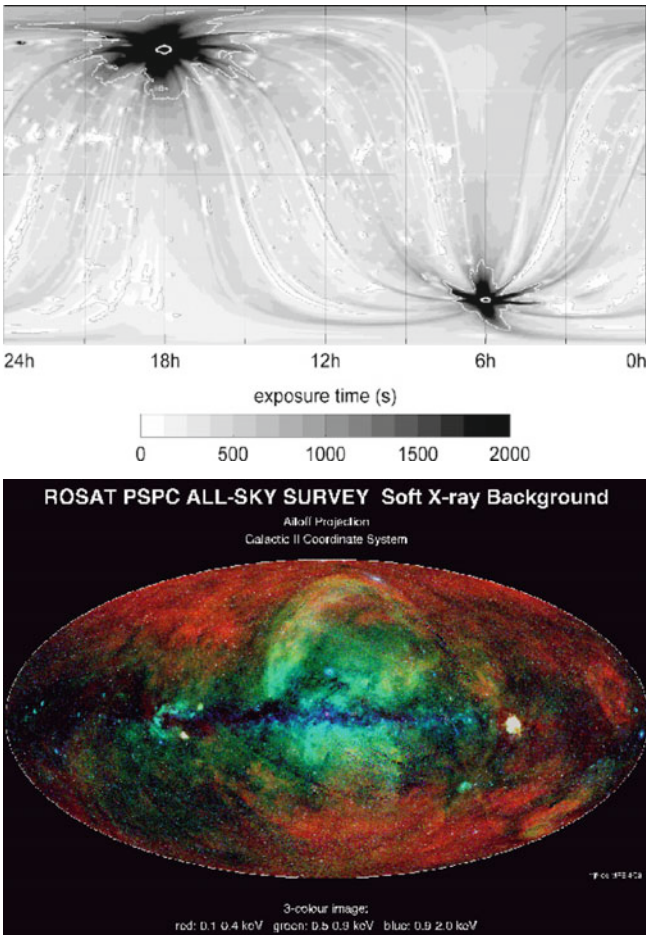


Fig. 6.40 The *top panel* shows the total exposure time in the ROSAT All-Sky Survey as a function of sky position. Near the ecliptic poles the exposure time is longest, as a consequence of the applied observing strategy. Because of the "South Atlantic Anomaly" (a region of enhanced cosmic ray flux over the South Atlantic Ocean, off the coast of Brazil, caused by the shape of the Earth's magnetosphere), the exposure time is generally higher in the North than in the South. The X-ray sky, as observed in the RASS, is shown in the *lower panel*. The colors indicate the shape of the spectral energy distribution, where *blue* indicates sources with a harder spectrum. Credit: Max-Planck-Institut für Extraterrestrische Forschung, Garching; journal article: S.L. Snowden et al. 1997, *ROSAT Survey Diffuse X-Ray Background Maps. II.*, ApJ 485, 125

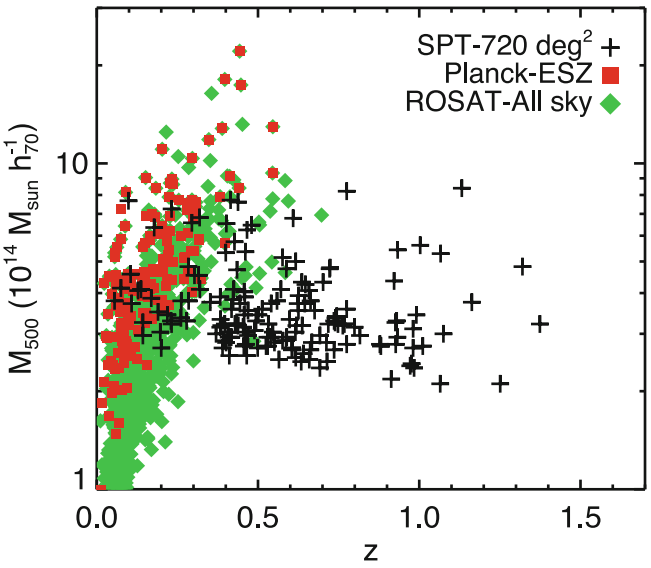


Fig. 6.41 Distribution of clusters in redshift and mass, as obtained from three different surveys. *Green* points show clusters selected through their X-ray emission, here from the ROSAT All-Sky Survey (see Fig. 6.40). The *crosses* show clusters selected by their SZ-signal, as obtained from the South Pole Telescope; its high angular resolution resolves the SZ-signal even for very distant clusters. Correspondingly, the mass distribution of the selected cluster does not show any marked redshift dependence, at least out to $z \sim 1$. *Red* points show clusters selected by Planck, also through their SZ-signal. Due to the lower angular resolution of Planck, the signal from higher-redshift clusters, which are smaller than the telescope beam, is diluted. Source: C.L. Reichardt et al. 2013, *Galaxy Clusters Discovered via the Sunyaev-Zel'dovich Effect in the First 720 Square Degrees of the South Pole Telescope Survey*, ApJ 763, 127, Fig. 4. ©AAS. Reproduced with permission

be seen by the example shown in Fig. 6.42. The strong polarization of the radio emission shows that the origin of the radiation is synchrotron emission. Therefore, these radio relics must contain relativistic electrons. About 50 radio relics have been found to date.

As we discussed in Sect. 5.1.3, relativistic particles are accelerated in shock fronts, such as occur in supernova remnants. Shock fronts are formed in plasmas with supersonic flow velocities. A natural way of explaining the occurrence of a shock in the outskirts of a cluster is to assume that the cluster has been subject to a recent merger event. If the intracluster gas of the two clusters run into each other, a shock front is formed. For the radio relic shown in Fig. 6.42, the predictions of the shock hypothesis were tested in quite some detail. The radio spectral index α varies strongly across the relic, from $\alpha \approx 0.6$ to about 2.0, indicating that the electron distribution closest to the shock has the flattest spectrum, whereas it continuously steepens away from the shock, due to energy losses. The high degree of polarization ($\sim 50\,\%$) shows the presence of a very well-ordered magnetic field in the emission region.

The merging hypothesis for the example shown in Fig. 6.42 is further supported by the strongly disturbed morphology of the intracluster gas, as shown by the X-ray contours, as well as by the indication of a second relic on the opposite side of the cluster center; such counter relics are predicted from numerical simulations of cluster mergers.

6.5 Scaling relations for clusters of galaxies

Our examination of galaxies revealed the existence of various scaling relations, for example the Tully–Fisher relation.

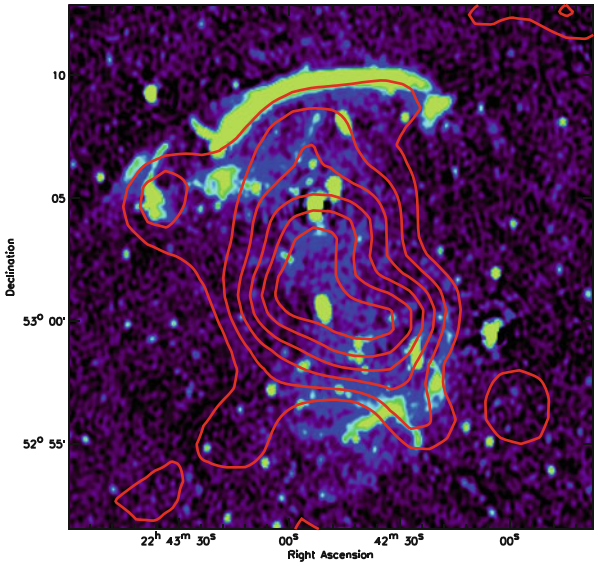


Fig. 6.42 1.4 GHz map of the galaxy cluster CIZA J2242.8+5301 at redshift $z = 0.1921$, taken with the Westerbork Synthesis Radio Telescope (WSRT), on which the X-ray contours (shown in *red*) are superposed. The most prominent radio source is the large elongated feature some 1.5 Mpc to the North of the cluster center, whose length is ~ 2 Mpc. Source: R.J. van Weeren et al. 2010, arXiv:1010.4306, Fig. 1. Reproduced by permission of the author

These have proven to be very useful not only for the distance determination of galaxies, but also because any successful model of galaxy evolution needs to be able to explain these empirical scaling relations—they must contain information about the formation of galaxies. Therefore, it is of great interest to examine whether clusters of galaxies also fulfill any such scaling relation. As we will see, the X-ray properties of clusters play a central role in this.

6.5.1 Mass-temperature relation

It is expected that the larger the spatial extent, velocity dispersion of galaxies, temperature of the X-ray gas, and luminosity of a cluster are, the more massive it is. In fact, from theoretical considerations one can deduce the existence of relations between these parameters. The X-ray temperature T specifies the thermal energy per gas particle, which should be proportional to the binding energy for a cluster in virial equilibrium,

$$T \propto \frac{M}{r} \,.$$

Since this relation is based on the virial theorem, r should be chosen to be the radius within which the matter of the cluster is virialized. This value for r is called the *virial radius* $r_{\rm vir}$. From theoretical considerations of cluster formation (see Chap. 7), one finds that the virial radius is defined such that within a sphere of radius $r_{\rm vir}$, the average mass density of the cluster is about $\Delta_{\rm c} \approx 200$ times as high as the critical density $\rho_{\rm cr}$ of the Universe (see also Problem 6.1). The mass within $r_{\rm vir}$ is called the *virial mass* $M_{\rm vir}$ which is, according to this definition,

$$\boxed{M_{\rm vir} = \frac{4\pi}{3}\,\Delta_{\rm c}\,\rho_{\rm cr}\,r_{\rm vir}^3} \,. \tag{6.51}$$

Combining the two above relations, one obtains

$$\boxed{T \propto \frac{M_{\rm vir}}{r_{\rm vir}} \propto r_{\rm vir}^2 \propto M_{\rm vir}^{2/3}} \,. \tag{6.52}$$

This relation can now be observationally tested by using a sample of galaxy clusters with known temperature and with mass determined by the methods discussed in Sect. 6.4.2. An example of this is displayed in Fig. 6.43, in which the mass is plotted versus temperature for clusters from the extended HIFLUGCS sample.[8] Since it is easier to determine the mass inside a smaller radius than the virial mass itself, the mass M_{500} within the radius r_{500}, the radius within which the average density is 500 times the critical density, is plotted here. The measured values clearly show a very strong correlation, and best-fit straight lines describing power laws of the form $M = AT^{\alpha}$ are also shown in the figure. The exact values of the two fit parameters depend on the choice of the cluster sample; the right-hand panel of Fig. 6.43 shows in particular that galaxy groups (thus, 'clusters' of low mass and temperature) are located below the power-law fit that is obtained from higher mass clusters. If one confines the sample to clusters with $M \geq 5 \times 10^{13} M_{\odot}$, the best fit is described by

$$M_{500} = 3.57 \times 10^{13} M_{\odot} \left(\frac{k_{\rm B} T}{1\,{\rm keV}}\right)^{1.58} \,, \tag{6.53}$$

with an uncertainty in the parameters of slightly more than 10 %. This relation is very similar to the one deduced from our theoretical consideration, $M \propto T^{1.5}$. With only small variations in the parameters, the relation (6.53) is obtained both from a cluster sample in which the mass was determined

[8] One of the cluster catalogs that were extracted from the RASS data is the HIFLUGCS catalog. It consists of the 63 X-ray-brightest clusters and is a strictly flux-limited survey, with $f_{\rm X}(0.1\text{-}2.4\,{\rm keV}) \geq 2.0 \times 10^{-11}\,{\rm erg\,s^{-1}\,cm^{-2}}$; it excludes the Galactic plane, $|b| \leq 20°$, as well as other regions around the Magellanic clouds and the Virgo cluster of galaxies in order to avoid large column densities of Galactic gas which lead to absorption, as well as Galactic and other nearby X-ray sources. The extended HIFLUGCS survey contains, in addition, several other clusters for which good measurements of the brightness profile and the X-ray temperature are available.

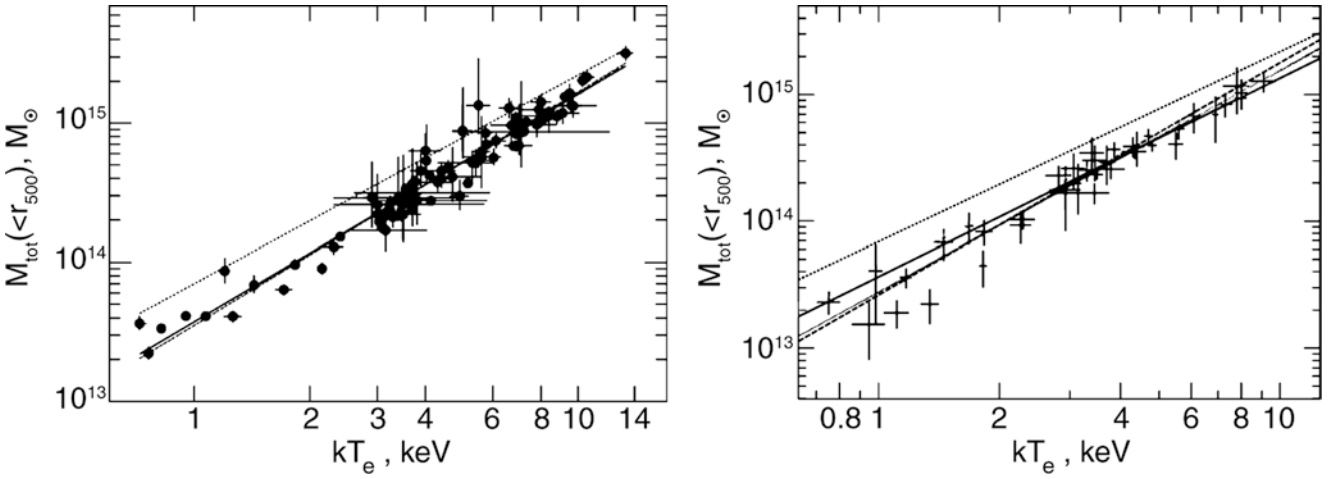


Fig. 6.43 For the clusters of galaxies from the extended HIFLUGCS sample, the mass within a mean overdensity of 500 is plotted as a function of X-ray temperature, where a dimensionless Hubble constant of $h = 0.5$ was assumed. In the *left-hand panel*, the mass was determined by applying an isothermal β-model, while in the *right-hand panel*, the radial temperature profile $T(r)$ was used to determine the mass, by means of (6.37). Most of the temperature measurements are from observations by the ASCA satellite. The *solid* and *dash-dotted curves* in the *left-hand panel* show the best fit to the data, where for the latter only the clusters from the original HIFLUGCS sample were used. In the *right-hand panel*, the *dashed line* is a fit to all the data in the plot, while the *solid line* takes into account only clusters with a mass $\geq 5 \times 10^{13} M_\odot$. In both panels, the *upper dotted line* shows the mass-temperature relation that was obtained from a simulation using simplified gas dynamics—the slope agrees with that found from the observations, but the amplitude is significantly too high. Source: A. Finoguenov et al. 2001, *Details of the mass-temperature relation for clusters of galaxies*, A&A 368, 749, p. 752, Figs. 1, 2. ©ESO. Reproduced with permission

based on an isothermal β-model, and from a cluster sample in which the measured radial temperature profile $T(r)$ was utilized in the mass determination [see (6.37)]. Constraining the sample to clusters with temperatures above 3 keV, one obtains a slope of 1.48 ± 0.1, in excellent agreement with theoretical expectations. Considerably steeper mass-temperature relations result from the inclusion of galaxy groups into the sample, from which we conclude that they do not follow in detail the scaling argument sketched above.

The X-ray temperature of galaxy clusters apparently provides a very accurate measure for their virial mass, better than the velocity dispersion (see below).

6.5.2 Mass-velocity dispersion relation

The velocity dispersion of the galaxies in a cluster also can be related to the mass: from (6.26) we find

$$M_{\rm vir} = \frac{3 r_{\rm vir} \sigma_v^2}{G} \,. \tag{6.54}$$

Together with $T \propto \sigma_v^2$ and $T \propto r_{\rm vir}^2$, it then follows that

$$\boxed{M_{\rm vir} \propto \sigma_v^3} \,. \tag{6.55}$$

This relation can now be tested using clusters for which the mass was determined using the X-ray method, and for which measurements of the velocity dispersion of the cluster galaxies are available. Alternatively, the relation $T \propto \sigma_v^2$ can be tested. One finds that these relations are essentially satisfied for the observed clusters. However, the relation between σ_v and M is not as tight as the M-T relation. Furthermore, numerous clusters exist which strongly deviate from this relation. These are clusters of galaxies that are not relaxed, as can be deduced from the velocity distribution of the cluster galaxies (which strongly deviates from a Maxwell distribution in these cases) or from a bimodal or even more complex galaxy distribution in the cluster. These outliers need to be identified, and removed, if one intends to apply the scaling relation between mass and velocity dispersion.

6.5.3 Mass-luminosity relation

The total X-ray luminosity that is emitted via bremsstrahlung is proportional to the squared gas density and the gas volume, hence it should behave as

$$L_{\rm X} \propto \rho_{\rm g}^2 \, T^{1/2} \, r_{\rm vir}^3 \propto \rho_{\rm g}^2 \, T^{1/2} \, M_{\rm vir} \,. \tag{6.56}$$

Estimating the gas density through $\rho_{\rm g} \sim M_{\rm g} \, r_{\rm vir}^{-3} = f_{\rm g} \, M_{\rm vir} \, r_{\rm vir}^{-3}$, where $f_{\rm g} = M_{\rm g}/M_{\rm vir}$ denotes the gas fraction

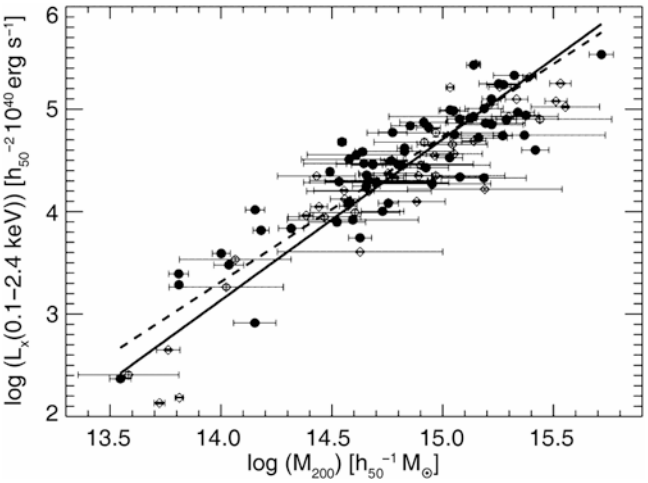


Fig. 6.44 For the galaxy clusters in the extended HIFLUGCS sample, the X-ray luminosity in the energy range of the ROSAT satellite is plotted versus the mass of the cluster. The solid points show the clusters of the HIFLUGCS sample proper. For the full sample and for the main HIFLUGCS sample, a best-fit power law is indicated by the *solid line* and *dashed line*, respectively. Source: T. Reiprich & H. Böhringer 2002, *The Mass Function of an X-Ray Flux-limited Sample of Galaxy Clusters*, ApJ 567, 716, p. 726, Fig. 6. ©AAS. Reproduced with permission

with respect to the total mass of the cluster, and using (6.52), we obtain

$$L_{\rm X} \propto f_{\rm g}^2 \, M_{\rm vir}^{4/3} \, . \tag{6.57}$$

This relation needs to be modified if the X-ray luminosity is measured within a fixed energy interval. Particularly for observations with ROSAT, which could only measure low-energy photons (below 2.4 keV), the received photons from massive clusters typically had $E_\gamma < k_{\rm B} T$, so that the measured X-ray luminosity becomes independent of T. Hence, one expects a modified scaling relation between the X-ray luminosity measured by ROSAT $L_{<2.4\,\rm keV}$ and the mass of the cluster,

$$L_{<2.4\,\rm keV} \propto f_{\rm g}^2 \, M_{\rm vir} \, . \tag{6.58}$$

This scaling relation can also be tested empirically, as shown in Fig. 6.44, where the X-ray luminosity in the energy range of the ROSAT satellite is plotted against the virial mass. One can immediately see that clusters of galaxies indeed show a strong correlation between luminosity and mass, but with a clearly larger scatter than in the mass-temperature relation.[9] Therefore, the temperature of the intergalactic gas is a better mass indicator than the X-ray luminosity or the velocity dispersion of the cluster galaxies.

[9]It should be noted, though, that the determination of $L_{\rm X}$ and M are independent of each other, whereas in the mass determination the temperature is an explicit parameter so that the measurements of these two parameters are correlated.

However, determining the slope of the relation from the data approximately yields $L_{<2.4\,\rm keV} \propto M^{1.5}$, instead of the expected behavior ($L_{<2.4\,\rm keV} \propto M^{1.0}$). Obviously, the above scaling arguments are not valid with the assumption of a constant gas fraction. This discrepancy between theoretical expectations and observations has been found in several samples of galaxy clusters and is considered well established. An explanation is found in models where the intergalactic gas has been heated not only by gravitational infall into the potential well of the cluster. Other sources of heating may have been present or still are. For cooler, less massive clusters, this additional heating should have a larger effect than for the very massive ones, which could also explain the deviation of low-M clusters from the mass-luminosity relation of massive clusters visible in Fig. 6.44. As already argued in the discussion of cooling flows in Sect. 6.4.3, an AGN in the inner regions of the cluster may provide such a heating. The kinetic energy provided by supernovae in cluster galaxies is also considered a potential source of additional heating of the intergalactic gas. Furthermore, the normalization of the luminosity-mass relation appears to be significantly different for cool-core clusters than for others, which supports the idea that heating affects this relation. It is obvious that solving this mystery will provide us with better insights into the formation and evolution of the gas component in clusters of galaxies.

Despite this discrepancy between the simple models and the observations, Fig. 6.44 shows a clear correlation between mass and luminosity, which can thus be used empirically after having been calibrated. Although the temperature is the preferred measure for a cluster's mass, one will in many cases resort to the relation between mass and X-ray luminosity because determining the luminosity (in a fixed energy range) is considerably simpler than measuring the temperature, for which significantly better photon statistics, i.e., longer exposure times are required.

6.5.4 The Y-parameter

The foregoing scaling relations connect the cluster mass with an observable. We have seen that the relation between cluster mass and gas temperature yields the smallest dispersion; hence, the X-ray temperature provides the best mass proxy of those that we considered above.

More recently, an alternative quantity for estimating the cluster mass was introduced. We saw in Sect. 6.4.4 above that the integrated SZ-effect is proportional to $Y = M_{\rm g} T_{\rm g}$, where is case of non-isothermal temperature distribution $T_{\rm g}$ needs to be interpreted as density-weighted temperature. In analogy, one considers

$$Y_{\rm X} := T_{\rm X} \, M_{\rm g} \, , \tag{6.59}$$

which is related to the thermal energy of gas in the cluster. The gas mass is related to the total mass by $M_g = f_g M$, where f_g is the gas-mass fraction. Assuming this to be constant, and using (6.52), we expect a scaling relation of the form

$$Y_X \propto M^{5/3} \,. \tag{6.60}$$

Numerical simulations including hydrodynamics first suggested that clusters should have a smaller scatter around the Y_X-M relation than for the other scaling relations. From X-ray observations, one indeed finds that clusters obey a relation very close to (6.60), with a slightly different slope, $M \propto Y^{0.55}$ (instead of 0.6). Furthermore, the scatter around the Y_X-M relation is smaller than that around the T_X-M relation, and in particular, at least for relaxed clusters, it is well described by a power-law behavior. Hence, Y_X provides a very useful mass proxy.

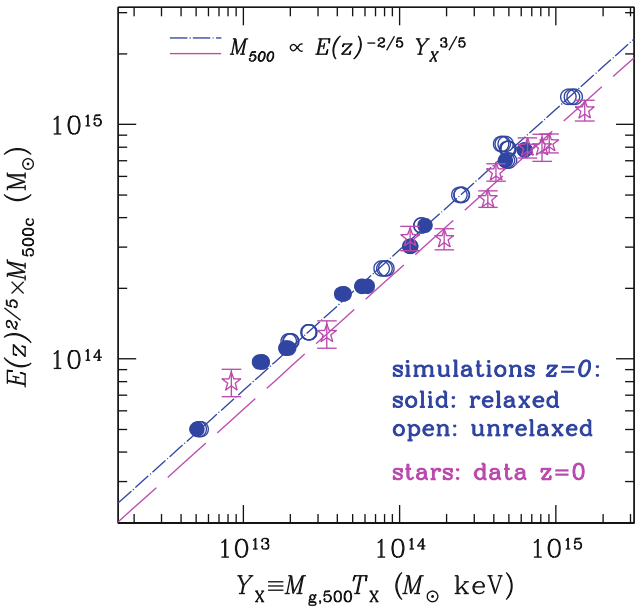


Fig. 6.45 The relation between the cluster mass M_{500} and the Y_X parameter. For this figure, the central region of the clusters has been cut out; specifically, the X-ray temperature was measured in the annulus $0.15 r_{500} \le r \le r_{500}$. The *blue* points show results from simulated clusters, where *filled dots* correspond to relaxed (cool-core) clusters, *open symbols* to non-relaxed clusters. It can be seen that both types of simulated clusters follow the same power law between M and Y_X. A sample of observed relaxed cluster is shown as *star symbols*; since these clusters are not at redshift zero, as the simulated ones, a redshift correction was applied to their masses (see Sect. 6.5.5 below). The observed clusters follow a very similar power law, though with a $\sim$ 15 % smaller normalization. Source: A.V. Kravtsov et al. 2006, *A New Robust Low-Scatter X-Ray Mass Indicator for Clusters of Galaxies*, ApJ 650, 128, p. 135, Fig. 7. ©AAS. Reproduced with permission

Center-excised scaling relations. We have seen that one can roughly divide clusters into cool-core clusters, which show a strong peak in the X-ray emission towards the center, and non-cool core clusters, which are presumably dynamically disturbed. At first sight, one would not expect that both types of clusters follow the same scaling relations. Indeed, not only the temperature profile of non-cool core clusters is different from that of cool-core clusters, but as a consequence also their pressure profiles, which is the relevant quantity for the SZ effect [see (6.47)]. Furthermore, we have seen that feedback processes from a central AGN affect the intracluster medium, causing (strong) deviations from spherical symmetry and locally varying temperature. Hence, we would expect that the central regions of clusters do not follow approximately some 'universal' behavior or a corresponding scaling relation.

However, if this central region is excised, the shape of the pressure profiles of clusters closely follow a universal profile, out to a radius of r_{500} (see Fig. 6.45); for larger radii, observations are difficult with the current X-ray observatories, due to the high background and the fact that the X-ray surface brightness decreases steeply for large radii. Hence, once the central region is cut out in the analysis, clusters seem to form a relatively homogeneous class of objects, despite all potential complications of hydrodynamic effects. It is for this reason why clusters can be used for cosmological studies, as we shall describe in Sect. 8.2.

6.5.5 Redshift dependence of scaling relations

The foregoing relations between the quantities of clusters were all obtained for very low-redshift objects. When considering clusters at higher redshift, one might expect that these scaling relations will evolve.

As mentioned before, the virial radius of a cluster is defined by the mean density inside of it being larger than the critical density by about a factor of $\Delta_c \sim 200$. The critical density is a function of redshift, $\rho_{cr}(z) = 3H^2(z)/(8\pi G) = E^2(z)\,\rho_{cr,0}$, where $\rho_{cr,0}$ is the critical density of the current universe [see (4.15)], and $E(z) = H(z)/H_0$. Since ρ_{cr} is larger at higher redshifts, the mean mass density inside clusters also increases, $\langle\rho\rangle = \Delta_c \rho_{cr}(z) \propto E^2(z)$. This affects also the scaling of the characteristic temperature of the cluster,

$$T \propto \frac{M_{vir}}{r_{vir}} = M_{vir}^{2/3} \left(\frac{M_{vir}}{r_{vir}^3}\right)^{1/3} \propto M_{vir}^{2/3} \langle\rho\rangle^{1/3} \propto [M_{vir} E(z)]^{2/3} \,. \tag{6.61}$$

The bolometric X-ray luminosity, assumed to be due to bremsstrahlung only, is the volume integral over the

emissivity (6.32), and so scales like

$$L_{\rm bol} \propto r_{\rm vir}^3 \left\langle n_{\rm e}^2 \right\rangle T^{1/2} \ ,$$

where we assumed a constant temperature of the intracluster gas. If the gas fraction of clusters is a constant, i.e., independent of mass and redshift, then we expect $r_{\rm vir}^3 \left\langle n_{\rm e}^2 \right\rangle \propto M_{\rm vir} \left\langle n_{\rm e}^2 \right\rangle / \langle \rho \rangle \propto M_{\rm vir} \left(\left\langle \rho^2 \right\rangle / \langle \rho \rangle^2 \right) \langle \rho \rangle \propto M_{\rm vir} E^2$, assuming the clumping factor $\left\langle \rho^2 \right\rangle / \langle \rho \rangle^2$ to stay constant. Together with (6.61), we then obtain the expected scaling of the luminosity,

$$\frac{L_{\rm bol}}{E(z)} \propto \left[M_{\rm vir} \, E(z) \right]^{4/3} \ . \tag{6.62}$$

Finally, for the Y-parameter, we have $Y = M_{\rm gas} T \propto M_{\rm vir} \left[M_{\rm vir} \, E(z) \right]^{2/3}$, where we again assumed a constant gas fraction and used (6.61); thus,

$$Y \, E(z) \propto \left[M_{\rm vir} \, E(z) \right]^{5/3} \ . \tag{6.63}$$

These relation generalize (6.52), (6.57) and (6.60) to higher redshifts (note that this redshift dependence was included in Fig. 6.45). If we assume that the profiles of clusters are self-similar, i.e., their density and temperature profiles all have the same shape, but different amplitude and scale length, then these relation do not only apply to the virial radius, but also for regions with different density thresholds.

In detail, these scaling relations are observed to be somewhat different. Whereas the redshift scaling in the mass-temperature relation (6.61) seems to apply to high-redshift clusters, the redshift scaling in the mass-luminosity relation (6.62) is not supported by observations; instead, the redshift evolution is observed to be slower than predicted by the self-similar model. Finally, Fig. 6.46 shows the mass-luminosity relation, where the mass determination was done via weak lensing analysis. Whereas there is a very tight correlation between luminosity and mass, the slope of the best-fitting power law is flatter than predicted by the self-similar behavior (6.62), which probably indicates that clusters are not truly self-similar.

6.5.6 Near-infrared luminosity as mass indicator

Whereas the optical luminosity of galaxies depends not only on the mass of the stars but also on the star formation history, the NIR light is much less dependent on the latter. As we have discussed before, the NIR luminosity is thus quite a reliable measure of the total mass in stars. For this reason, we would expect that the NIR luminosity of a cluster is tightly correlated with its total stellar mass. Furthermore, if the latter is closely related to the total cluster mass, as would be the case if the stellar mass is a fixed fraction of the cluster's total mass, the NIR luminosity can be used to estimate the masses of clusters.

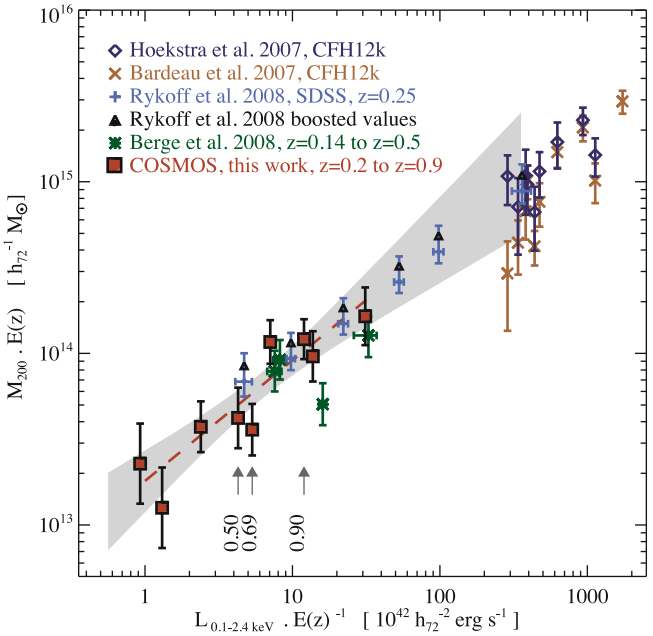


Fig. 6.46 The cluster mass-X-ray luminosity relation, where the mass was determined with weak gravitational lensing methods. *Dark blue diamonds* and *sienna crosses* indicate mass measurements of individual clusters. This is feasible only for high-mass objects. The other data points indicate mean masses of sets of clusters within given L-bins. In particular, the *red squares* were obtained from a weak lensing analysis of X-ray clusters in the COSMOS field, for which the three highest redshift bins are explicitly indicated. The *shaded region* indicates the range which power-law fits within the 1-σ range to the data would cover. Source: A. Leauthaud et al. 2010, *A Weak Lensing Study of X-ray Groups in the Cosmos Survey: Form and Evolution of the Mass-Luminosity Relation*, ApJ 709, 97, p. 109, Fig. 6. ©AAS. Reproduced with permission

The Two Micron All Sky Survey (2MASS) provides the first opportunity to perform such an analysis on a large sample of galaxy clusters. One selects clusters of galaxies for which masses were determined by X-ray methods, and then measures the K-band luminosities of the galaxies within the cluster. Figure 6.47 presents the resulting mass-luminosity diagram within r_{500} for 93 galaxy clusters and groups, where the mass was derived from the clusters' X-ray temperatures (plotted on the top axis) by means of (6.53). A surprisingly close relation between these two parameters is seen, which can be described by a power law of the form

$$\frac{L_{500}}{10^{12} L_\odot} = 3.95 \left(\frac{M_{500}}{2 \times 10^{14} M_\odot} \right)^{0.69} \ , \tag{6.64}$$

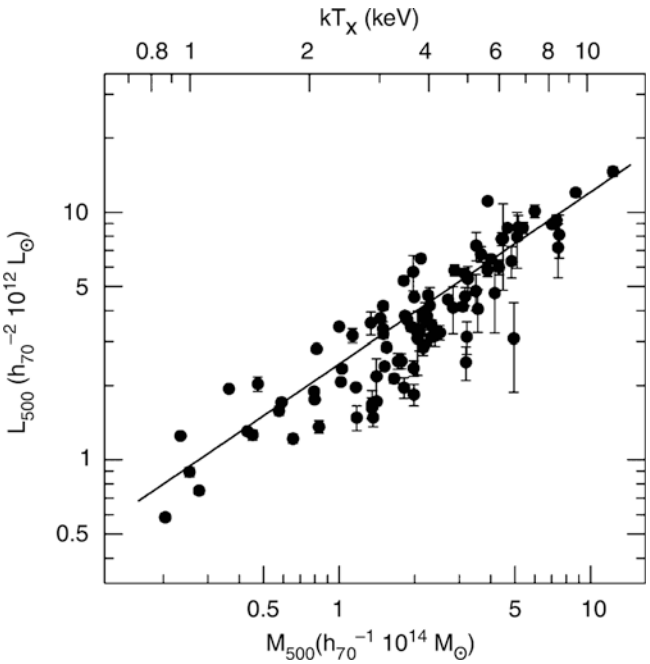


Fig. 6.47 The correlation between K-band luminosity and the mass of galaxy clusters, measured within the radius inside of which the mean density is 500 times the critical density of the Universe. The cluster mass was determined by the relation (6.53) between mass and temperature. Source: Y.-T. Lin et al. 2004, *K-Band Properties of Galaxy Clusters and Groups: Luminosity Function, Radial Distribution, and Halo Occupation Number*, ApJ 610, 745, p. 753, Fig. 3. ©AAS. Reproduced with permission

where a Hubble constant of $h = 0.7$ is assumed. The dispersion of individual clusters around this power law is about 32 %, where at least part of this scatter originates in uncertainties in the mass determination—thus, the intrinsic scatter is even smaller. This result is of great potential importance for future studies of galaxy clusters, and it renders the NIR luminosity a competitive method for the determination of cluster masses, which is of considerable interest in view of the current and future generation of NIR wide-field instruments (like VISTA on Paranal, for instance).

6.6 Clusters of galaxies as gravitational lenses

6.6.1 Luminous arcs

In 1986, two groups independently discovered unusually stretched, arc-shaped sources in two clusters of galaxies at high redshift (see Figs. 6.48 and 6.49). The nature of these sources was unknown at first; they were named *arcs*, or *giant luminous arcs*, which did not imply any interpretation originally. Different hypotheses for the origin of these arcs were formulated, like for instance emission by shock fronts in the ICM, originating from explosive events. All these scenarios were disproven when the spectroscopy of the arc in the cluster Abell 370 showed that the source is at a much higher redshift than the cluster itself. Thus, the arc is a background source, subject to the gravitational lens effect (see Sect. 3.11) of the cluster. By differential light deflection, the light beam of the source can be distorted in such a way that highly elongated arc-shaped images are produced.

The discovery that clusters of galaxies may act as strong gravitational lenses came as a surprise at that time. Based on the knowledge about the mass distribution of clusters, derived from X-ray observations before ROSAT, it was estimated that the central surface mass density of clusters is not sufficiently high for strong effects of gravitational light deflection to occur. This incorrect estimate of the central surface mass density in clusters originated from analyses utilizing the β-model which, as briefly discussed above, starts with some heavily simplifying assumptions.[10]

Hence, arcs are strongly distorted and highly magnified images of galaxies at high redshift. In some massive clusters several arcs were discovered and the unique angular resolution of the HST played a crucial role in such observations. Some of these arcs are so thin that their width is unresolved even by the HST, indicating an extreme length-to-width ratio. For many arcs, additional images of the same source were discovered, sometimes called 'counter arcs'. The identification of multiple images is performed either by optical spectroscopy (which is difficult in general, because one arc is highly magnified while the other images of the same source are considerably less strongly magnified and therefore much fainter in general, together with the fact that spectroscopy of faint sources is very time-consuming), by multi-color photometry (all images of the same source should have the same color), or by common morphological properties.

Lens models. Once again, the simplest mass model for a galaxy cluster as a lens is the singular isothermal sphere (SIS). This lens model was discussed previously in Sect. 3.11.2. Its characteristic angular scale is specified by

[10] Another lesson that can be learned from the discovery of the arcs is one regarding the psychology of researchers. After the first observations of arcs were published, several astronomers took a second look at their own images of these two clusters and clearly detected the arcs in them. The reason why this phenomenon, which had been observed much earlier, was not published before can be explained by the fact that researchers were not completely sure about whether these sources were real. A certain tendency prevails in not recognizing phenomena that occur unexpectedly in data as readily as results which are expected. However, there are also researchers who behave in exactly the opposite manner and even interpret phenomena expected from theory in some unusual way.

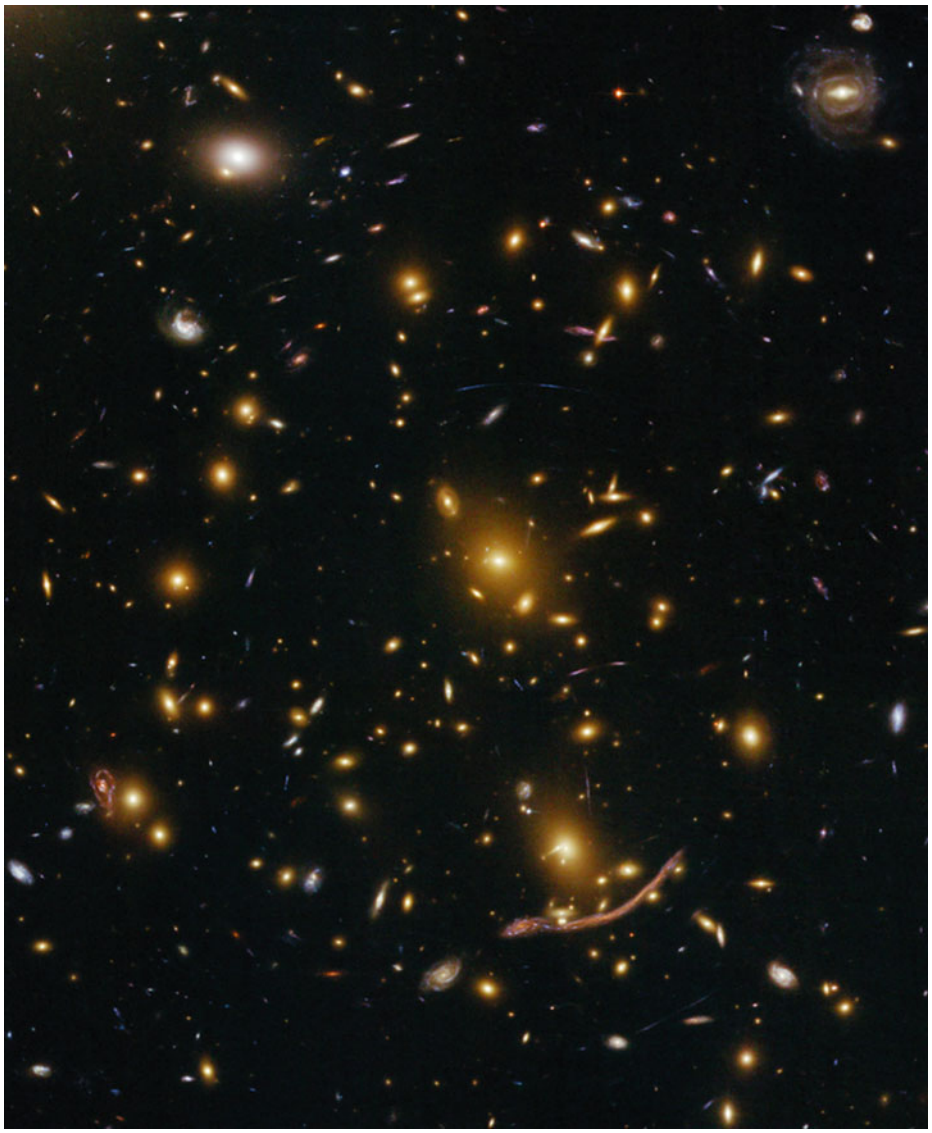


Fig. 6.48 The cluster of galaxies A 370 at redshift $z = 0.375$ is one of the first two clusters in which giant luminous arcs were found in 1986. In this HST image, the arc is clearly visible; it is about 20″ long, tangentially oriented with respect to the center of the cluster which is located roughly halfway between the two brightest cluster galaxies, and curved towards the center of the cluster. The arc is the image of a galaxy at $z_s = 0.724$. Several other very thin arcs are clearly seen, many of which appear in close pairs; in these cases, they are multiple images of the same source galaxy. Only with HST images was it realized how thin these arcs are. One also notes the presence of a 'radial arc', not far from the bright galaxy closest to the major arc. Credit: NASA, ESA, the Hubble SM4 ERO Team, and ST-ECF

the Einstein radius (3.75), or

$$\theta_E = 28''.8 \left(\frac{\sigma_v}{1000\,\mathrm{km/s}} \right)^2 \left(\frac{D_{ds}}{D_s} \right) . \qquad (6.65)$$

Very high magnifications and distortions of images can occur only very close to the Einstein radius. This immediately yields an initial mass estimate of a cluster, by assuming that the Einstein radius is about the same as the angular separation of the arc from the center of the cluster. The projected mass within the Einstein radius can then be derived, using (3.81). Since clusters of galaxies are, in general, not spherically symmetric and may show significant substructure, so that the separation of the arc from the cluster center may deviate significantly from the Einstein radius, this mass estimate is not very accurate in general; the uncertainty is presumably $\sim 30\,\%$. Models with asymmetric mass distributions predict a variety of possible morphologies for the arcs and the positions of multiple images, as is demonstrated in Fig. 6.50 for an elliptical lens. If several arcs are discovered in a cluster, or several images of the source of an arc, we can investigate detailed mass models for such a cluster. The accuracy of these models depends on the number and positions of the observed lensed images; e.g., on how many arcs and how many multiple image systems are available for modeling. The resulting mass models are not unambiguous, but they are robust. Clusters that contain many lensed images have very well-determined mass properties, for instance the mass and the mass profile within the radii at which arcs are found, or the ellipticity of the mass distribution and its substructure.

Figure 6.51 shows two clusters of galaxies which contain several arcs. The lensing mass estimate of the cluster Cl 0024+17 is quite different from the mass obtained through an X-ray analysis. Spectroscopy of its member galaxies show

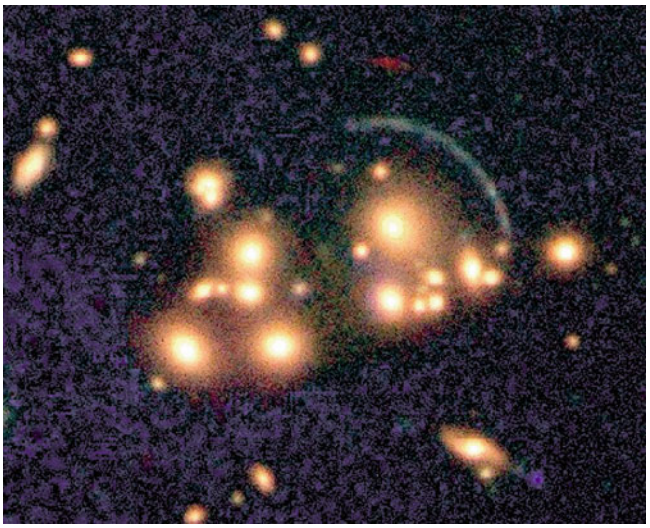


Fig. 6.49 The cluster of galaxies Cl 2244−02 at redshift $z = 0.33$ is the second cluster in which an arc was discovered. Spectroscopic analysis of this arc revealed the redshift of the corresponding source to be $z_s = 2.24$—at the time of discovery in 1987, it was the first normal galaxy detected at a redshift > 2. This image was observed with the near-IR camera ISAAC at the VLT. Above the arc, one can see another strongly elongated red source which is probably associated with a galaxy at very high redshift as well. Credit: European Southern Observatory

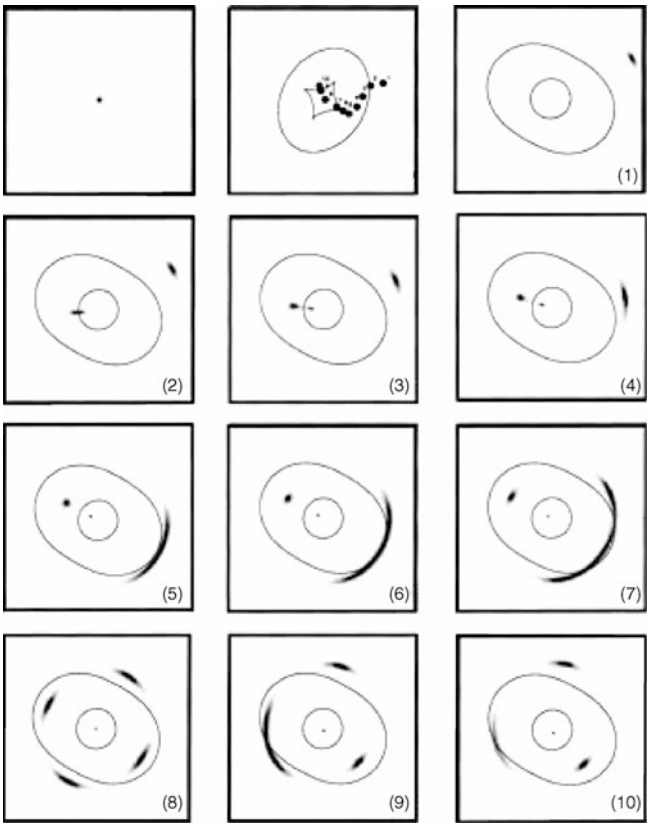


Fig. 6.50 Distortions by the lens effect of an elliptical potential, as a function of the source position. The *first panel* shows the source itself. The *second panel* displays ten positions of the source in the source plane (numbered from 1 to 10) relative to the center of the lens; the *solid curves* show the inner and outer caustics. The remaining panels (numbered from 1 to 10) show the inner and outer critical curves and the resulting images of the source. Source: B. Fort & Y. Mellier 1994, *Arc(let)s in clusters of galaxies*, A&AR 5, 239. © Springer-Verlag. Reproduced with permission

evidence for a bimodal distribution in velocity space which is interpreted as the cluster being indeed a pair of clusters, colliding along the line-of-sight. If this interpretation is correct, the mass estimate from X-rays is expected to yield discordant values.

For a long time, A 2218 was the classic example of the existence of numerous arcs in a single galaxy cluster. Then after the installation of the ACS camera on-board HST in 2002, a spectacular image of the cluster A 1689 was obtained in which more than 100 arcs and multiple images were identified (see Fig. 6.52). Several zoomed sections of this image are shown in the lower part of the figure. For clusters of galaxies with such a rich inventory of lens phenomena, very detailed mass models can be constructed.

Such mass models have predictive power, allowing an iterative modeling process. An initial simple mass model is fitted to the most prominent lensed images in the observation, i.e., either giant arcs or clearly recognizable multiple images. In general, this model then predicts further images of the source producing the arc. Close to these predicted positions, these additional images are then searched for, utilizing the morphology of the light distribution and the color. If this initial model describes the overall mass distribution quite well, such images are found. The exact positions of the new images provide further constraints on the lens model which is then refined accordingly. Again, the new model will predict further multiple image systems, and so on. By this procedure, very detailed models can sometimes be obtained. Since the lens properties of a cluster depend on the distance or the redshift of the source, the redshift of lensed sources can be predicted from the identification of multiple image systems in clusters if a detailed mass model is available. These predictions can then be verified by spectroscopic analysis, and the success of this method gives us confidence in the accuracy of the lens models.

Results. We can summarize the most important results of the examination of clusters using arcs and multiple images as follows: the mass of galaxy clusters is indeed much larger than the mass of their luminous matter. The lensing method yields a mass which is in very good agreement with mass estimates from the X-ray method or from dynamical methods. However, the core radius of clusters, i.e., the scale on which the mass profile flattens inwards, is significantly smaller than determined from X-ray observations. If the mass distribution in clusters has a core, its size is estimated to be $r_c \sim 30h^{-1}$ kpc, in contrast to $\sim 150h^{-1}$ kpc obtained from

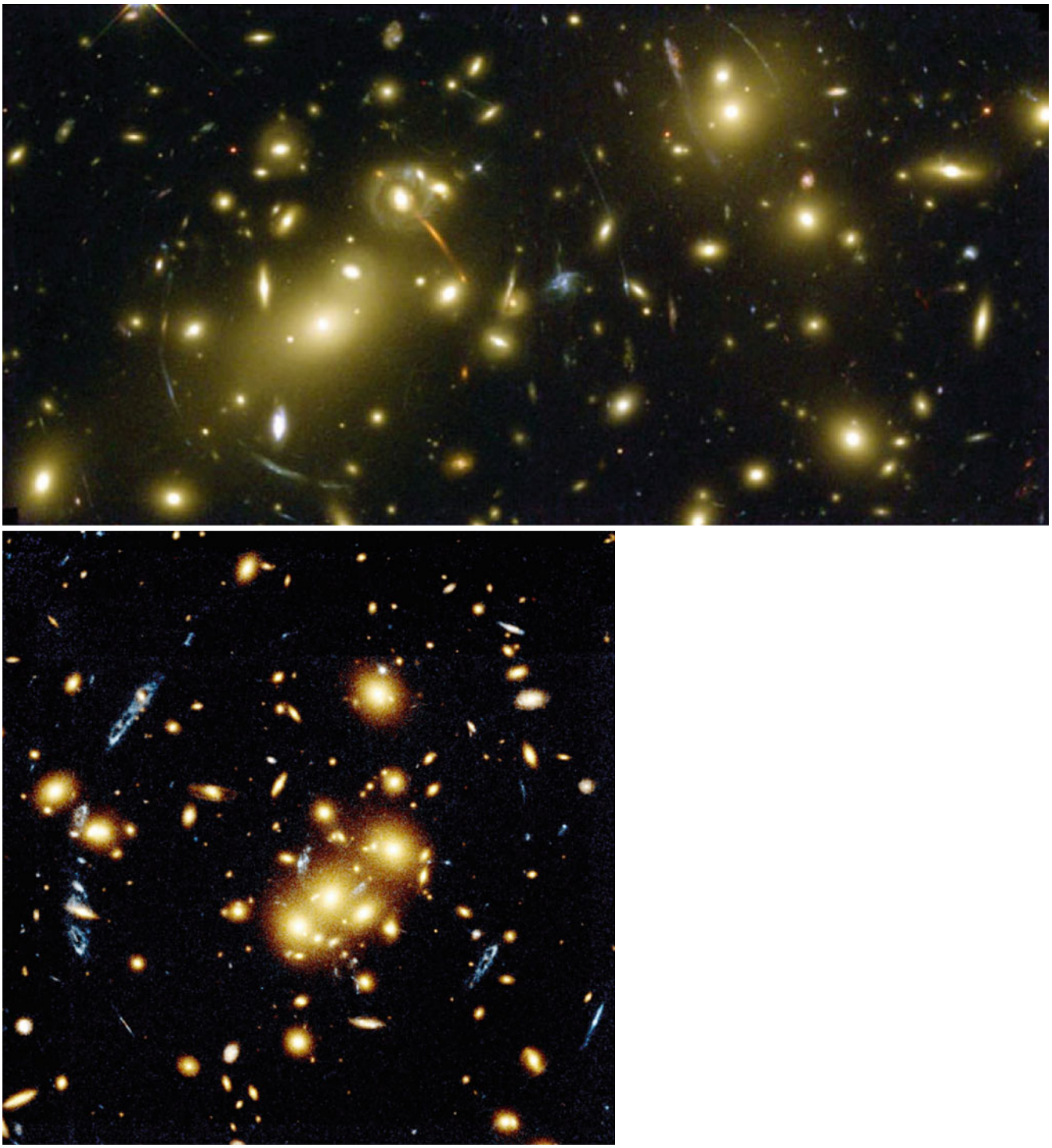


Fig. 6.51 *Top*: the cluster of galaxies A 2218 ($z_d = 0.175$) contains one of the most spectacular arc systems. The majority of the galaxies visible in the image are associated with the cluster, and the redshifts of many of the strongly distorted arcs have now been measured. *Bottom*: the cluster of galaxies Cl 0024+17 ($z_d = 0.39$) contains a rich system of arcs. The arcs appear bluish, stretched in a direction which is tangential to the cluster center. The three arcs to the left of the cluster center, and the arc to the right of it and closer to the center, are images of the same background galaxy which has a redshift of $z_s = 1.62$. Another image of the same source was found close to the cluster center. Also note the identical ('pretzel'-shaped) morphology of the images. Credit: *Top*: W. Couch/University of New South Wales, R.S. Ellis/Cambridge University and NASA. *Bottom*: W.N. Colley and E. Turner/Princeton University, J.A. Tyson/Bell Labs, Lucent Technologies, and NASA

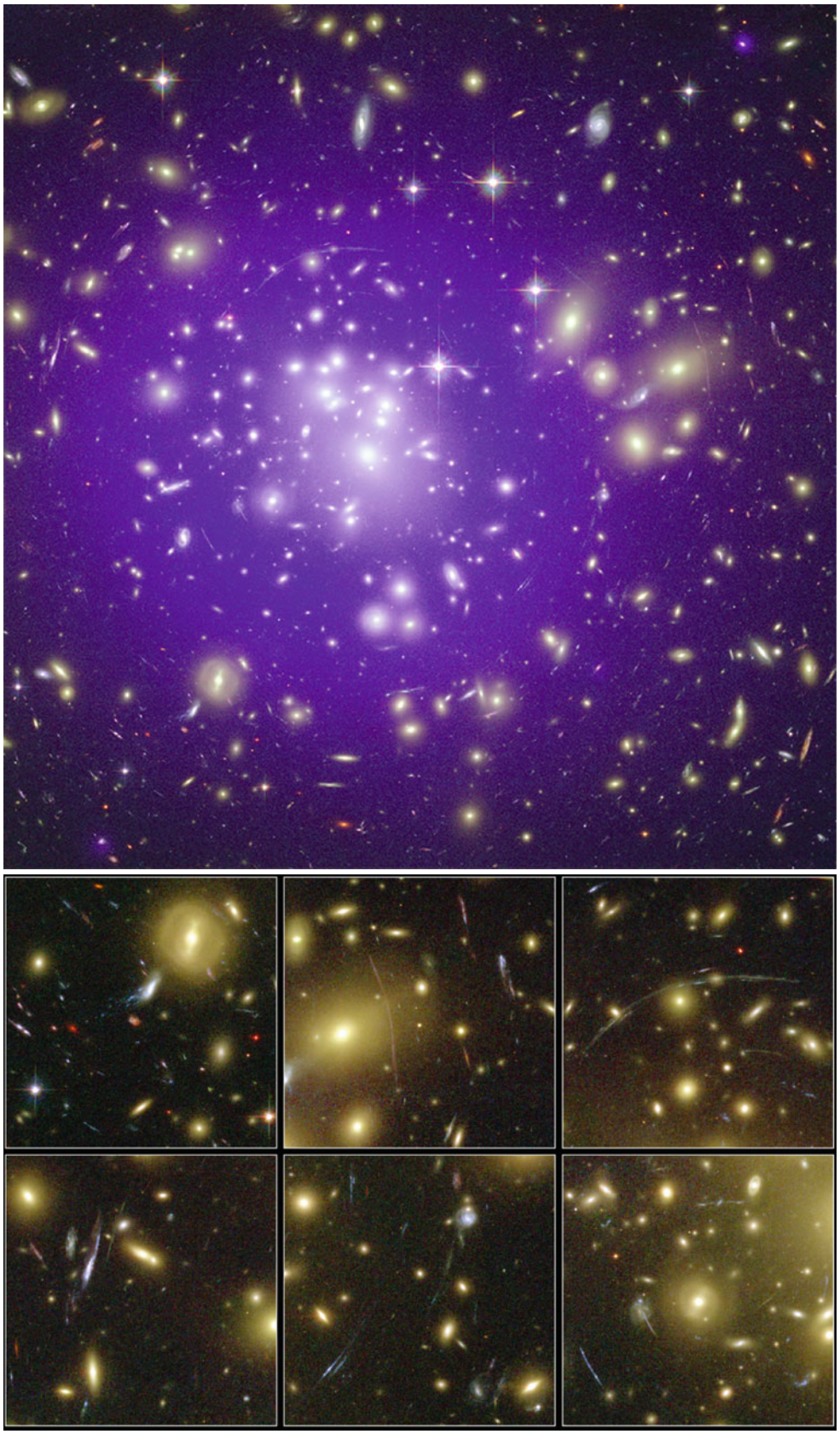


Fig. 6.52 The cluster of galaxies Abell 1689 has one of the richest system of arcs and multiple images found to date. In a deep ACS exposure of this cluster, shown *on top* with the X-ray emission superimposed in *blue*, more than a hundred such lensed images were detected. Six zoomed sections of this ACS image are shown in the *bottom part*, in which various arcs are visible, some with an extreme length-to-width ratio, indicating very high magnification factors. Credit: X-ray: NASA/CXC/MIT/E.-H. Peng et al.; optical: NASA, N. Benitez/JHU, T. Broadhurst/Racah Institute of Physics/The Hebrew University, H. Ford/JHU, M. Clampin/STScI, G. Hartig/STScI, G. Illingworth/UCO/Lick Observatory, the ACS Science Team and ESA

early X-ray analyses.[11] This difference leads to a discrepancy in the mass determination between the two methods on scales below $\sim 200h^{-1}$ kpc.

We emphasize that, at least in principle, the mass determination based on arcs and multiple images is substantially more accurate because it does not require any assumptions about the symmetry of the mass distribution, about hydrostatic equilibrium of the X-ray gas, or about an isothermal temperature distribution. On the other hand, the lens effect measures the mass in cylinders because the lens equation contains only the projected mass distribution, whereas the X-ray method determines the mass inside spheres. The conversion between the two methods introduces uncertainties, in particular for clusters which deviate significantly from spherical symmetry. Overestimating the core radius was the main reason why the discovery of the arcs was a surprise because clusters with core radii like the ones determined from the early X-ray measurements would in fact not act as strong gravitational lenses. Hence, the mere existence of arcs shows that the core radius must be small.

A closer analysis of galaxy clusters with a cool core shows that, in these clusters, the mass profile estimated from X-ray observations is compatible with the observed arcs. Such clusters are considered dynamically relaxed, so that for them the assumption of a hydrostatic equilibrium is well justified. The X-ray analysis has to account explicitly for the existence of an inner cool region, though, and the accordingly modified X-ray emission profile is more sophisticated than the simple β-model. Clusters without a cool core are distinctly more complex dynamically. Besides the discrepancy in mass determination, lensing and X-ray methods can lead to different estimates of the center of mass in such unrelaxed clusters, which may indicate that the gas has not had enough time since the last strong interaction or merging process to settle into an equilibrium state.

As we mentioned before, the cluster Cl 0024+17 in Fig. 6.51 most likely has a complex structure along the line-of-sight, and hence it is not surprising that X-ray estimated masses deviate significantly from those obtained by lensing. As a second example, the cluster A 1689 (Fig. 6.52) presumably also has a complex structure. A combined lensing, X-ray and SZ-study of this cluster suggests that it is highly elongated, with an axis ratio of ~ 2 and the long axis close to the line of sight. This orientation maximizes the lensing strength of a triaxial mass distribution, and clusters selected by their strong lensing features will be biased towards these orientations. In order to account for the differences in mass estimates from these different methods, one concludes that the gas is not in hydrostatic equilibrium, but there must be substantial fraction of non-thermal pressure in that cluster, e.g., from turbulent motions of the intracluster gas.

The mass distribution in clusters often shows significant substructure. Many clusters of galaxies in which arcs are observed are not relaxed. These clusters still undergo dynamical evolution—they are young systems with an age not much larger than $t_{\rm cross}$, or systems whose equilibrium was disturbed by a fairly recent merger process. For such clusters, the X-ray method is not well founded because the assumptions about symmetry and equilibrium are not satisfied. The distribution of arcs in the cluster A2218 (Fig. 6.51) clearly indicates a non-spherical mass distribution. Indeed, this cluster seems to consist of at least two massive components around which the arcs are curved, indicating that the cluster is currently undergoing a strong merging event. This is further supported by measurements of the temperature distribution of the intracluster gas, which shows a strong peak in the center, where the temperature is about a factor of two higher than in its surrounding region.

From lens models, we find that for clusters with a central cD galaxy, the orientation of the mass distribution follows that of the cD galaxy quite closely. We conclude from this result that the evolution of the brightest cluster galaxy must be closely linked to the evolution of the cluster, e.g., by accretion of a cooling flow onto the BCG or due to mergers with other cluster members. Often, the shape of the mass distribution very well resembles the galaxy distribution and the X-ray emission.

> The investigation of galaxy clusters with the gravitational lens method provides a third, completely independent method of determining cluster masses. It confirms that the mass of galaxy clusters significantly exceeds that of the visible matter in stars and in the intracluster gas. We conclude from this result that clusters of galaxies are dominated by dark matter.

6.6.2 The weak gravitational lens effect

The principle of the weak lensing effect. In Sect. 3.11 we saw that gravitational light deflection does not only deflect light beams as a whole, but also that the size and shape of light beams are distorted by differential light deflection. This differential light deflection leads, e.g., to sources appearing brighter than they would be without the lens effect. The giant arcs discussed above are a very clear example of these distortions and the corresponding magnifications.

If some background sources exist which are distorted in such an extreme way as to appear as giant luminous arcs, then it is plausible that many more background galaxies should exist which are less strongly distorted. Typically, these are

[11] We will see in the next chapter that the density of dark matter halos is predicted to increase right to the center, i.e., no core is formed in this case.

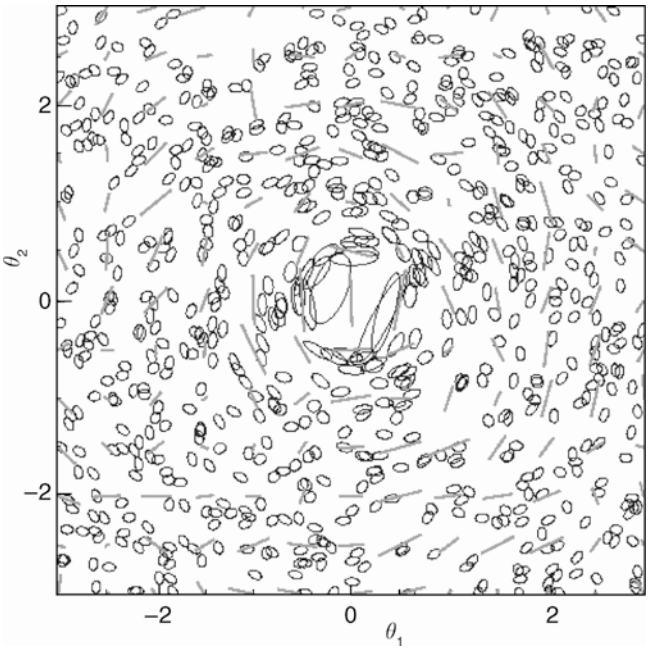


Fig. 6.53 The principle of the weak gravitational lensing effect is illustrated here with a simulation. Due to the tidal component of the gravitational field in a cluster, the shape of the images (ellipses) of background galaxies get distorted and, as for arcs, the galaxy images will be aligned, on average, tangentially to the cluster center. By local averaging over the ellipticities of galaxy images, a local estimate of the tidal gravitational field can be obtained (the direction of the sticks indicates the orientation of the tidal field, and their length is proportional to its strength). From this estimated tidal field, the projected mass distribution can then be reconstructed. Source: C. Seitz, *The determination of the mass distribution in clusters of galaxies by gravitationally distorted images of background galaxies*, Dissertation, LMU München, 1996

located at larger angular separations from the cluster center, where the lens effect is weaker than at the locations of the luminous arcs. Their distortion then is so weak that it cannot be identified in an individual galaxy image. The reason for this is that the intrinsic light distribution of galaxies is not circular; rather, the observed image shape is a superposition of the intrinsic shape and the gravitational lens distortion. The former is considerably larger than the latter, in general, and acts as a kind of noise in the measurement of the lensing effect. However, the distortion of adjacent galaxy images should be similar since the gravitational field their light beams are traversing is similar. By averaging over many such galaxy images, the distortion can then be measured (see Fig. 6.53) because no preferred direction exists in the intrinsic orientation of galaxies; it is expected to be random. After the results from the Hubble Deep Field (Fig. 1.37) became available, if not before, we have known that the sky is densely covered by small, faint galaxies. In deep optical images, one therefore finds a high number density of such galaxies located in the background of a galaxy cluster. Their measured shapes can be used for investigating the weak lensing effect of the cluster.

The distortion, obtained by averaging over image ellipticities, reflects the contribution of the tidal forces to the local gravitational field of the cluster. In this context, it is denoted as *shear*. It is given by the projection of the tidal component of the gravitational field along the line-of-sight. The shear results from the derivative of the deflection angle, where the deflection angle (3.62) depends linearly on the surface mass density of the lens. Hence, it is possible to reconstruct the surface mass density of galaxy clusters in a completely parameter-free way using the measured shear: it can be used to map the (total, i.e., dark plus luminous) matter in a cluster.

Observations. Since shear measurements are based on averaging over image ellipticities of distant galaxies, this method of *weak gravitational lensing* requires optical images with as high a galaxy density as possible. This implies that the exposures need to be very deep to reach very faint magnitudes. But since very faint galaxies are also very distant and, as a consequence, have small angular extent, the observations need to be carried out under very good observing conditions, to be able to accurately measure the shape of galaxy images without them being smeared into circular images by atmospheric turbulence, i.e., the seeing. Typically, to apply this method images from 4-m class telescopes are used, with exposure times of 1–3 h. This way, we reach a density of about 30 galaxies per square arcminute (thus, 10^5 per square degree) of which shapes are sufficiently well measurable. This corresponds to a limiting magnitude of about $R \sim 25$. The seeing during the exposure should not be larger than $\sim 0''.8$ to still be able to correct for effects of the point-spread function. The smaller the seeing, the more galaxy images can be used for the weak lensing analysis.

Systematic observations of the weak lensing effect only became feasible with the development of wide-field cameras.[12] This, together with the improvement of the dome seeing at many telescopes and the development of dedicated software for data analysis, rendered quantitative observational studies with weak lensing possible; the best telescopes at the best observatories regularly accomplish seeing below $0''.8$, and the dedicated software is specifically designed for measuring the shapes of extremely faint galaxy images and for correcting for the effects of seeing and anisotropy of the point spread function.

Mass reconstruction of galaxy clusters. By means of this method, the reconstruction of the mass density of a large

[12] Prominent examples for such cameras are Megacam at CFHT, the first square degree camera with $\sim (18000)^2$ pixels, and OmegaCAM (with the same field-of-view) at the VLT Survey Telescope on Paranal. The largest camera currently in operation is that of Pan-STARRS1, covering $\sim 6\,\mathrm{deg}^2$

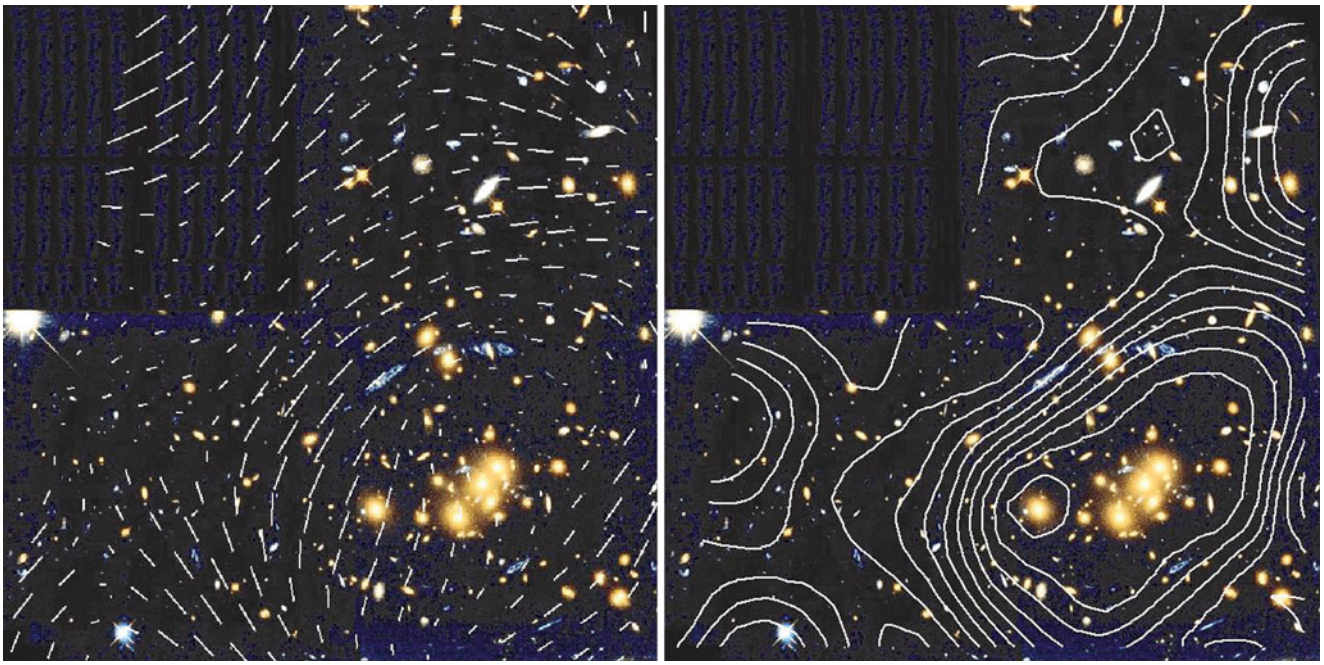


Fig. 6.54 *On the left*, the tidal (or shear) field of the cluster Cl 0024+17 (Fig. 6.51) is indicated by sticks whose length and direction represent the strength and orientation of the tidal gravitational field. *On the right*, the surface mass density is shown, reconstructed by means of the weak gravitational lens effect. The bright galaxies in the cluster are seen to follow the (dark) matter distribution; the orientation of the isodensity contours is the same as the orientation of the light in the center of the cluster. Credit: optical image: HST/NASA, Colley et al.; shear field and mass reconstruction: C. Seitz, *The determination of the mass distribution in clusters of galaxies by gravitationally distorted images of background galaxies*, Dissertation, LMU München, 1996

number of clusters became possible. The most important results of these investigations are as follows: the center of the mass distribution corresponds to the optical center of the cluster (see Fig. 6.54). If X-ray information is available, the mass distribution is, in general, found to be centered on the X-ray maximum, except in clusters which are heavily disturbed. The shape of the mass distribution—e.g., its ellipticity and orientation—is in most cases very similar to the distribution of bright cluster galaxies. The comparison of the mass profile determined by this method and that determined from X-ray data agree well, typically within a factor of ~ 1.5; we will discuss mass profiles of clusters from weak lensing further below (Sect. 7.7) where we compare them with cosmological predictions.

Through the weak lensing effect, substructure in the mass distribution is also detected in some clusters (Fig. 6.55) which does not in all cases reflect the distribution of cluster galaxies. However, in general a good correspondence between light and mass exists (Fig. 6.56). From these lensing studies, we obtain a mass-to-light ratio for clusters that agrees with that found from X-ray analyses, about $M/L \sim 250h$ in Solar units. Clusters of galaxies that strongly deviate from this average value do exist, however. Two independent analyses for the cluster MS 1224+20 resulted in a mass-to-light-ratio of $M/L \approx 800h$ in Solar units, more than twice the value normally found in clusters.

The similarity of the mass and galaxy distributions is not necessarily expected because the lens effect measures the total mass distribution, and therefore mainly the dark matter in a cluster of galaxies. The similar distributions then imply that the galaxies in a cluster seem to basically follow the distribution of the dark matter (Sect. 7.6.1), although there are some exceptions.

Bullet clusters, and the nature of dark matter. Figure 6.34 shows the X-ray emission and the galaxy distribution of the so-called Bullet cluster, actually a system of two clusters which have recently collided. Besides the hydrodynamic effects of the collision, this system is of great interest to test for the nature of the discrepancy between the observed mass in clusters (from the stars in the member galaxies and the intracluster gas) and the gravitational mass deduced from the X-ray emission, the galaxy peculiar velocities in clusters, and the lensing results. If this discrepancy is due to a modification of the law of gravity on large scales, as was speculated, then we expect that the bulk of the mass in the Bullet cluster is centered on the X-ray emitting gas, since its mass is several times higher than that of the stars in the cluster galaxies. On the other hand, if the discrepancy is due to (collisionless) dark matter, then this dark matter should behave similarly to the galaxies in a collision: the scattering cross section is very small, and

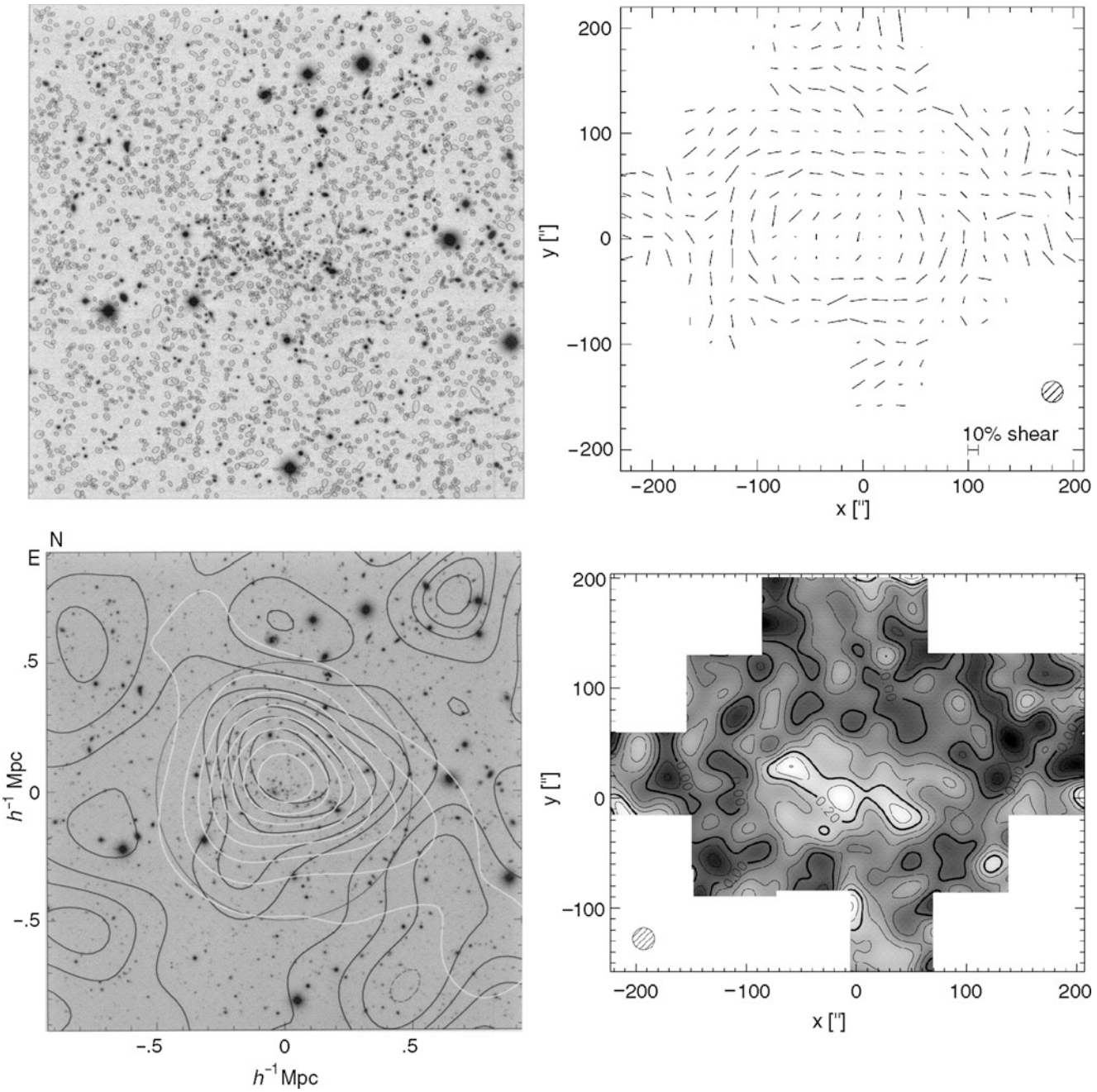


Fig. 6.55 Analysis of the cluster of galaxies MS 1054−03 by the weak lensing effect. In the *upper left panel*, a ground-based image is shown with a field size of 7′.5 × 7′.5. In this image, about 2400 faint objects are detected, the majority of which are galaxies at high redshift. From the measured ellipticities of the galaxies, the tidal field of the cluster can be reconstructed, and from this the projected mass distribution $\Sigma(\boldsymbol{\theta})$, presented in the *lower left panel*; the latter is indicated by the black contours, while the white contours represent the smoothed light distribution of the cluster galaxies. A mosaic of HST images allows the ellipticity measurement of a significantly larger number of galaxies, and with better accuracy. The tidal field resulting from these measurements is displayed in the *upper right panel*, with the reconstructed surface mass density shown in the *lower right panel*. One can clearly see that the cluster is strongly structured, with three density maxima which correspond to regions with bright cluster galaxies. This cluster seems to be currently in the process of formation through a merger of smaller entities. Source: *Left*: G.A. Luppino & N. Kaiser 1997, *Detection of Weak Lensing by a Cluster of Galaxies at* $z = 0.83$, ApJ 475, 20, PLATE 2, 3, Figs. 3, 5. ©AAS. Reproduced with permission. *Right*: H. Hoekstra et al. 2000, *Hubble Space Telescope Weak-Lensing Study of the* $z = 0.83$ *Cluster MS 1054–03*, ApJ 532, 88, p. 94, 96, Figs. 9, 12. ©AAS. Reproduced with permission

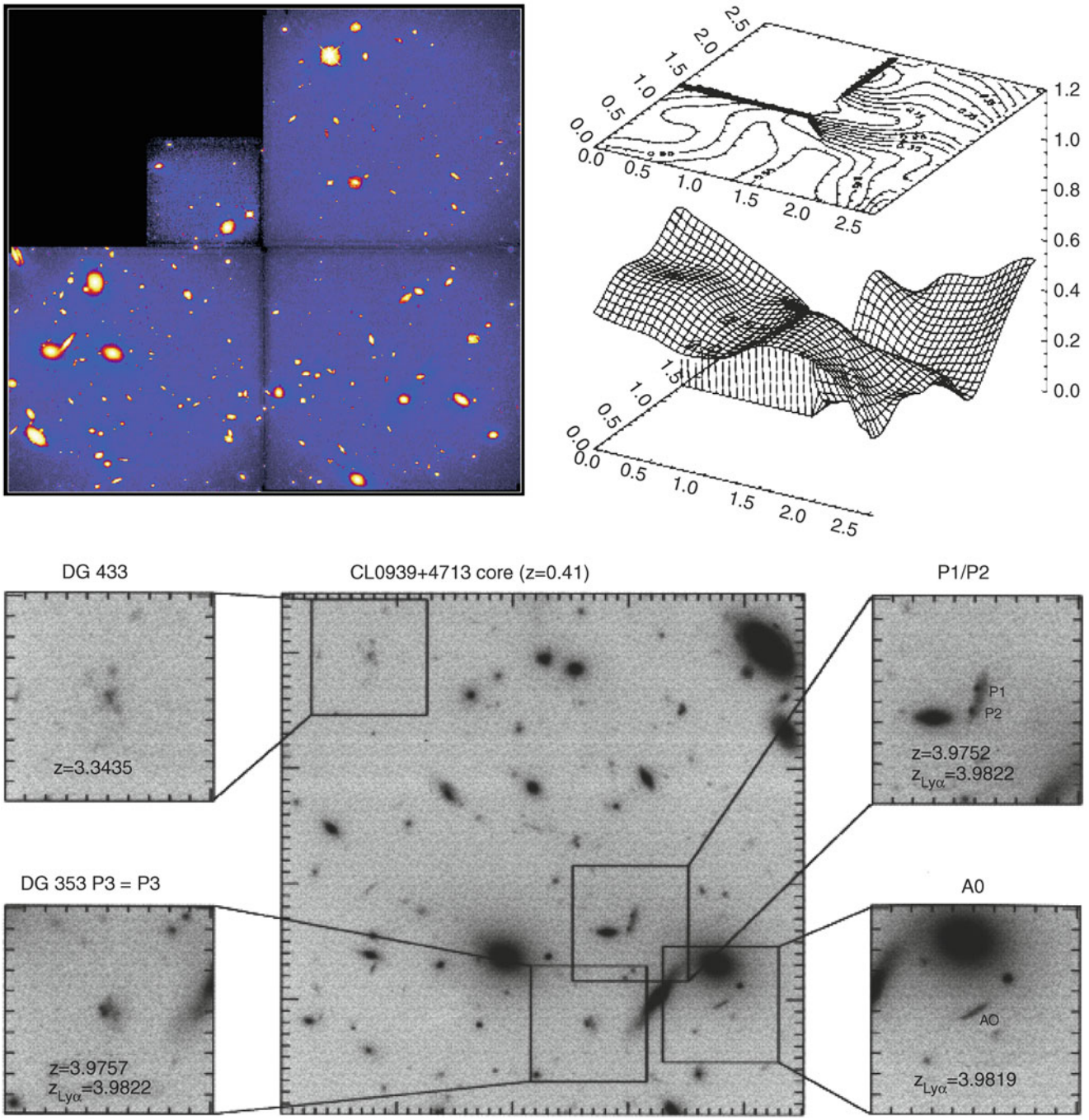


Fig. 6.56 The cluster of galaxies Cl 0939+4713 (A851) is the cluster with the highest redshift in the Abell catalog. The HST image in the *upper left panel* was obtained shortly after the refurbishment of the HST in 1994; in this image, North is down, whereas it is up in the *bottom images*. The mass distribution of the cluster was reconstructed from this image and is shown in the *upper right panel*, both as the level surface and by the contours on top. We see that the distributions of bright galaxies and of (dark) matter are very similar: their respective centers are aligned, a secondary maximum exists in both the light and the matter distribution, as does the prominent minimum in which no bright galaxies are visible either. A higher-resolution mass map from the weak lensing data predicts that the cluster is critical near its center, i.e., able to produce strong lensing features. Indeed, these were observed, as can be seen from the image at the *bottom*: a triple image system at $z \approx 3.98$ and an arc with $z = 3.98$ were confirmed spectroscopically. Source: *Top*: C. Seitz et al. 1996, *The mass distribution of CL0939+4713 obtained from a 'weak' lensing analysis of a WFPC2 image.*, A&A 314, 707, p. 708, 712, Figs. 1, 5. ©ESO. Reproduced with permission. *Bottom*: S.C. Trager et al. 1997, *Galaxies at* $z \approx 4$ *and the Formation of Population II*, ApJ 485, 92, p. 93, Fig. 1. ©AAS. Reproduced with permission

Fig. 6.57 The Bullet cluster, shown before in Fig. 6.34. In *red*, the X-ray emission is shown, superposed on the HST image of the cluster. In *blue*, the weak lensing mass reconstruction is shown, which is located on the galaxy concentrations of the two colliding clusters. Credit: X-ray: NASA/CXC/CfA/M. Markevitch et al.; Optical: NASA/STScI; Magellan/U. Arizona/D. Clowe et al.; Lensing Map: NASA/STScI; ESO WFI; Magellan/U.Arizona/D. Clowe et al.

thus the dark matter components of both colliding clusters (like the galaxies) just run through each other. So: where's the mass?

This question can be answered unambiguously with the weak lensing effect; the result is displayed in Fig. 6.57. The mass is centered on the galaxy distributions of the two clusters, and clearly displaced from the X-ray emitting gas. Therefore, the mass discrepancy cannot be ascribed to a modification of the law of gravity; instead, the main mass component of the clusters must behave very similarly as the galaxies, i.e., collisionless. Thus, the Bullet cluster provides the clearest direct proof for the existence of dark matter. Whereas the Bullet cluster was the first of its kind, in the meantime other cluster collisions have been investigated as well. Two other clusters with very similar behavior are shown in Fig. 6.58.

> The weak lensing results from clusters in collision cannot be explained without the existence of collisionless dark matter in galaxy clusters, even if modifications of the law of gravity on large scales were allowed for.

Dark matter filaments between cluster pairs. The hierarchical structure growth in the Universe, a consequence of the Cold Dark Matter model, predicts that mass concentrations form near the intersection points of dark matter filaments, through which they accrete mass (see Chap. 7). These filaments in general have a density too low to produce a detectable weak lensing signal. However, if two large mass concentrations, such as two clusters, are relatively close to each other, the density of the filament between them is expected to be considerably higher than in the mean. Therefore, pairs of galaxy clusters are the best locations to look for dark matter filaments that connect the members of the pair. Indeed, several attempts for the detection of such filaments have been made, using weak lensing techniques; in most of the cases the significance of the detection is rather low. However, for the best case up to now, the double cluster Abell 222+223, there is a convincing detection of a filament (see Fig. 6.59). In addition, the location of the filament coincides with an overdensity of galaxies and soft X-ray emission. Therefore, this system may be the first detection of a dark matter filament.

Mass calibration of clusters. We saw that the mass determination of clusters from their X-ray emission is far from trivial, owing to the complexity of the baryonic physics affecting the intracluster gas, at least in the inner region of clusters. Any systematic uncertainty in the mass estimate affects the scaling relations, which are an essential tool for cluster cosmology (see Sect. 8.2). Gravitational lensing offers an alternative possibility for determining cluster masses which is independent of equilibrium and symmetry assumptions. The accuracy of weak lensing masses for individual clusters, out to large radii (e.g., within r_{500}), is not impressive, and is affected by the noise caused by intrinsic galaxy ellipticity, light deflection by other masses along the line-of-sight to the source population, and the mass-sheet degeneracy (see Problem 3.5). However, the mean mass of a statistical sample of clusters can be obtained by using the superposition of the individual lensing signals. The mass estimate from such a 'stacked' cluster profile is viewed as

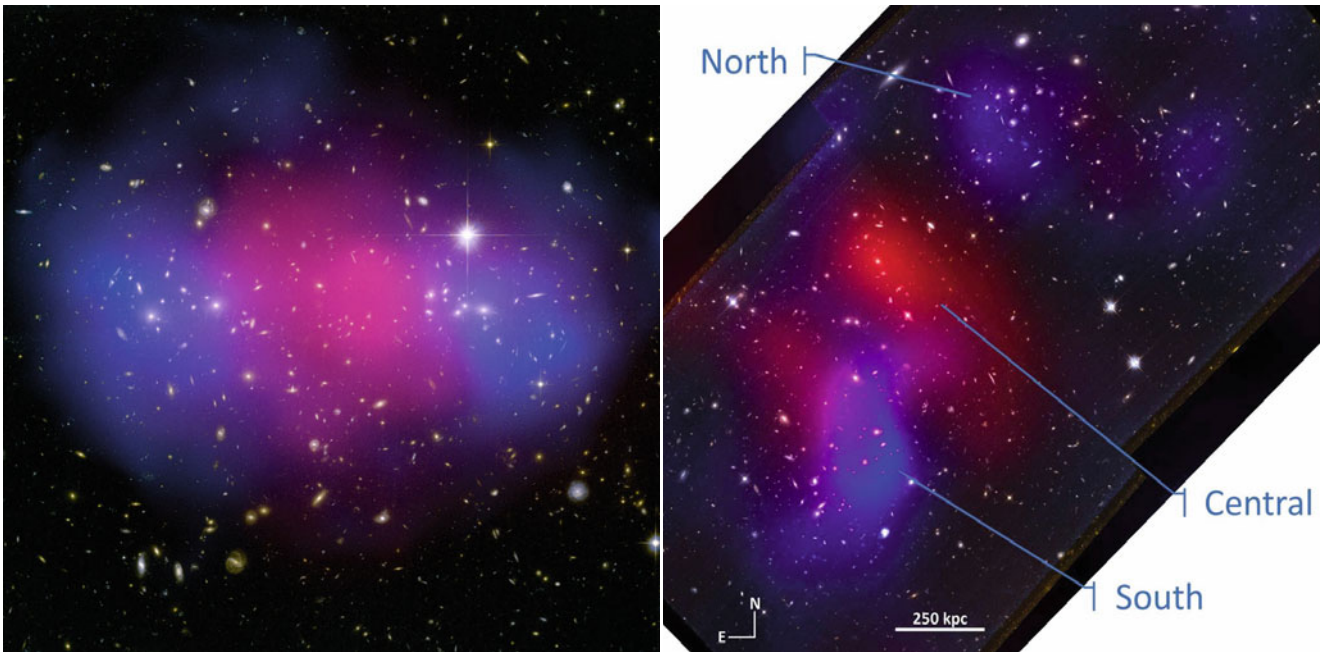


Fig. 6.58 Two clusters which are similar to the Bullet cluster. *Left panel*: The cluster MACS J0025.4−1222 has undergone a recent collision between two separate clusters. The two galaxy concentrations are seen at the left and right side of the optical image, which was taken with the HST. Shown in *pink* is the X-ray emission from this cluster, observed with Chandra. Using the weak lensing technique, the mass distribution of this cluster was reconstructed and is shown in *blue*. As for the Bullet cluster, the total mass distribution traces that of the galaxies, and is very different from the distribution of the hot intracluster gas. *Right panel*: The merging cluster DLSCL J0916.2+2951 at $z = 0.53$ also has its hot gas (indicated by the X-ray emission shown in *red*) displaced from the two concentrations of cluster galaxies, denoted by 'North' and 'South' in the image. In *blue*, the weak lensing mass reconstruction of this cluster is shown, again centered on the galaxy concentrations and clearly displaced from the intracluster gas. Credit: *Left*: NASA, ESA, CXC, M. Bradac (University of California, Santa Barbara), and S. Allen (Stanford University). *Right*: W.D. Dawson et al. 2012, *Discovery of a Dissociative Galaxy Cluster Merger with Large Physical Separation*, ApJ 747, L42, p. 2, Fig. 1. ©AAS. Reproduced with permission

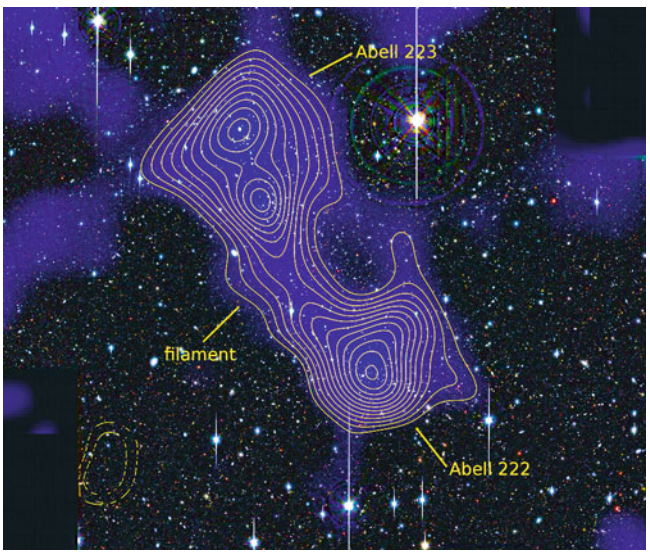


Fig. 6.59 The pair of clusters Abell 222 and 223, as observed with the Subaru telescope, with the superimposed matter distribution (*blue shades* and *yellow contours*). The filament between the pair of clusters is clearly indicated. Credit: Jörg Dietrich, University of Michigan/University Observatory Munich

being least affected by systematic effects, and therefore frequently employed in constructing scaling relations. In Sect. 7.7 we will present results of such studies.

The search for clusters of galaxies with weak gravitational lensing. The weak lensing effect can not only be used to map the matter distribution of known clusters, but it can also be employed to search for clusters. Mass concentrations generate a tangentially oriented shear field in their vicinity, which can specifically be searched for. The advantage of this method is that it detects cluster-mass matter concentrations based solely on their mass properties, in contrast to all other methods which rely on the emission of electromagnetic radiation, whether in the form of optical light from cluster galaxies or as X-ray emission from a hot intracluster medium. In particular, if clusters with atypically low gas or galaxy content exist, they could be detected in this way.

Quite a number of galaxy clusters have been detected with this method—see Fig. 6.60. Further candidates exist, where the shear signal indicates a significant mass concentration but it cannot be identified with any concentration of galaxies on

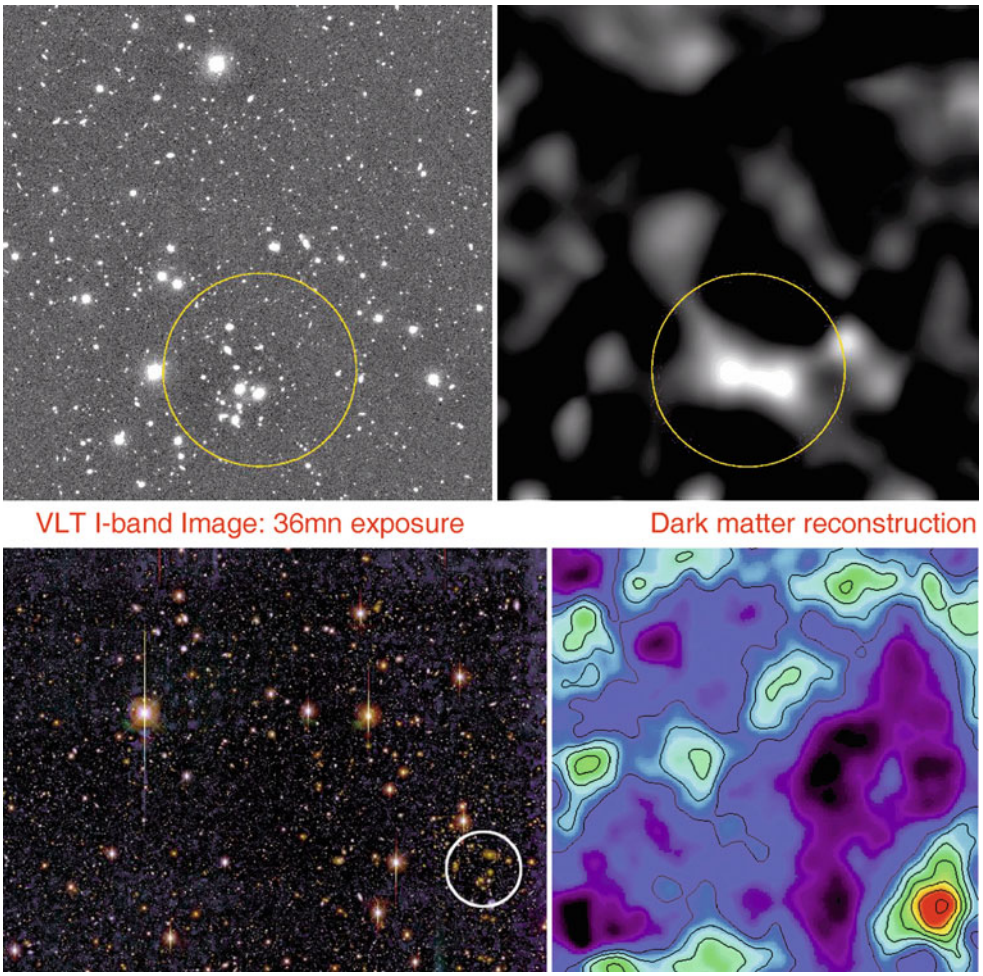


Fig. 6.60 *Top left*: a VLT/FORS1 image, taken as part of a survey of 'empty fields'. *Top right*: the mass reconstruction, as was obtained from the optical data by employing the weak lensing effect. Clearly visible is a peak in the mass distribution; the optical image shows a concentration of galaxies in this region. Hence, in this field a cluster of galaxies was detected for the first time by its lens properties. *Bottom*: as above, here a galaxy cluster was also detected through its lensing effect. *On the left*, an optical wide-field image is shown, obtained by the Big Throughput Camera, and the mass reconstruction is displayed *on the right*. The location of the peak in the latter coincides with a concentration of galaxies. Spectroscopic measurements yield that these form a cluster of galaxies at $z = 0.276$. Credit: *Top*: R. Maoli et al., ESO. *Bottom*: D. Wittman/Lucent Technologies' Bell Labs, et al., NOAO, AURA, NSF

optical images. The clarification of the nature of these lens signals is of great importance: if in fact matter concentrations do exist which correspond to the mass of a cluster but which do not contain luminous galaxies, then our understanding of galaxy evolution needs to be revised. However, we cannot exclude the possibility that these statistically significant signals are statistical outliers, or result from projection effects—remember, lensing probes the line-of-sight integrated matter density. Together with the search for galaxy clusters by means of the SZ-effect (Sect. 6.4.4), the weak lensing effect provides an interesting alternative for the detection of mass concentrations compared to the more traditional methods.

6.7 The galaxy population in clusters

6.7.1 Luminosity function of cluster galaxies

The luminosity function of galaxies in a cluster is defined in a similar way as in Sect. 3.10 for the total galaxy population. In many clusters, the Schechter luminosity function (3.52) represents a good fit to the data if the brightest galaxy is disregarded in each cluster (see Fig. 3.51 for the Virgo cluster of galaxies). The slope α at the faint end is not easy to determine, since projection effects become increasingly important for fainter galaxies. The value of α seems to vary between clusters, but it is not entirely clear whether this result may also be affected by projection effects of differing strength in different clusters. Thus, no final conclusion has been reached as to whether the luminosity function has a steep increase at $L \ll L^*$ or not, i.e., whether many more faint galaxies exist than luminous $\sim$L*-galaxies (compare the galaxy content in the Local Group, Sect. 6.1.1, where even in our close neighborhood it is difficult to obtain a complete census of the galaxy population). L^* is very similar for many clusters, which is the reason why the distance estimate by apparent brightness of cluster members, as done by Abell for his cluster catalog, is quite reliable, though a number of clusters exists with a clearly deviating value of L^*.

However, when averaging over many clusters, the picture becomes much clearer. The luminosity function of galaxies changes as one moves from field galaxies, through groups, to clusters. This change is already noticeable for poor groups with just a few galaxies, such as the Local Group. The

faint-end slope of the luminosity function is flatter in group and cluster environments than for field galaxies, and flattens with increasing cluster mass or richness. Most galaxies with $L \lesssim L^*$ live either in isolation or in relatively poor groups, whereas the most luminous galaxies are predominantly found in the central regions of rich groups and clusters. Whereas the fraction of elliptical galaxies is much smaller in low-density regions than in the inner part of clusters, they nevertheless dominate the galaxy mix at the highest luminosities.

Brightest Cluster Galaxies. Of special interest in clusters is the Brightest Cluster Galaxy (BCG), which in most cases is located near the center of the cluster. In many cases, the BCG is a *cD galaxy*; these differ from large ellipticals in several respects. They have a very extended stellar envelope, whose size may exceed $R \sim 100\,\mathrm{kpc}$ and whose surface brightness profile is much broader than that of a de Vaucouleurs-profile (see Fig. 3.11). cD galaxies are found only in the centers of clusters or groups, thus only in regions of strongly enhanced galaxy density. However, the extended stellar envelope merges into the intracluster light, and there is some ambiguity to distinguish between cD envelope and ICL, as discussed in Sect. 6.3.4. Many cD galaxies have multiple cores, which is a rather rare phenomenon among the other cluster members.

BCGs are very luminous and massive; their typical stellar mass is $M_* \sim 2 \times 10^{11} h^{-2} M_\odot$, with a spread of about a factor of 1.5. One finds that the more massive clusters contain the more luminous BCGs. At first sight, this is not too surprising, since the more galaxies a cluster contains, the larger is the probability to find one of them with very high stellar mass (or luminosity). However, there are several indications that this cannot be the sole reason. First, in most clusters, the second brightest cluster galaxy is about 1 mag or more fainter than the BCG, which cannot be explained purely by statistical arguments. Second, the BCGs appear to be not simply the extension of early-type galaxies towards the highest luminosity; for example, BCGs do not fall right on the fundamental plane for ellipticals. Their size-luminosity relation is different from that of ellipticals, in the sense that they are larger than ellipticals with the same luminosity.

6.7.2 The morphology-density relation

As mentioned several times before, the mixture of galaxy types in clusters seems to differ from that of isolated (field) galaxies. Whereas about 70 % of luminous field galaxies are spirals, clusters are dominated by early-type galaxies, in particular in their inner regions. Furthermore, the fraction of spirals in a cluster depends on the distance from the center and increases for larger r. Obviously, the local density has an effect on the morphological mix of galaxies. As in clusters, the fraction of group members which are spirals is lower than the fraction of spirals among field (i.e., isolated) galaxies, and the relative abundance of spiral galaxies decreases with increasing σ_v of the group.

More generally, one may ask whether the mixture of the galaxy population depends on the local galaxy density. While earlier studies of this effect were frequently confined to galaxies within and around clusters, extensive redshift surveys like the 2dFGRS and the SDSS (see Sect. 8.1.2) allow us to systematically investigate this question with very large and carefully selected samples of galaxies. The morphological classification of such large samples is performed by automated software tools, which basically measure the light concentration in the galaxies, or, alternatively, the best-fitting Sérsic-index n [see (3.39)]. A comparison of galaxies classified this way with visual classifications shows very good agreement.

Results from the Sloan Digital Sky Survey. As an example of such an investigation, results from the Sloan Digital Sky Survey are shown in Fig. 6.61. Galaxies were morphologically classified, based on SDSS photometry, and separated into four classes, corresponding to elliptical galaxies, S0-galaxies, and early (Sa) and late (Sc) types of spiral. In this analysis, only galaxies were included for which the redshift was spectroscopically measured. Therefore, the spatial galaxy density can be estimated. However, one needs to take into account the fact that the measured redshift is a superposition of the cosmic expansion and the peculiar velocity of a galaxy. The peculiar velocity may have rather large values ($\sim 1000\,\mathrm{km/s}$), in particular in clusters of galaxies. For this reason, for each galaxy in the sample the surface number density of galaxies which have a redshift within $\pm 1000\,\mathrm{km/s}$ of the target galaxy was determined. The left panel in Fig. 6.61 shows the fraction of the different galaxy classes as a function of this local galaxy density. A very clear dependence, in particular of the fraction of late-type spirals, on the local density can be seen: in regions of higher galaxy density Sc-spirals contribute less than 10 % of the galaxies, whereas their fraction is about 30 % in low-density regions. Combined, the fraction of spirals decreases from $\sim$ 65 % in the field to about 35 % in regions of high galaxy density. In contrast, the fraction of ellipticals and S0-galaxies increases towards higher densities, with the increase being strongest for ellipticals.

In the right-hand panel of Fig. 6.61, the mixture of galaxy morphologies is plotted as a function of the distance to the center of the nearest cluster, where the distance is scaled by the virial radius of the corresponding cluster. As expected, a very strong dependence of the fraction of ellipticals and spirals on this distance is seen. Sc-spirals contribute a mere 5 % of galaxies in the vicinity of cluster centers, whereas

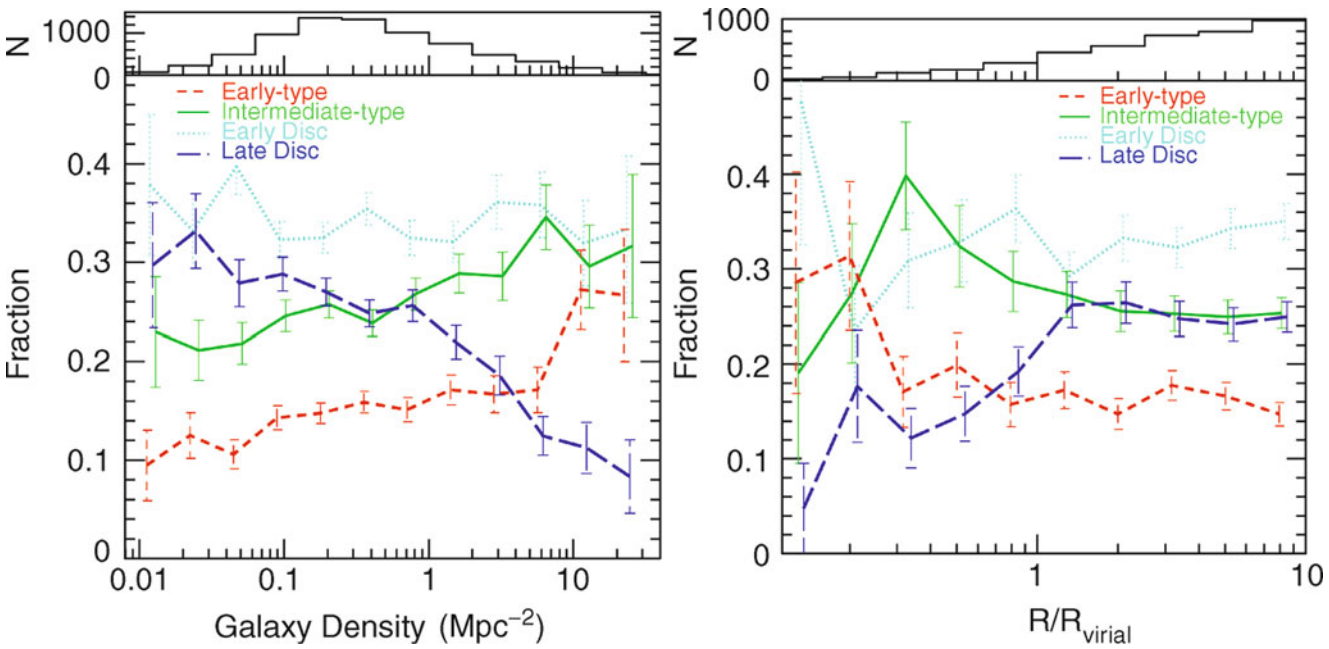


Fig. 6.61 The number fraction of galaxies of different morphologies is plotted as a function of the local galaxy density (*left panel*), and for galaxies in clusters as a function of the distance from the cluster center, scaled by the corresponding virial radius (*right panel*). Galaxies are divided into four different classes. 'Early-types' contain mainly ellipticals, 'intermediates' are mainly S0-galaxies, 'early and late discs' are predominantly Sa- and Sc-spirals, respectively. In both representations, a clear dependence of the galaxy mix on the density or on the distance from the cluster center, respectively, is visible. In the histograms at the top of each panel, the number of galaxies in the various bins is plotted. Source: T. Goto et al. 2003, *The morphology-density relation in the Sloan Digital Sky Survey*, MNRAS 346, 601, p. 607, 608, Figs. 12, 15. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

the fraction of ellipticals and S0-galaxies strongly increases inwards.

The two diagrams in Fig. 6.61 are of course not mutually independent: a region of high galaxy density is very likely to be located in the vicinity of a cluster center, and the opposite is valid accordingly. Therefore, it is not immediately clear whether the mix of galaxy morphologies depends primarily on the respective density of the environment of the galaxies, or whether it is caused by morphological transformations in the inner regions of galaxy clusters.

The morphology-density relation is also seen in galaxy groups. The fraction of late-type galaxies decreases, and the fraction of early-type galaxies increases with decreasing distance from the group center, as is also the case in clusters. When considering the morphological mix of visually classified early- and late-type galaxies, averaged over the whole group or cluster, i.e., up to the virial radius, then it seems to be constant for group/cluster halo masses in excess of $\sim 10^{13} M_{\odot}$.

Alternative consideration: the color-density relation. We pointed out in Sect. 3.1.3 that galaxies at fixed luminosity seem to have a bimodal color distribution (see Fig. 3.7). Using the same data set as that used for Fig. 3.7, the fraction of galaxies that are contained in the red population can be studied as a function of the local galaxy density. The result of this study is shown in the left-hand panel of Fig. 6.62, where the fraction of galaxies belonging to the red population is plotted against the local density of galaxies, measured in terms of the fifth-nearest neighboring galaxy within a redshift of $\pm 1000\,\mathrm{km/s}$. The fraction of red galaxies increases towards higher local number density, and the relative increase is stronger for the less luminous galaxies. If we identify the red galaxies with the early-type galaxies in Fig. 6.61, these two results are in qualitative agreement. Surprisingly, the fraction of galaxies in the red sample seems to be a function of a combination of the local galaxy density and the luminosity of the galaxy, as is shown in the right-hand panel of Fig. 6.62.

Interpretation. A closer examination of Fig. 6.61 may provide a clue as to what physical processes are responsible for the dependence of the morphological mix on the local number density. We consider first the right-hand panel of Fig. 6.61. Three different regimes in radius can be identified: for $R \gtrsim R_{\mathrm{vir}}$, the fraction of the different galaxy types remains basically constant. In the intermediate regime, $0.3 \lesssim R/R_{\mathrm{vir}} \lesssim 1$, the fraction of S0-galaxies strongly increases inwards, whereas the fraction of late-type spirals decreases accordingly. This result is compatible with the interpretation that in the outer regions of galaxy clusters spirals lose gas,

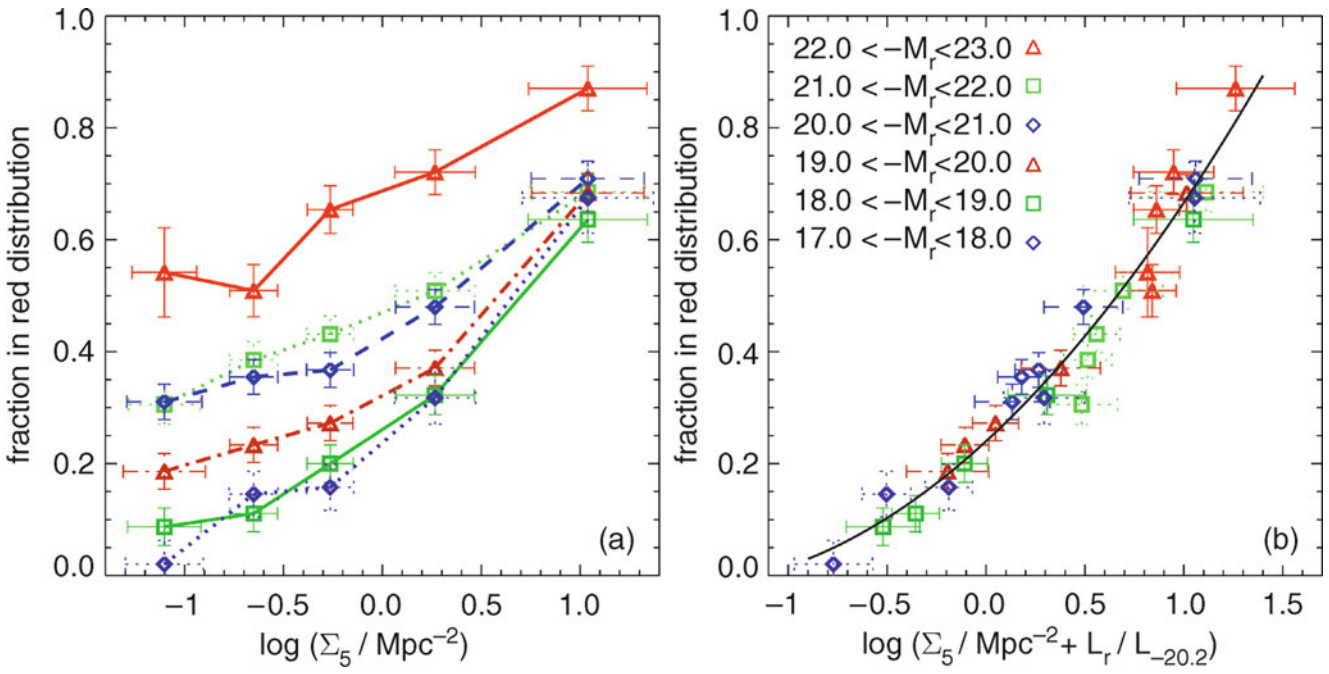


Fig. 6.62 *On the left*, the fraction of galaxies in the red population (see Sect. 3.1.3) is shown as a function of Σ_5, an estimator of the local galaxy number density based on the projected distance of the fifth-nearest spectroscopically confirmed neighbor galaxy within ± 1000 km/s. Different symbols correspond to different luminosity bins, as indicated. *On the right*, the same red fraction is plotted against a combination of the local galaxy density Σ_5 and the luminosity of the galaxy. Source: I.K. Baldry et al. 2004, *Color bimodality: Implications for galaxy evolution*, astro-ph/0410603, Fig. 6. Reproduced by permission of the author

for instance by their motion through the intergalactic medium (see Fig. 6.63), and these galaxies then transform into passive S0-galaxies. Below $R \lesssim 0.3 R_{\rm vir}$, the fraction of S0-galaxies decreases strongly, and the fraction of ellipticals increases substantially.

In fact, the ratio of the number densities of S0-galaxies and ellipticals, for $R \lesssim 0.3 R_{\rm vir}$, strongly decreases as R decreases. This may hint at a morphological transformation in which S0 galaxies are turned into ellipticals, probably by collisions or mergers. Such gas-free mergers, also called 'dry mergers', may be the preferred explanation for the generation of elliptical galaxies. One of the essential properties of dry mergers is that such a merging process would not be accompanied by a burst of star formation, unlike the case of gas-rich collisions of galaxies. The existence of a population of newly-born stars in ellipticals would be difficult to reconcile with the generally old stellar population actually observed in these galaxies. We will discuss these issues more thoroughly in Chap. 10.

Considering now the dependence on local galaxy density (the left-hand panel of Fig. 6.61), a similar behavior of the morphological mix of galaxies is observed: there seem to exist two characteristic values for the galaxy density where the relative fractions of galaxy morphologies change noticeably. Interestingly, the relation between morphology and density seems to evolve only marginally between $z = 0.5$ and the local Universe.

One clue as to the origin of the morphological transformation of galaxies in clusters, as a function of distance from the cluster center, comes from the observation that the velocity dispersion of very bright cluster galaxies seems to be significantly smaller than that of less luminous ones. Assuming that the mass-to-light ratio does not vary substantially among cluster members, this then indicates that the most massive galaxies have smaller velocity dispersions. One way to achieve this trend in the course of cluster evolution is by dynamical interactions between cluster galaxies. Such interactions tend to 'thermalize' the velocity distribution of galaxies, so that the mean kinetic energy of galaxies tends to become similar. This then causes more massive galaxies to become slower on average. If this interpretation holds, then the morphology-density relation may be attributed to these dynamical interactions, rather than to the (so-called ram-pressure) stripping of the interstellar medium as the galaxies move through the intracluster medium.

However, ram-pressure stripping of the interstellar medium has been clearly observed in clusters. In Fig. 6.63, two edge-on spiral galaxies in the Virgo cluster are shown, both in the optical and in the radio. The impact of the intracluster medium, which acts like a 'wind' in the rest-frame of the galaxies, can be clearly seen. This effect mostly acts on the atomic gas of spirals, whereas the molecular gas seems to be less affected; we recall that the molecular gas is more densely concentrated towards the

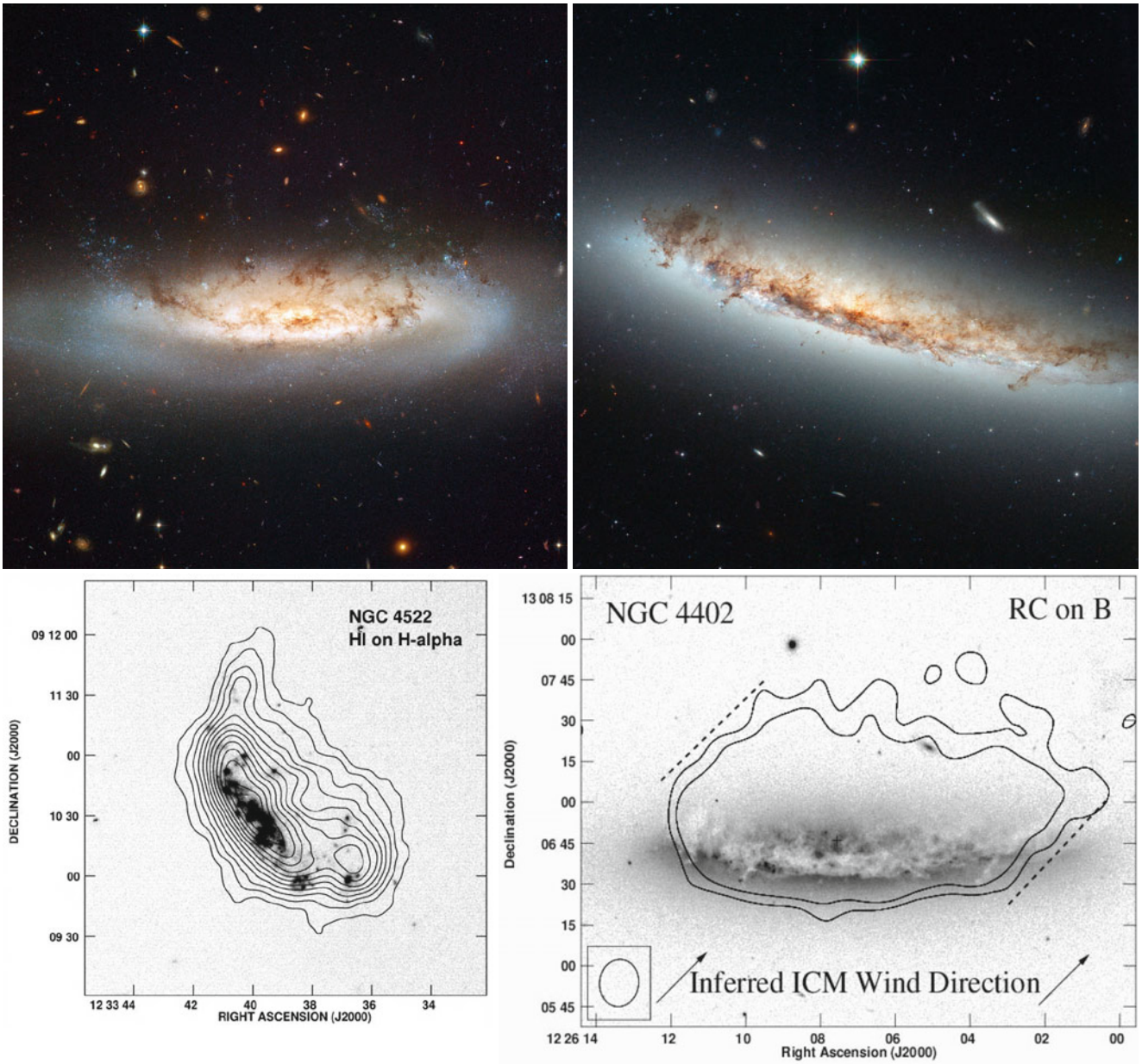


Fig. 6.63 Two edge-on spiral galaxies in the Virgo cluster, NGC 4522 (*left*) and NGC 4402 (*right*). The *top panels* show multi-color composite images of these two galaxies, taken with the HST. The high velocity of the galaxies moving through the intracluster gas removes part of the interstellar medium through ram pressure stripping; in addition, the interaction of the intracluster gas with the ISM of the galaxies triggers star formation, as can be seen by the blue knots of emission downstream (i.e., above the disk). The *bottom panels* show the neutral hydrogen gas (*left*, for NGC 4522) and the 20 cm radio emission (*right*, for NGC 4022) of these two galaxies (note that the orientation of these images are different from those of the HST images). The gas is distributed asymmetrically with respect to the galactic disk, emphasizing the stripping of gas from the galaxies. In the *bottom right*, the *arrows* indicate the direction of the ICM velocity as seen in the rest-frame of the galaxy (i.e., the galaxy moves in the opposite direction to these *arrows*). Credit: *Top panels*: NASA & ESA. *Bottom left*: B. Vollmer et al. 2008, *Ram-pressure stripped molecular gas in the Virgo spiral galaxy NGC 4522*, A&A 491, 455, p. 461, Fig. 13. ©ESO. Reproduced with permission. Bottom right: H.H. Crowl 2005, *Dense Cloud Ablation and Ram Pressure Stripping of the Virgo Spiral NGC 4402*, AJ 130, 65, p. 68, Fig. 3. ©AAS. Reproduced with permission

galactic disk, and thus more strongly bound. In fact, it has been known for a long time that there are spiral galaxies in groups and clusters which are deficient in neutral hydrogen, relative to field galaxies of the same stellar luminosity. If ram-pressure stripping, or other effects in the dense cluster environment, removes the interstellar medium from spirals, then the result could be a disk galaxy without ongoing star formation—something that may resemble an S0-galaxy. If that were the case, than S0s would be passively fading former spirals. Whereas the original spirals satisfied the

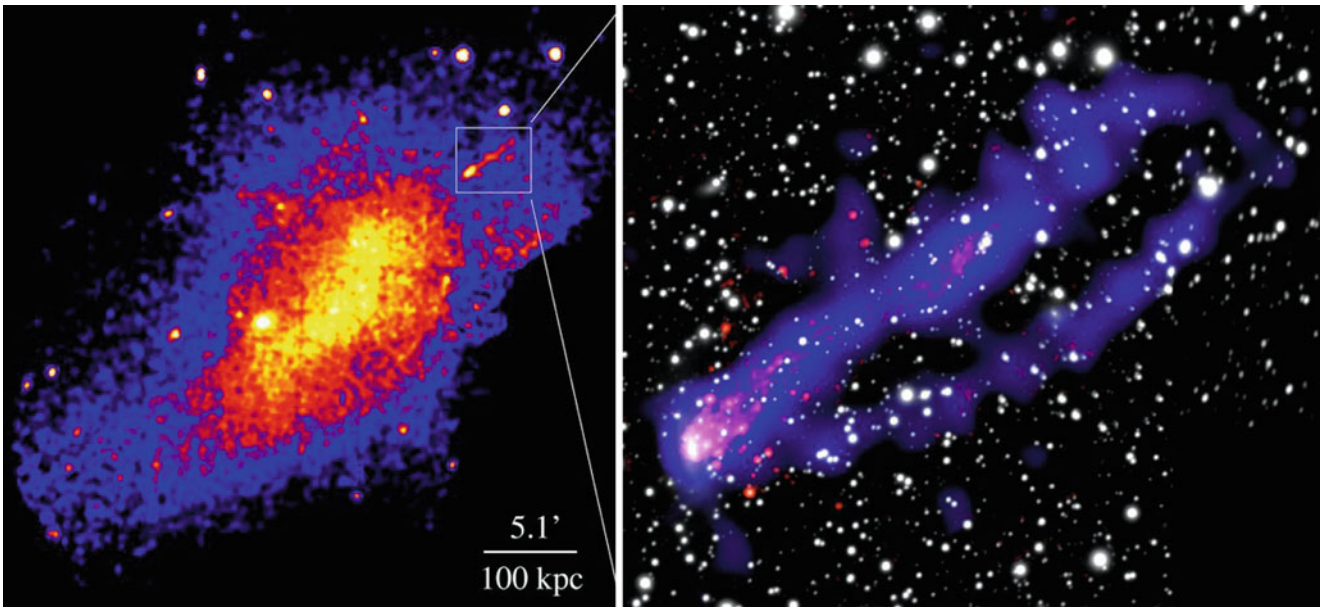


Fig. 6.64 The *left panel* shows an XMM-Newton image of the cluster Abell 3627. *On the right*, the region highlighted in the *left panel* is zoomed in, showing the X-ray emission as observed with Chandra (in *blue*), a narrow-band optical image centered on the Hα line in *red*, both of which are superposed on the broad-band optical image of the cluster (in *white*; obtained with the SOAR telescope in Chile). The X-ray image clearly shows two tails, with some 80 kpc length, originating from the galaxy ESO 137-001 located on the left in this image. These tails are generated by ram-pressure stripping caused by the hot intracluster medium (seen through its X-ray emission in the *left panel*), as the galaxy moves towards the central region of the cluster. The stripped gas is heated to temperatures of $\sim$ 1 keV, about a factor of 10 smaller than the temperature of the hot intracluster medium. The Hα emission is mainly caused by H II regions and thus indicates regions of star formation. Hence, new stars are formed in the gas that is removed from the galaxy. These stars will contribute to the intracluster light. Furthermore, 19 X-ray point sources were found around the X-ray tails, some of them being candidates for ultra-luminous X-ray sources (ULX). Together, this observation shows a clear connection between the evolution of galaxies in clusters and the intracluster medium. Credit: X-ray: NASA/CXC/UVa/M. Sun, et al.; H-alpha/Optical: SOAR (UVa/NOAO/UNC/CNPq-Brazil). Corresponding journal article: M. Sun et al. 2010, *Spectacular X-ray Tails, Intracluster Star Formation, and ULXs in A3627*, ApJ 708, 946, p. 949, Fig. 2. ©AAS. Reproduced with permission

Tully–Fisher relation, the S0s would then be expected to be considerably fainter than spirals, at fixed rotational velocity; this indeed is the case.

Ram-pressure stripping not only removes gas from cluster galaxies, but can heat the gas and also trigger local star formation, though the compression of gaseous regions by the pressure. A spectacular example for this is shown in Fig. 6.64. The newly formed stars are no longer gravitationally bound to the galaxy, and will thus contribute to the ICL in the cluster.

E + A galaxies. Galaxy clusters contain a class of galaxies which is defined in terms of spectral properties. These galaxies, which are rare in number, show strong Balmer line absorption in their spectra, characteristic of A stars, but no [OII] or Hα emission lines. The latter indicates that these galaxies are not undergoing strong star formation at present (since there are no HII regions around O- and B-stars), whereas the former shows that there was an episode of star formation within the past $\sim$ 1 Gyr, about as long ago as the main-sequence lifetime of A stars. These galaxies have been termed E + A galaxies since their spectra appears like a superposition of that of A-stars and that of otherwise normal elliptical galaxies. They are interpreted as being post-starburst galaxies. Since they were first seen in clusters, the interpretation of the origin of E + A galaxies was originally centered on the cluster environment—for example star-forming galaxies falling into a cluster and having their interstellar medium removed by tidal forces caused by the cluster potential well and/or stripping as the galaxies move through the intracluster medium. However, E + A galaxies were later also found in different environments, making the above interpretation largely obsolete. By investigating the spatial correlation of these galaxies with other galaxies shows that the phenomenon is not associated with the large-scale environment. An overdensity of neighboring galaxies can be seen only out to scales of $\sim$ 100 kpc. If the sudden turn-off of the star-formation activity is indeed caused by an external perturbation, it is therefore likely that it is caused by the dynamical interaction of close neighboring galaxies or by mergers. Indeed, about 30 % of E + A galaxies are found to have morphological signatures of perturbations, such as tidal tails, supporting the interaction hypothesis.

In fact, the spiral galaxies in clusters seem to differ statistically from those of field spirals, in that the fraction of disk galaxies with absorption-line spectra, and thus no ongoing star formation, seems to be larger in clusters than in the field by a factor ~ 4, indicating that the cluster environment has a marked impact on the star-formation ability of these galaxies.

6.8 Evolutionary effects

Today, we are able to discover and analyze clusters of galaxies at redshifts $z \sim 1$ and higher; thus the question arises whether these clusters have the same properties as local clusters. At $z \sim 1$ the age of the Universe is only about half of that of the current Universe. One might therefore expect an evolution of cluster properties.

Luminosity function. First, we shall consider the comoving number density of clusters as a function of redshift or, more precisely, the evolution of the luminosity function of clusters with z. As Fig. 6.65 demonstrates, such evolutionary effects are not very pronounced for moderate redshifts, and only at the highest luminosities or the most massive clusters, respectively, does an evolution become visible. The situation is somewhat different when one considers the evolution of the cluster abundance as a function of temperature: at redshift $z \sim 0.5$, the number of clusters above a given temperature is lower than at the current epoch by a factor ~ 3. Hence, there obviously is an evolution of the cluster population, in the sense that at high redshift, clusters of very high luminosity, high temperature, or very high mass are less abundant than they are today. The interpretation and the relevance of this fact will be discussed later (see Sect. 8.2.1). We also note that a weak redshift evolution of the X-ray luminosity function of clusters is compatible with a strong evolution of the cluster mass function, since according to (6.62), the relation between luminosity and mass evolves significantly with z

Butcher–Oemler effect. We saw in Chap. 3 that early-type galaxies are predominantly found in clusters and groups, whereas spirals are mostly field galaxies. For example, a massive cluster like Coma contains only 10 % spirals, the other luminous galaxies are ellipticals or S0 galaxies (see also Sect. 6.7.2). Besides these morphological differences, the colors of galaxies are very useful for a characterization: early-type galaxies (ellipticals and S0 galaxies) have little ongoing star formation and therefore consist mainly of old, thus low-mass and cool stars. Hence they are red, whereas spirals feature active star formation and are therefore distinctly bluer. The fraction of blue galaxies in nearby clusters is very low.

Butcher and Oemler found that this changes if one examines clusters of galaxies at higher redshifts: these contain a larger fraction of blue galaxies, thus of spirals (see Fig. 6.66). This means that the mixture of galaxies changes over time. In clusters, spirals must become scarcer with increasing cosmic time, e.g., by transforming into early-type galaxies. With 8-m class telescopes, such studies can be extended to much higher redshifts. It was found that the fraction of blue galaxies increases further, until at $z \gtrsim 1.3$ the blue fraction in clusters is essentially the same as in the field.

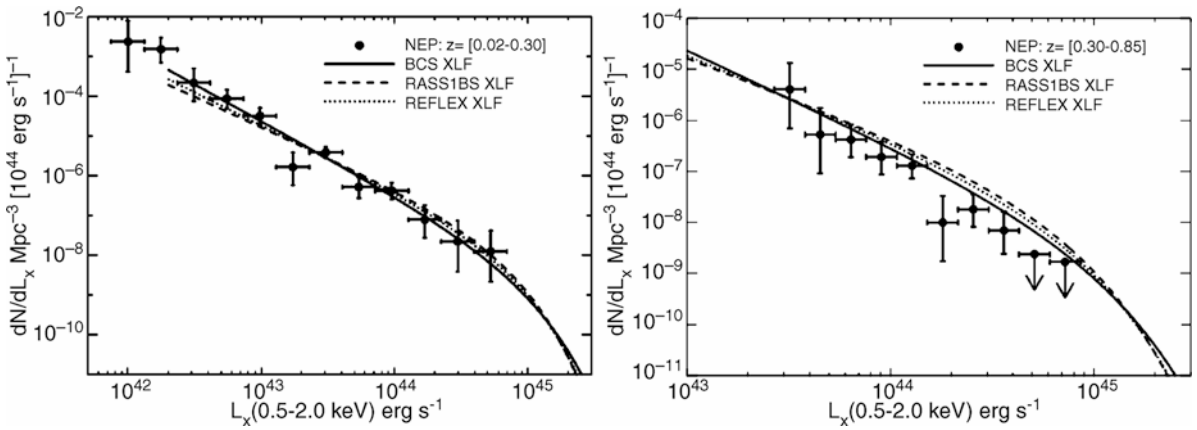


Fig. 6.65 X-ray luminosity function of galaxy clusters, as was obtained from a region around the north ecliptic pole (NEP), the region with the longest exposure time in the ROSAT All-Sky Survey (see Fig. 6.40). Plotted is $\mathrm{d}N/\mathrm{d}L_\mathrm{X}$, the (comoving) number density per luminosity interval, for clusters with $0.02 \leq z \leq 0.3$ (*left panel*) and $0.3 \leq z \leq 0.85$ (*right panel*), respectively. The luminosity was derived from the flux in the photon energy range from 0.5 to 2 keV. The *three different curves* specify the local luminosity function of clusters as found in other cluster surveys at lower redshifts. We see that evolutionary effects in the luminosity function are relatively small and become visible only at high L_X. Source: I.M. Gioia et al. 2001, *Cluster Evolution in the ROSAT North Ecliptic Pole Survey*, ApJ 553, L105, p. L106, Fig. 1.

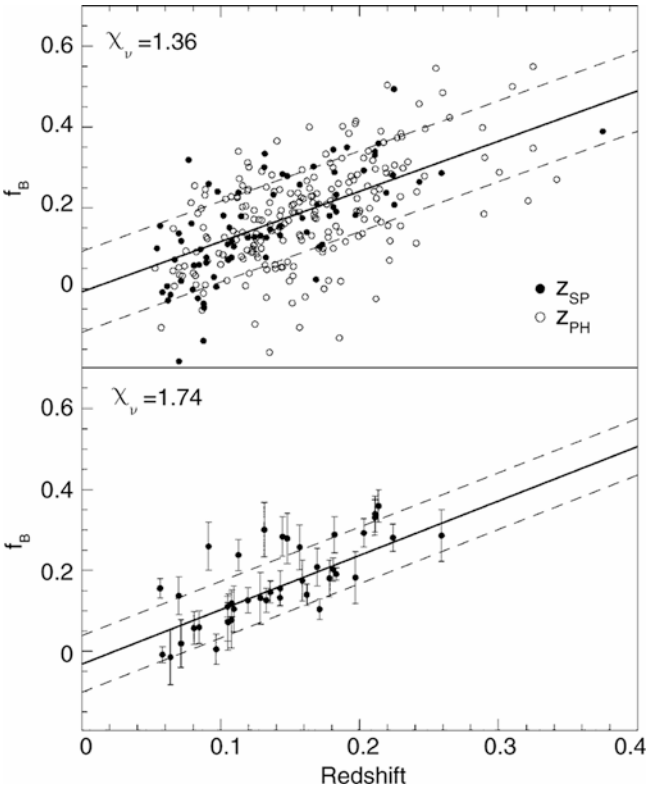


Fig. 6.66 Butcher–Oemler effect: in the *upper panel*, the fraction of blue galaxies f_b in a sample of 195 galaxy clusters is plotted as a function of cluster redshift, where *open* (*filled*) *circles* indicate photometric (spectroscopic) redshift data for the clusters. The *lower panel* shows a selection of clusters with spectroscopically determined redshifts and well-defined red cluster sequence. For the determination of f_b, foreground and background galaxies need to be statistically subtracted using control fields, which can also result in negative values for f_b. A clear increase in f_b with redshift is visible, and a line of regression yields $f_b = 1.34z - 0.03$. Source: V.E. Margoniner et al. 2001, *The Butcher-Oemler Effect in 295 Clusters: Strong Redshift Evolution and Cluster Richness Dependence*, ApJ 548, L143, p. L144, Fig. 1. ©AAS. Reproduced with permission

A possible and plausible explanation is that spirals lose their interstellar gas through ram-pressure stripping (Fig. 6.63), which is then mixed with the ICM. This is plausible because the ICM also has a high metallicity. These metals can only originate in a stellar population, thus in the enriched material in the ISM of galaxies.

Indeed, from the colors of cluster galaxies it is possible to derive very strict upper limits on their star formation in recent times. The color of cluster galaxies at high redshifts even provides interesting constraints on cosmological parameters—only those models are acceptable which have an age of the Universe, at the respective redshift, larger than the estimated age of the stellar population. One interesting example of this is presented in Fig. 6.67.

Therefore, we conclude from these observations that the stars in cluster galaxies formed at very early times in the Universe. But this does not necessarily mean that the galaxies themselves are also this old, because galaxies can be transformed into each other by merger processes (see Fig. 6.68). This changes the morphology of galaxies, but may leave the stellar populations largely unchanged.

At higher redshift (say, $z \gtrsim 0.5$), the fraction of clusters which show a cool core and which are therefore believed to be relaxed is smaller than in the current Universe. This observational finding is not unexpected, since the age of the Universe was smaller back then, and thus there was considerably less time for clusters to settle into an equilibrium state. Furthermore, the metallicity of the intracluster gas decreases towards higher redshifts, indicating that the enrichment of the gas is an ongoing process.

Clusters of galaxies at very high redshift. The search for clusters at high redshift is of great cosmological interest. As will be demonstrated in Sect. 7.5.2, the expected number density of clusters as a function of z strongly depends on the cosmological model. Hence, this search offers an opportunity to constrain cosmological parameters by the statistics of galaxy clusters.

The search for clusters in the optical (thus, by galaxy overdensities) becomes increasingly difficult at high z because of projection effects. Nevertheless, several groups have managed to detect clusters at $z \sim 0.8$ with this technique. In particular, the overdensity of galaxies in three-dimensional space can be analyzed if, besides the angular coordinates on the sphere, the galaxy colors are also taken into account. Because of the red cluster sequence, the overdensity is much more prominent in this space than in the sky projection alone.

Projection effects play a considerably smaller role in X-ray searches for clusters. With ROSAT, some clusters with $z \sim 1.2$ were found (see Fig. 6.39). The current X-ray satellites Chandra and XMM-Newton are more sensitive, and have detected several clusters with redshifts up to 2; two examples for clusters at $z \sim 1.4$ are shown in Figs. 6.69 and 6.70. These examples demonstrate that combining deep X-ray images with observations in the optical and the NIR is an efficient method of compiling samples of distant clusters. More recently, cluster detections using the Sunyaev–Zeldovich effect have obtained an increasingly prominent role in finding high-redshift clusters.

Through optical methods, it is also possible to identify galaxy concentrations at very high redshift. One approach is to assume that luminous AGNs at high redshift are found preferentially in regions of high overdensity, which is also expected from models of galaxy formation. With the redshift of the AGN known, the redshift at which one should search for an overdensity of galaxies near the AGN is defined. Those searches have proven to be quite successful; for instance, they are performed using narrow-band filter

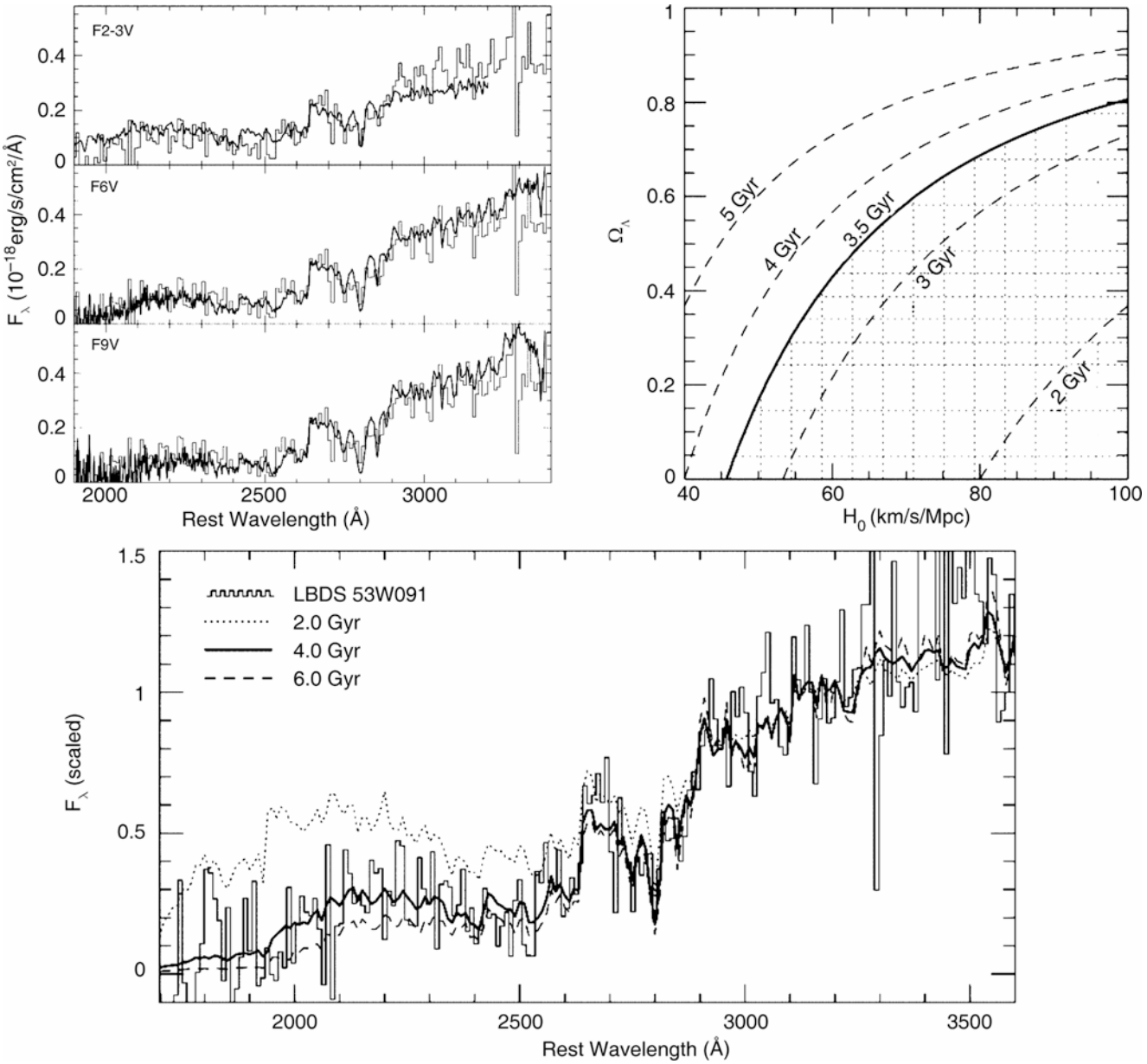


Fig. 6.67 The radio galaxy LBDS 53W091 has a redshift of $z = 1.552$, and it features a very red color ($R - K \approx 5.8$). Optical spectroscopy of the galaxy provides us with the spectral light distribution of the UV emission in the galaxy's rest frame. The UV light of a stellar population is almost completely due to stars on the upper main sequence—see Fig. 3.34. In the *upper left panel*, the spectrum of LBDS 53W091 is compared to those of different F stars; one can see that F6 stars match the spectral distribution of the galaxy nearly perfectly. In the *bottom panel*, synthetic spectra from population synthesis calculations are compared to the observed spectrum. A population with an age of about 4 Gyr represents the best fit to the observed spectrum; this is also comparable to the lifetime of F6 stars: the most luminous (still existing) stars dominate the light distribution of a stellar population in the UV. In combination, this reveals that this galaxy at $z = 1.552$ is at least 3 Gyr old. Phrased differently, the age of the Universe at $z = 1.55$ must be at least 3 Gyr. In the *upper right panel*, the age of the Universe at $z = 1.55$ is displayed as a function of H_0 and Ω_Λ (for $\Omega_\mathrm{m} + \Omega_\Lambda = 1$). Hence, this single galaxy provides significant constraints on cosmological parameters. Source: H. Spinrad et al. 1997, *LBDS 53W091: an Old, Red Galaxy at* $z = 1.552$, ApJ 484, 581, p. 587, 595, 599, Figs. 8, 17, 18. ©AAS. Reproduced with permission

photometry, with the filter centered on the redshifted Lyα line, tuned to the redshift of the AGN. Candidates need to be verified spectroscopically afterwards. One example of a strong galaxy concentration at $z = 4.1$ is presented in Fig. 6.71. The identification of a strong spatial concentration of galaxies is not sufficient to have identified a cluster of galaxies though, because it is by no means clear whether one has found a gravitationally bound system of galaxies (and the corresponding dark matter). Rather, such galaxy concentrations are considered to be the predecessors of galaxy clusters

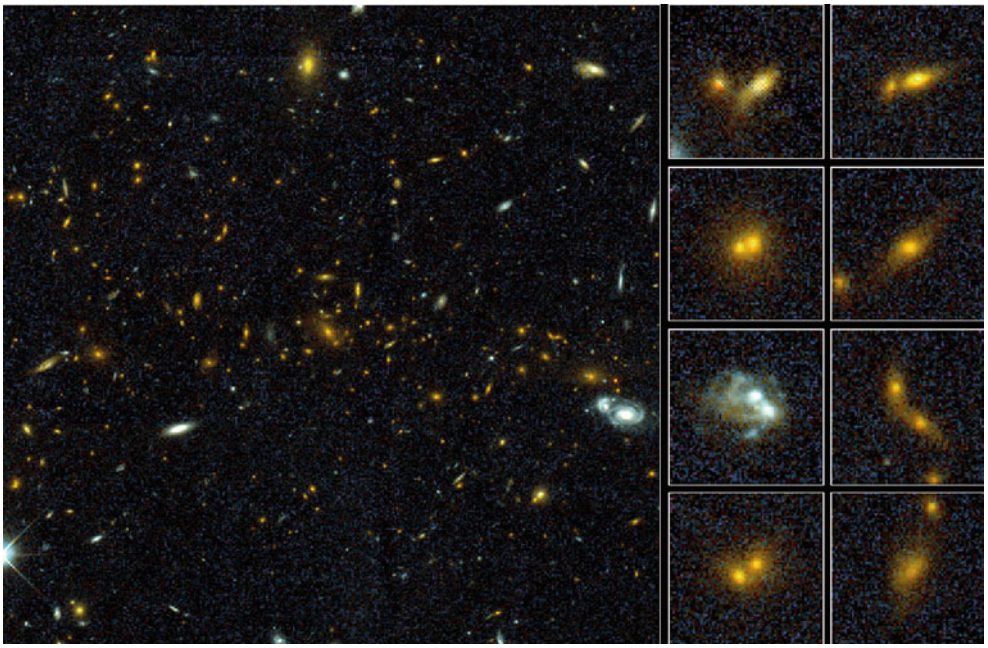


Fig. 6.68 The cluster of galaxies MS 1054−03, observed with the HST, is the most distant cluster in the EMSS X-ray survey ($z = 0.83$). The reddish galaxies in the image on the left form a nearly linear structure. This cluster is far from being spherical, as we have also seen from its weak lensing results (Fig. 6.55)—it is not relaxed. The smaller images on the right show blow-ups of selected cluster fields where mergers of galaxies become visible: in this cluster, the merging of galaxies is directly observable. At least six of the nine merging pairs found in this cluster have been shown to be gravitationally bound systems. Credit: P. van Dokkum, M. Franx/U. Groningen & U. Leiden, ESA, NASA

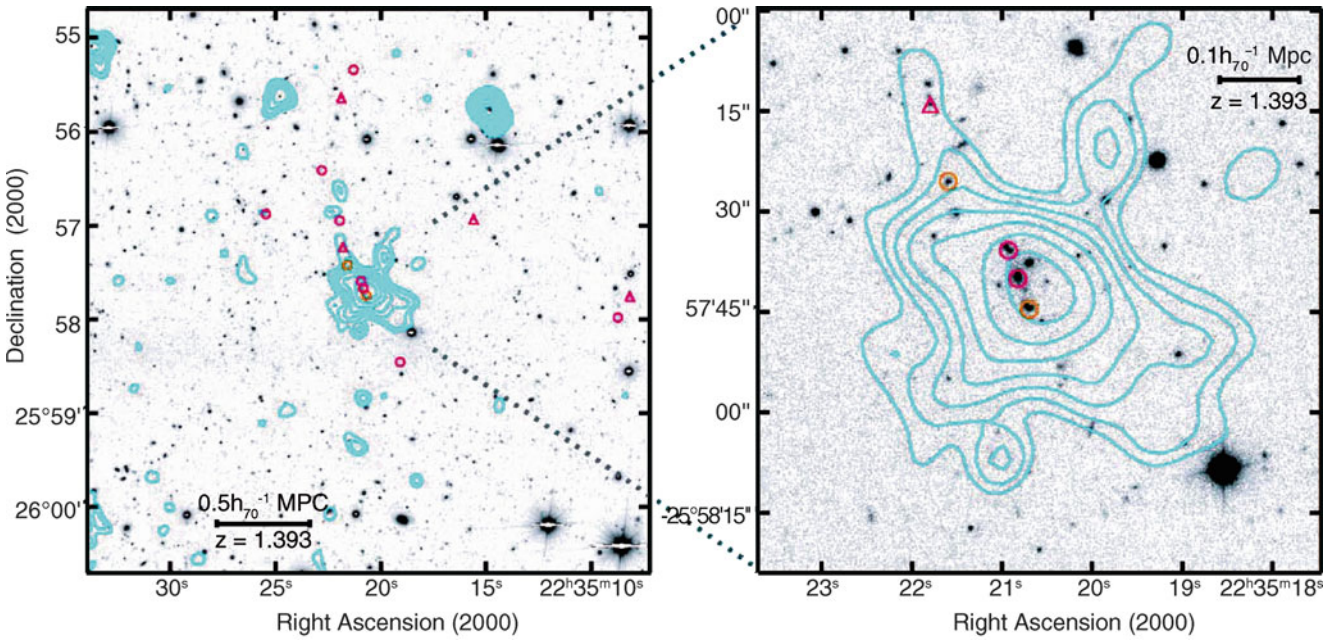


Fig. 6.69 The cluster of galaxies XMMU J2235.2−2557 was discovered in the field-of-view of an XMM-Newton image for which a different source was the original target. The image *on the left* shows the X-ray contours, superposed on an R-band image, while the image *on the right* shows the central section, here superposed on a K-band image. Galaxies in the field follow a red cluster sequence if the color is measured in $R - z$. The *symbols* denote galaxies at redshift $1.37 < z < 1.40$. The strong X-ray source to the upper right of the cluster center is a Seyfert galaxy at lower redshift. Until 2005, this cluster was the most distant X-ray selected cluster known, with a temperature of ~ 6 keV and a velocity dispersion of $\sigma \sim 750$ km/s. Source: C.R. Mullis et al. 2005, *Discovery of an X-Ray-luminous Galaxy Cluster at z=1.4*, ApJ 623, L85, p. L86, Fig. 1.

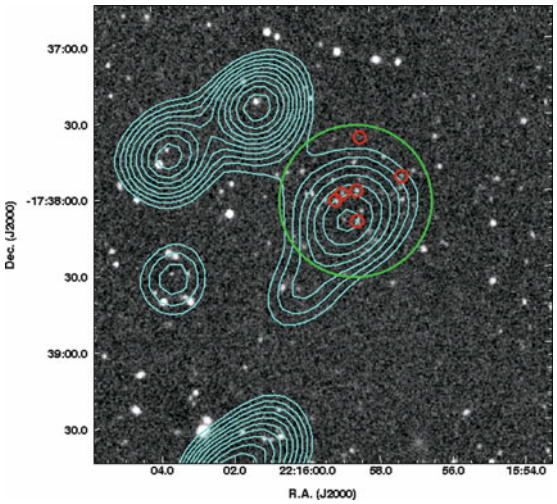


Fig. 6.70 A K_s-band image of the field containing the cluster XMMXCS J2215.9−1738, superposed by the contours of the X-ray emission observed with XMM-Newton. The cluster was discovered by XMM-Newton as an extended X-ray source, close to the center of this $3' \times 3'$ field; the other three nearby X-ray sources are unrelated point sources. The spectra of six member galaxies identified this as a cluster at $z = 1.45$, one of the highest redshift X-ray detected clusters known. The temperature of the cluster is measured to be $\sim 7\,\mathrm{keV}$, signalling a very massive object, a conclusion also obtained from the high X-ray luminosity. Source: S.A. Stanford et al. 2006, *The XMM Cluster Survey: A Massive Galaxy Cluster at z = 1.45*, ApJ 646, L13, p. L14, Fig. 1. ©AAS. Reproduced with permission

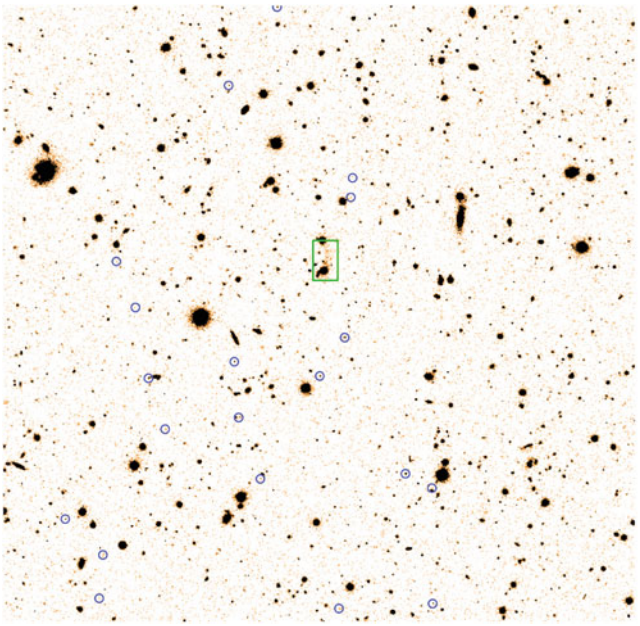
which will only evolve into bound systems during later cosmological evolution; they are thus often called 'proto-clusters'.

6.9 Problems

6.1. The virial radius. In Sect. 7.5.1 we will show, on the basis of the spherical collapse model, that the mean density within a virialized region is about 200 times the critical density of the Universe at the time of collapse. We will derive here an alternative and simple argument for that statement.

1. Consider a circular orbit with radius r_{200} of a particle in a dark matter halo of mass M_{200}. Calculate the orbital time and show that it depends only on the mean density of the halo inside the orbit, $t_{\rm orbit} \propto \bar{\rho}^{-1/2}$. Thus, this orbital time does not depend on the mass (or radius) of the halo. Recall that the free-fall time $t_{\rm ff}$ (see Problem 4.7) has the same dependence on $\bar{\rho}$—what is their ratio?
2. By setting the mean density to be $200\rho_{\rm cr}(z)$, show that the orbital time is, up to a factor of order unity, the age of the Universe at the time of the halo formation, i.e., the time of collapse!

Thus, for a halo characterized by a mean density of 200 times the critical density $\rho_{\rm cr}(z)$, a particle at the outer edge can finish one orbit within the age of the Universe at redshift z. Since the time scale for relaxation cannot be smaller than the



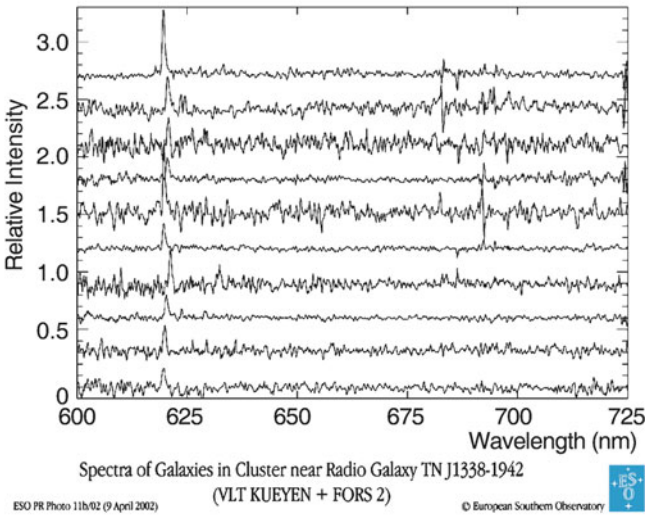


Fig. 6.71 The most distant known group of galaxies. The region around the radio galaxy TN J1338−1942 ($z = 4.1$) was scanned for galaxies at the same redshift; 20 such galaxies were found with the VLT, marked by *circles* in *the left image*. For ten of these galaxies, the spectra are shown *on the right*; in all of them, the Lyα emission line is clearly visible. Hence, groups of galaxies were already formed in an early stage of the Universe. Credits: B. Venemans, G. Miley et al., European Southern Observatory

time a particle needs to orbit through the region, the choice $\bar{\rho} = 200\rho_{\rm cr}(z)$ is indeed plausible.

6.2. Cosmological surface brightness dimming. Consider a source of bolometric luminosity L and radius R, and show that its bolometric surface brightness decreases with redshift $\propto (1+z)^{-4}$. Calculate the redshift dependence of the specific surface brightness I_ν if the source has a power-law spectrum, $L_\nu \propto \nu^{-\alpha}$.

Cosmology II: Inhomogeneities in the Universe

7

7.1 Introduction

In Chap. 4, we discussed homogeneous world models and introduced the standard model of cosmology. It is based on the cosmological principle, the assumption of a (spatially) homogeneous and isotropic universe. Of course, the assumption of homogeneity is justified only on large scales because observations show us that our Universe is inhomogeneous on small scales—otherwise no galaxies or stars would exist.

The distribution of galaxies on the sky is not uniform or random (see Fig. 6.1), rather they form clusters and groups of galaxies. Also clusters of galaxies are not distributed uniformly, but their positions are correlated, grouped together in superclusters. The three-dimensional distribution of galaxies, obtained from redshift surveys, shows an interesting large-scale structure, as can be seen in Fig. 7.1 which shows the spatial distribution of galaxies in the two-degree-Field Galaxy Redshift Survey (2dFGRS).

Even larger structures have been discovered. The Great Wall is a galaxy structure with an extent of $\sim 100h^{-1}$ Mpc, which was found in a redshift survey of galaxies (Fig. 7.2). The largest structure discovered up to now is the Sloan Great Wall, also shown in Fig. 7.2. Such surveys also led to the discovery of the so-called *voids*, nearly spherical regions which contain virtually no (bright) galaxies, and which have a diameter of typically $30h^{-1}$ Mpc. The discovery of these large-scale inhomogeneities raises the question of whether

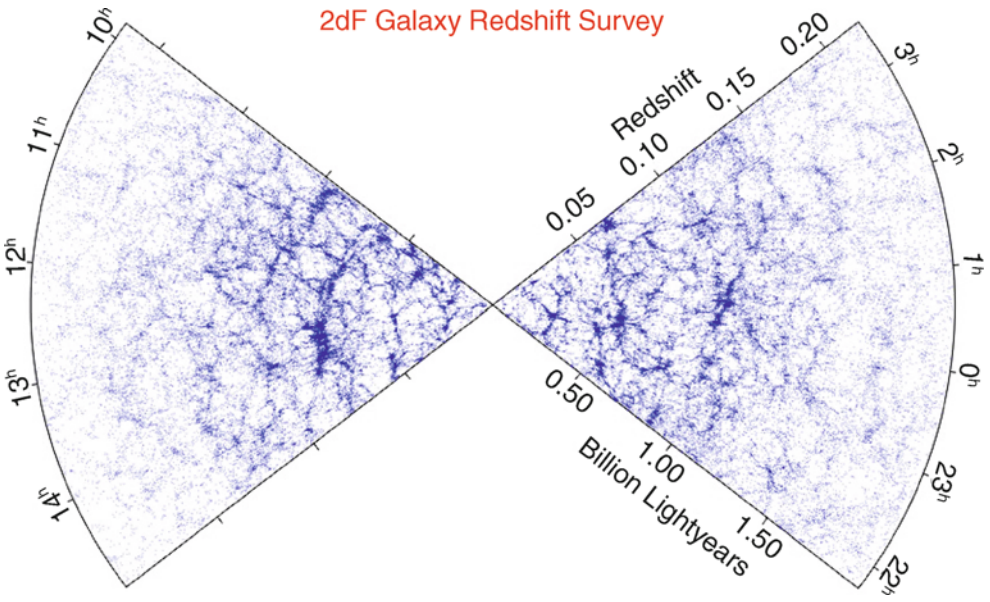


Fig. 7.1 The distribution of galaxies in the complete 2dF Galaxy Redshift Survey. In the radial direction, the escape velocity, or redshift, is plotted, and the polar angle is the right ascension. According the Hubble law, the redshift is directly related to the distance of an object, so that redshift surveys map the three-dimensional distribution of galaxies, with our Galaxy at the center of the figure. In the 2dFGRS, more than 350 000 spectra were taken between 1997 and 2002; plotted here is the distribution of more than 200 000 galaxies with reliable redshift measurements. The data from the complete survey are publicly available. Credit: M. Colless and the 2dF Galaxy Redshift Survey team

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_7,

even larger structures might exist in the Universe, or more precisely: does a scale exist, averaged over which the Universe appears homogeneous? The existence of such a scale is a requirement for the homogeneous world models to provide a realistic description of the mean cosmic behavior.

To date, no evidence of structures with linear dimension well above $100h^{-1}$ Mpc have been found, as can also be seen from Fig. 7.1. Hence, the Universe seems to be basically homogeneous if averaged over scales of $R \sim 200h^{-1}$ Mpc. This 'homogeneity scale' needs to be compared to the Hubble radius $R_{\rm H} \equiv c/H_0 \approx 3000h^{-1}$ Mpc. This implies $R \ll R_{\rm H}$, so that after averaging, $(R_{\rm H}/R)^3 \sim (15)^3 \sim 3000$ independent volume elements exist per Hubble volume. This justifies the approximation of a homogeneous world model when considering the mean cosmic history.

On small scales, the Universe is inhomogeneous. Evidence for this is the galaxy distribution projected on the sky, the three-dimensional galaxy distribution determined by redshift surveys, and the existence of clusters of galaxies, superclusters, 'Great Walls', and voids. In addition, the anisotropy of the cosmic microwave background (CMB), with relative fluctuations of $\Delta T/T \sim 10^{-5}$, indicates that the Universe already contained small inhomogeneities at redshift $z \sim 1000$, which we will discuss more thoroughly in Sect. 8.6. In this chapter, we will examine the evolution of such density inhomogeneities and their description.

7.2 Gravitational instability

7.2.1 Overview

The smallness of the CMB anisotropy suggests that the density inhomogeneities at redshift $z \sim 1000$—this is the epoch where most of the CMB photons interacted with matter for the last time—must have had very small amplitudes. Today, the amplitudes of the density inhomogeneities are considerably larger; for example, a massive cluster of galaxies contains within a radius of $\sim 1.5h^{-1}$ Mpc more than 200 times more mass than an average sphere of this radius in the Universe. Thus, these are no longer small density fluctuations.

Obviously, the Universe became more inhomogeneous in the course of its evolution; as we will see, density perturbations grow over time. One defines the *relative density contrast*

$$\boxed{\delta(\boldsymbol{r},t) := \frac{\rho(\boldsymbol{r},t) - \bar{\rho}(t)}{\bar{\rho}(t)}}\,, \tag{7.1}$$

where $\bar{\rho}(t)$ denotes the mean cosmic matter density at time t. From the definition of δ, one can immediately see that $\delta \geq -1$, because $\rho \geq 0$. The smallness of the CMB anisotropy suggests that at $z \sim 1000$, $|\delta| \ll 1$. The dynamics of the cosmic Hubble expansion is controlled by the gravitational field of the average matter density $\bar{\rho}(t)$, whereas the density fluctuations $\Delta\rho(\boldsymbol{r},t) = \rho(\boldsymbol{r},t) - \bar{\rho}(t)$ generate an additional gravitational field.

We shall here be interested only in very weak gravitational fields, for which the Newtonian description of gravity can be applied. Since the Poisson equation, which specifies the relation between matter density and the gravitational potential, is linear, the effects of the homogeneous matter distribution and of density fluctuations can be considered separately. The gravitational field of the total matter distribution is then the sum of the average matter distribution and that of the density fluctuations.

We consider a region in which $\Delta\rho > 0$, hence $\delta > 0$, so that the gravitational field in this region is stronger than the cosmic average. An overdense region produces a stronger gravitational field than that corresponding to the mean Hubble expansion. By this additional self-gravity, the overdense region will expand more slowly than the average Hubble expansion. Because of the delayed expansion, the density in this region will also decrease more slowly than in the cosmic mean, $\bar{\rho}(t) = (1+z)^3\rho_0 = a^{-3}(t)\rho_0$, and hence the density contrast in this region will increase. As a consequence, the relative density will increase, which again produces an even stronger gravitational field. ... It is obvious that this situation is unstable. Of course, the argument also works the other way round: in an underdense region with $\delta < 0$, the gravitational field generated is weaker than in the cosmic mean, therefore the self-gravity is weaker than that which corresponds to the Hubble expansion. This implies that the expansion is decelerated less than in the cosmic mean, the underdense region expands faster than the Hubble expansion, and thus the local density will decrease more quickly than the mean density of the Universe. In this way, the absolute value of the density contrast increases, i.e., δ becomes more negative over the course of time.

> Density fluctuations grow over time due to their self-gravity; overdense regions increase their density contrast over the course of time, while underdense regions decrease their density contrast. In both cases, $|\delta|$ increases. Hence, this effect of *gravitational instability* leads to an increase of density fluctuations with increasing time. The evolution of structure in the Universe is described by this effect of *gravitational instability*.

Structure growth in the Universe can be understood in the framework of this model. In this chapter we will describe structure formation quantitatively. This includes the analysis of the time evolution of density perturbations, as well as a statistical description of such density fluctuations. We will

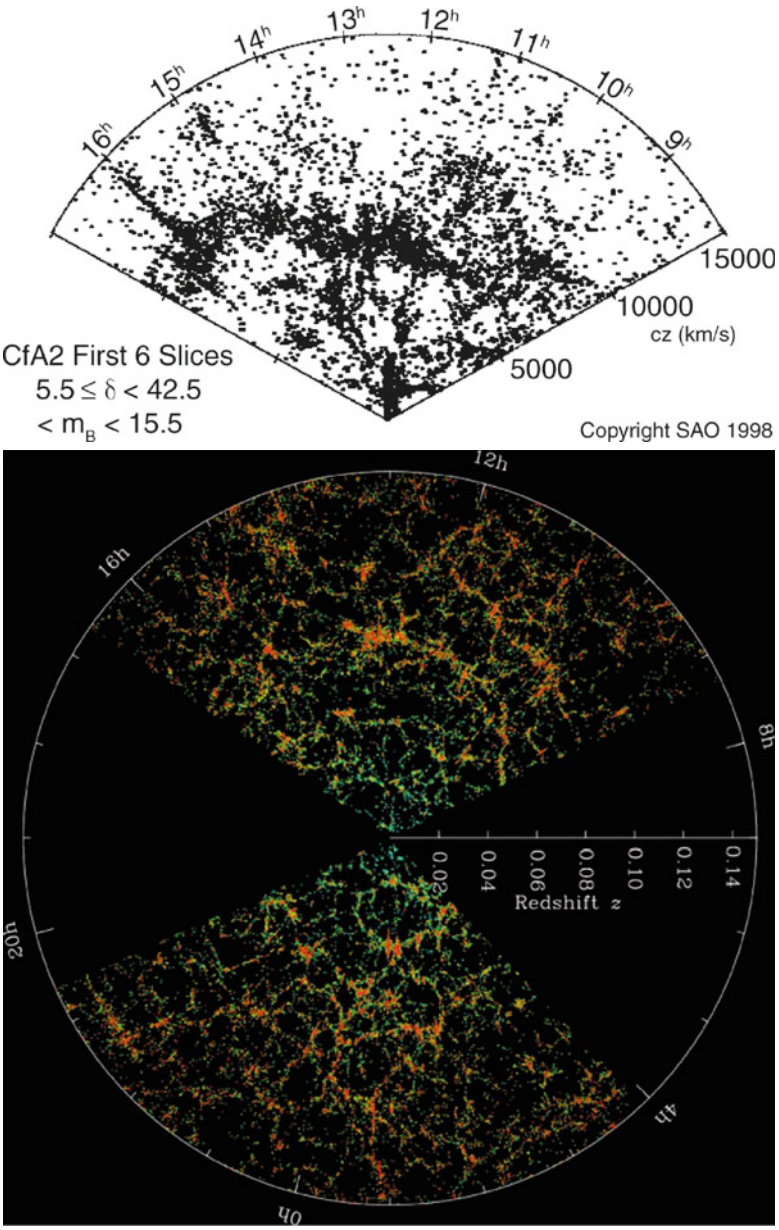


Fig. 7.2 *Top:* In the CfA galaxy redshift survey, carried out in the 1980s, a large coherent structure of galaxies was found, called the Great Wall. Shown are galaxies with radial velocities of $cz \leq 15\,000\,\mathrm{km/s}$, with declination $8.5^\circ \leq \delta \leq 42^\circ$. The Great Wall is located at a redshift of $cz \sim 6000\,\mathrm{km/s}$, extending in right ascension between $9^\mathrm{h} \leq \alpha \leq 16^\mathrm{h}$. *Bottom:* The distribution of galaxies as measured from the Sloan Digital Sky Survey. Plotted are galaxies in the narrow declination range $-1.25^\circ \leq \delta \leq 1.25^\circ$. Note that this distribution extends to considerably larger distances than the one in the CfA survey. The *color of the points* indicates the color of the galaxies. The most remarkable feature seen is the long filament of galaxies near the center of the upper part of the figure, called the Sloan Great Wall. Credit: *Top*: J. Huchra, M. Geller, Harvard-Smithsonian Center for Astrophysics. *Bottom*: M. Blanton and the Sloan Digital Sky Survey

then see that the evolution of inhomogeneities is directly observable, and that the Universe was less inhomogeneous at high redshift than it is today. Since the history of perturbations depends on the cosmological model, we need to examine whether this evolution can be used to obtain an estimate of cosmological parameters. In Chap. 8, we will give an affirmative answer to this question. Finally, we will briefly discuss the origin of density fluctuations.

7.2.2 Linear perturbation theory

We first will examine the growth of density perturbations. For this discussion, we will concentrate on length-scales that are substantially smaller than the Hubble radius. On these scales, structure growth can be described in the framework of the Newtonian theory of gravity. The effects of space-time curvature and thus of General Relativity need to be accounted for only for density perturbations on length-scales comparable to, or larger than the Hubble radius. In addition, we assume for simplicity that the matter in the Universe consists only of dust (i.e., pressure-free matter), with density $\rho(\boldsymbol{r}, t)$. The matter distribution will be described in the *fluid approximation*, where the velocity field of this fluid shall be denoted by $\boldsymbol{v}(\boldsymbol{r}, t)$.[1]

[1] Strictly speaking, the cosmic dust cannot be described as a fluid because the matter is assumed to be collisionless. This means that no interactions occur between the particles, except for gravitation. Two

Equations of motion. The behavior of this fluid is described by the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0 \,, \tag{7.2}$$

which expresses the fact that matter is conserved: the density decreases if the fluid has a diverging velocity field (thus, if particles are moving away from each other). In contrast, a converging velocity field will lead to an increase in density. Furthermore, the Euler equation applies,

$$\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v} \cdot \nabla)\, \boldsymbol{v} = -\frac{\nabla P}{\rho} - \nabla \Phi \,, \tag{7.3}$$

which describes the conservation of momentum and the behavior of the fluid under the influence of forces. The left-hand side of (7.3) is the time derivative of the velocity as would be measured by an observer moving with the flow, because $\partial \boldsymbol{v}/\partial t$ is the derivative at a fixed point in space, whereas the total left-hand side of (7.3) is the time derivative of the velocity measured along the flow lines. The latter is affected by the gravitational field Φ and the pressure gradient. However, since we are only considering pressureless matter, the pressure vanishes, $P \equiv 0$. The gravitational potential satisfies the Poisson equation

$$\nabla^2 \Phi = 4\pi G \rho - \Lambda \,. \tag{7.4}$$

which has been modified compared to the usual Poisson equation to account for the presence of a cosmological constant Λ. We will see in a short while why this form of the Λ-term is chosen.

These three equations for the description of a self-gravitating fluid can in general not be solved analytically. However, we will show that a special, cosmologically relevant exact solution can be found, and that by linearization of the system of equations around this exact solution approximate solutions can be constructed for small relative density contrasts, $|\delta| \ll 1$.

Exact solution: The Hubble expansion. The special exact solution is the flow that we have already encountered in Chap. 4: the homogeneous expanding cosmos. By substituting into the above equations it is immediately shown that

$$\boldsymbol{v}(\boldsymbol{r}, t) = H(t)\boldsymbol{r}$$

is a solution of the equations if ρ is homogeneous and satisfies (4.11), and if the Friedmann equation (4.19) for the scale factor applies (see problem 7.1). Note that the particular form of the Λ-term in the Poisson equation was chosen because with this modification of the 'standard' Poisson equation we can obtain the second Friedmann equation from a Newtonian treatment.

As long as the density contrast $|\delta| \ll 1$, the deviations of the velocity field from the Hubble expansion will be small. We expect that in this case, physically relevant solutions of the above equations are those which deviate only slightly from the homogeneous case.

It is convenient to consider the problem in comoving coordinates; hence we define, as in (4.4),

$$\boldsymbol{r} = a(t)\, \boldsymbol{x} \,.$$

In a homogeneous cosmos, $\boldsymbol{x}$ is a constant for every matter particle, and its spatial position $\boldsymbol{r}$ changes only due to the Hubble expansion. Likewise, the velocity field is written in the form

$$\boldsymbol{v}(\boldsymbol{r}, t) = \frac{\dot{a}}{a}\boldsymbol{r} + \boldsymbol{u}\left(\frac{\boldsymbol{r}}{a}, t\right) \,, \tag{7.5}$$

where $\boldsymbol{u}(\boldsymbol{x}, t)$ is a function of the comoving coordinate $\boldsymbol{x}$. In (7.5), the first term represents the homogeneous Hubble expansion, whereas the second term describes the deviations from this homogeneous expansion. For this reason, $\boldsymbol{u}$ is called the *peculiar velocity*. Next, we will show how the above equations read in comoving coordinates.

Transforming the fluid equations to comoving coordinates. We first note that the partial derivative $\partial/\partial t$ in (7.2) means a time derivative at fixed $\boldsymbol{r}$. If the equations are to be written in comoving coordinates, this partial time derivative needs to be transformed into one where $\boldsymbol{x}$ is kept fixed. For example,

$$\begin{aligned}\left(\frac{\partial}{\partial t}\right)_r \rho(\boldsymbol{r}, t) &= \left(\frac{\partial}{\partial t}\right)_r \rho_x\left(\frac{\boldsymbol{r}}{a}, t\right) \\ &= \left(\frac{\partial}{\partial t}\right)_x \rho_x(\boldsymbol{x}, t) - \frac{\dot{a}}{a}\, \boldsymbol{x} \cdot \nabla_x \rho_x(\boldsymbol{x}, t) \,,\end{aligned} \tag{7.6}$$

where ∇_x is the gradient with respect to comoving coordinates, and where we define the function $\rho_x(\boldsymbol{x}, t) \equiv \rho(a\boldsymbol{x}, t)$. Note that $\rho_x(\boldsymbol{x}, t)$ and $\rho(\boldsymbol{x}, t)$ both describe the same *physical* density field, but that ρ and ρ_x are different *mathematical* functions of their arguments. After these transformations, (7.2) becomes

$$\frac{\partial \rho}{\partial t} + \frac{3\dot{a}}{a}\rho + \frac{1}{a}\nabla \cdot (\rho \boldsymbol{u}) = 0 \,, \tag{7.7}$$

where from now on all spatial derivatives are to be considered with respect to $\boldsymbol{x}$. For notational simplicity, from now on we also set $\rho \equiv \rho_x$ and $\delta \equiv \delta(\boldsymbol{x}, t)$, and note that the partial time derivative is to be understood to be at fixed $\boldsymbol{x}$. Writing $\rho = \bar{\rho}(1 + \delta)$ and using $\bar{\rho} \propto a^{-3}$, (7.7) reads in comoving coordinates

flows of such dust can thus penetrate each other. This situation can be compared to that of a fluid whose molecules are interacting by collisions. Through these collisions, the velocity distribution of the molecules will, at each position, assume an approximate Maxwell distribution, with a well-defined average velocity that corresponds to the flow velocity at this point. Such an unambiguous velocity does not exist for dust in general. However, at early times, when deviations from the Hubble flow are still very small, no multiple flows are expected, so that in this case, the velocity field is well defined.

$$\frac{\partial\delta}{\partial t} + \frac{1}{a}\nabla\cdot[(1+\delta)\,\boldsymbol{u}] = 0\,. \tag{7.8}$$

Accordingly, the gravitational potential Φ is written as

$$\Phi(\boldsymbol{r},t) = \left(\frac{2\pi}{3}G\bar{\rho}(t) - \frac{\Lambda}{6}\right)|\boldsymbol{r}|^2 + \phi(\boldsymbol{x},t) = \frac{\ddot{a}a}{2}\,|\boldsymbol{x}|^2 + \phi(\boldsymbol{x},t)\,; \tag{7.9}$$

the first term is the Newtonian potential for a homogeneous density field, and ϕ satisfies the Poisson equation for the density inhomogeneities,

$$\begin{aligned}\nabla^2\phi(\boldsymbol{x},t) &= 4\pi G a^2(t)\bar{\rho}(t)\delta(\boldsymbol{x},t)\\ &= \frac{3H_0^2\Omega_{\rm m}}{2a(t)}\,\delta(\boldsymbol{x},t)\,,\end{aligned} \tag{7.10}$$

where in the last step we used $\bar{\rho}\propto a^{-3}$ and the definition of the density parameter $\Omega_{\rm m}$. Then, the Euler equation (7.3) becomes

$$\frac{\partial\boldsymbol{u}}{\partial t} + \frac{\boldsymbol{u}\cdot\nabla}{a}\boldsymbol{u} + \frac{\dot{a}}{a}\boldsymbol{u} = -\frac{1}{\bar{\rho}\,a}\nabla P - \frac{1}{a}\nabla\phi\,, \tag{7.11}$$

where (4.13) has been utilized.

Linearization. In the homogeneous case, $\delta\equiv 0$, $\boldsymbol{u}\equiv 0$, $\phi\equiv 0$, $\rho=\bar{\rho}$, and (7.7) then implies $\dot{\bar{\rho}}+3H\bar{\rho}=0$, which also follows immediately from (4.17) in the case of $P=0$. Now we will look for approximate solutions of the above set of equations which describe only small deviations from this homogeneous solution. For this reason, in these equations we only consider first-order terms in the small parameters δ and $\boldsymbol{u}$, i.e., we disregard terms that contain $\boldsymbol{u}\,\delta$ or are quadratic in the velocity $\boldsymbol{u}$, i.e., the second term on the l.h.s. of (7.11). The linearized continuity and Euler equations then read

$$\frac{\partial\delta}{\partial t} + \frac{1}{a}\nabla\cdot\boldsymbol{u} = 0\,, \tag{7.12}$$

$$\frac{\partial\boldsymbol{u}}{\partial t} + \frac{\dot{a}}{a}\boldsymbol{u} = -\frac{1}{a}\nabla\phi\,, \tag{7.13}$$

where we set $P=0$, and the Poisson equation (7.10) is linear. Combining this set of equations, we can eliminate the peculiar velocity $\boldsymbol{u}$ and the gravitational potential ϕ from the equations[2] and then obtain a second-order differential equation for the density contrast δ,

$$\frac{\partial^2\delta}{\partial t^2} + \frac{2\dot{a}}{a}\frac{\partial\delta}{\partial t} = 4\pi G\bar{\rho}\delta\,. \tag{7.14}$$

It is remarkable that neither does this equation contain derivatives with respect to spatial coordinates, nor do the coefficients in the equation depend on $\boldsymbol{x}$. Therefore, (7.14) has solutions of the form

$$\delta(\boldsymbol{x},t) = D(t)\,\tilde{\delta}(\boldsymbol{x})\,,$$

i.e., the spatial and temporal dependencies factorize in these solutions. Here, $\tilde{\delta}(\boldsymbol{x})$ is an arbitrary function of the spatial coordinate, and $D(t)$ satisfies the equation

$$\ddot{D} + \frac{2\dot{a}}{a}\dot{D} - 4\pi G\bar{\rho}(t)\,D = 0\,. \tag{7.15}$$

The growth factor. The differential equation (7.15) has two linearly independent solutions. One can show that one of them increases with time, whereas the other decreases (we will see a special example of this below). If, at some early time, both functional dependencies were present, the increasing solution will dominate at later times, whereas the solution decreasing with t will become irrelevant. Therefore, we will consider only the increasing solution, which is denoted by $D_+(t)$, and normalize it such that $D_+(t_0)=1$. Then, the density contrast becomes

$$\boxed{\delta(\boldsymbol{x},t) = D_+(t)\,\delta_0(\boldsymbol{x})}\,. \tag{7.16}$$

This mathematical consideration allows us to draw immediately a number of conclusions. First, the solution (7.16) implies that in linear perturbation theory *the spatial shape of the density fluctuations is frozen in comoving coordinates*, only their amplitude increases. The *growth factor* $D_+(t)$ of the amplitude follows a simple differential equation that is readily solvable for any cosmological model. In fact, one can show (see problem 7.2) that for arbitrary values of the density parameters in matter and vacuum energy, the growth factor has the form

$$D_+(a) \propto \frac{H(a)}{H_0}\int_0^a \frac{{\rm d}a'}{\left[\Omega_{\rm m}/a' + \Omega_\Lambda a'^2 - (\Omega_{\rm m}+\Omega_\Lambda-1)\right]^{3/2}}\,, \tag{7.17}$$

where the factor of proportionality is determined from the condition $D_+(t_0)=1$.

In accordance with $D_+(t_0)=1$, $\delta_0(\boldsymbol{x})$ would be the distribution of density fluctuations today if the evolution was indeed linear until the present epoch. Therefore, $\delta_0(\boldsymbol{x})$ is denoted as the *linearly extrapolated density fluctuation field*. However, the linear approximation breaks down if $|\delta|$ is no longer $\ll 1$. In this case, the terms that have been neglected in the above derivations are no longer small and have to be included. The problem then becomes *considerably* more difficult and defies analytical treatment. Instead one needs, in general, to rely on numerical procedures for analyzing the growth of density perturbations. Furthermore, it shall be

[2] This is done by first taking the divergence of (7.13), $\nabla\cdot\dot{\boldsymbol{u}} = -(\dot{a}/a)\nabla\cdot\boldsymbol{u} - (1/a)\nabla^2\phi = \dot{a}\delta - (1/a)\nabla^2\phi$, where we made use of (7.12) in the second step. Taking the time derivative of (7.12) yields $\ddot{\delta} = (\dot{a}/a^2)\nabla\cdot\boldsymbol{u} - (1/a)\nabla\cdot\dot{\boldsymbol{u}} = -2(\dot{a}/a)\dot{\delta} + (1/a^2)\nabla^2\phi$. Finally, the Poisson equation is employed.

noted once again that, for large density perturbations, the fluid approximation is no longer valid, and that up to now we have assumed that the pressure can be neglected. At early times, i.e., for $z \gtrsim z_{\rm eq}$ [see (4.58)], this assumption becomes invalid, so that the above equations need to be modified for these early epochs.

Example: Einstein–de Sitter model. In the special case of a world model with $\Omega_{\rm m} = 1$, $\Omega_\Lambda = 0$, (7.15) can be solved explicitly. In this case, $a(t) = (t/t_0)^{2/3}$, so that

$$\left(\frac{\dot a}{a}\right) = \frac{2}{3t}\ , \text{ and } \bar\rho(t) = a^{-3}\rho_{\rm cr} = \frac{3H_0^2}{8\pi G}\left(\frac{t}{t_0}\right)^{-2}\ ;$$

furthermore, in this model $t_0\, H_0 = 2/3$, so that (7.15) reduces to

$$\ddot D + \frac{4}{3t}\dot D - \frac{2}{3t^2} D = 0\ . \tag{7.18}$$

This equation is easily solved by making the ansatz $D \propto t^q$; this ansatz is suggested because (7.18) is equidimensional in t, i.e., each term has the dimension $D/(\text{time})^2$. Inserting into (7.18) yields a quadratic equation for q,

$$q(q-1) + \frac{4}{3}q - \frac{2}{3} = 0\ ,$$

with solutions $q = 2/3$ and $q = -1$. The latter corresponds to fluctuations decreasing with time and will be disregarded in the following. So, for the Einstein–de Sitter model, the increasing solution

$$D_+(t) = \left(\frac{t}{t_0}\right)^{2/3} = a(t) \tag{7.19}$$

is found, i.e., in this case the growth factor equals the scale factor. For different cosmological parameters this is not the case, but the qualitative behavior is quite similar, which is demonstrated in Fig. 7.3 for three models. In particular, fluctuations were able to grow by a factor ~ 1000 from the epoch of recombination at $z \sim 1000$, from which the CMB photons originate, to the present day.

Evidence for dark matter on cosmic scales. At the present epoch, $\delta \gg 1$ certainly on scales of clusters of galaxies (~ 2 Mpc), and $\delta \sim 1$ on scales of superclusters (~ 10 Mpc). Hence, because of the law of linear structure growth (7.16) and the behavior of $D_+(t)$ shown in Fig. 7.3, we would expect $\delta \gtrsim 10^{-3}$ at $z = 1000$ for these structures to be able to grow to non-linear structures at the current epoch. For this reason, we should also expect CMB fluctuations to be of comparable magnitude, $\Delta T/T \gtrsim 10^{-3}$. The observed

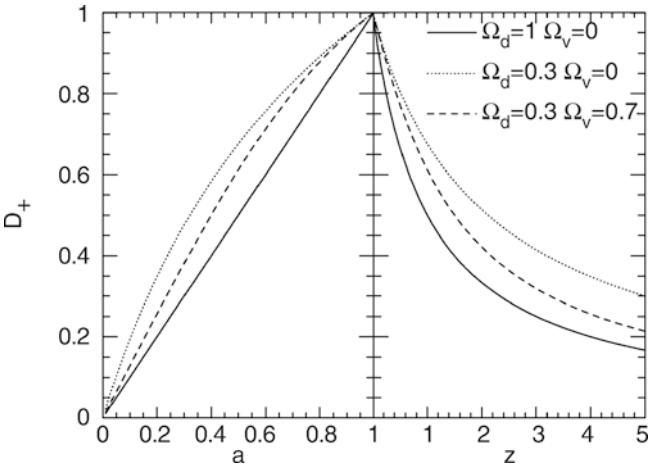


Fig. 7.3 Growth factor D_+ for three different cosmological models, as a function of the scale factor a (*left panel*) and of redshift (*right panel*). It is clearly visible how quickly D_+ decreases with increasing redshift in the EdS-model, in comparison to the models of lower density

fluctuation amplitude is much smaller, $\Delta T/T \sim 10^{-5}$, however. The corresponding density fluctuations therefore cannot have grown sufficiently strongly up to today to form non-linear structures.

This contradiction can be resolved by the dominance of dark matter. Since photons interact with baryonic matter only, the CMB anisotropies basically provide (at least on angular scales below $\sim 1°$) information on the density contrast of *baryons*. Dark matter may have had a higher density contrast at recombination, but the baryons, which are strongly coupled to the radiation field before recombination, are prevented from strong clustering due to the radiation pressure. Only after recombination, when the electrons have combined with the atomic nuclei and essentially no free electrons remain, the coupling to the radiation field ends, after which the baryons may fall into the potential wells formed by the dark matter. We will return to this issue in Sect. 7.4.3.

7.2.3 Peculiar velocities

As mentioned on several occasions before, cosmic sources do not exactly follow the Hubble expansion, but have an additional peculiar velocity. Deviations from the Hubble flow are caused by local gravitational fields, and such fields are in turn generated by local density fluctuations. These inevitably lead to an acceleration, which affects the matter and generates peculiar velocities. Indeed, we see from (7.12) that a time-dependent density contrast δ implies the presence of peculiar velocities. In numerical simulations, the peculiar velocities of individual particles are followed in the computations automatically. In this brief section, we will investigate

the large-scale peculiar velocities as they are derived from linear perturbation theory.

Since the spatial dependence of the density contrast δ is constant in time, $\delta(\boldsymbol{x},t) = \delta_0(\boldsymbol{x})\,D_+(t)$ [see (7.16)], the acceleration vector $\boldsymbol{g}$ has a constant direction in the framework of linear perturbation theory. Hence, one obtains the peculiar velocity in the form

$$\boldsymbol{u}(\boldsymbol{x}) \propto \int \mathrm{d}t\ \boldsymbol{g}(\boldsymbol{x},t)\ ,$$

i.e., parallel to $\boldsymbol{g}(\boldsymbol{x})$. On the other hand, $\boldsymbol{g}(\boldsymbol{x})$ is the gradient of the gravitational potential, $\boldsymbol{g} \propto -\nabla\phi$. This implies that $\boldsymbol{u}(\boldsymbol{x})$ is a gradient field, i.e., a scalar function $\psi(\boldsymbol{x})$ exists such that $\boldsymbol{u} = \nabla\psi$, where the gradient is taken with respect to the comoving spatial coordinate $\boldsymbol{x}$. Therefore, $\nabla\cdot\boldsymbol{g} \propto -\nabla^2\phi \propto -\delta$, where the Poisson equation (7.10) has been utilized. Thus we conclude that $\nabla\cdot\boldsymbol{u} \propto -\delta$, i.e., the divergence of the peculiar velocity field can be expressed in terms of the density contrast. In the following, we will obtain this result in a more quantitative way.

We consider the growing mode of density perturbations, for which (7.16) implies that

$$\frac{\partial\delta}{\partial t} = \frac{\dot{D}_+}{D_+}\,\delta\ .$$

Inserting this result into the linearized continuity equation (7.12) yields

$$\begin{aligned}\nabla\cdot\boldsymbol{u} &= -a\,\frac{\dot{D}_+}{D_+}\,\delta = -a\,\dot{a}\,\frac{1}{D_+}\,\frac{\mathrm{d}D_+}{\mathrm{d}a}\,\delta \\ &= -a\,H(a)\,f(a)\,\delta\ ,\end{aligned} \tag{7.20}$$

where we replaced the time derivative of D_+ by a derivative with respect to the scale factor. In the last step, we defined the function

$$f(a) := \frac{a}{D_+}\,\frac{\mathrm{d}D_+}{\mathrm{d}a} = \frac{\mathrm{d}\log D_+}{\mathrm{d}\log a}\ . \tag{7.21}$$

The function f can be calculated explicitly from (7.17). It turns out that the resulting expression can be very accurately approximated by

$$f(a) \approx \Omega_\mathrm{m}^\gamma(a)\ , \tag{7.22}$$

where $\Omega_\mathrm{m}(a)$ is the redshift-dependent matter density parameter,

$$\Omega_\mathrm{m}(a) = \frac{\rho_\mathrm{m}(a)}{\rho_\mathrm{cr}(a)} = \Omega_{\mathrm{m},0}\left[a^3\left(\frac{H}{H_0}\right)^2\right]^{-1}\ . \tag{7.23}$$

Here we used the expression for the redshift-dependent critical density given in (4.80), and we explicitly wrote $\Omega_{\mathrm{m},0}$ for the current-epoch value of the density parameter, usually simply called Ω_m. The parameter γ is in the range of 0.55–0.6 for a large variety of cosmological parameters. Therefore, one usually approximates $f(a) \approx \Omega_\mathrm{m}^{0.6}(a)$. Introducing, as before, the velocity potential ψ by $\boldsymbol{u} = \nabla\psi$, we obtain from (7.20)

$$\nabla^2\psi \approx -a\,H(a)\,\Omega_\mathrm{m}^{0.6}(a)\,\delta\ . \tag{7.24}$$

This Poisson equation for ψ can be solved, and by computing the gradient of the solution, the peculiar velocity field can be calculated,

$$\boldsymbol{u}(\boldsymbol{x},t) = \frac{\Omega_\mathrm{m}^{0.6}(a)}{4\pi}\,a\,H(a)\int \mathrm{d}^3y\ \delta(\boldsymbol{y},t)\,\frac{\boldsymbol{y}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|^3}\ . \tag{7.25}$$

This equation shows that the velocity field can be derived from the density field. If the density field in the Universe was observable, one would obtain a direct prediction for the corresponding velocity field from the above relations. This depends on the matter density Ω_m, so that from a comparison with the observed velocity field, one could estimate the value for Ω_m. We will come back to this in Sect. 8.1.8.

7.3 Description of density fluctuations

We will now examine the question of how to describe an inhomogeneous universe quantitatively, i.e., how to quantify the structures it contains. This task sounds easier at first sight than it is in reality. One has to realize that the aim of such a theoretical description cannot be to describe the complete function $\delta(\boldsymbol{x},t)$ for a particular universe. No cosmological model will be able to describe, for instance, the matter distribution in the vicinity of the Milky Way in detail. No model based on the laws of physics alone will be able to predict that at a distance of $\sim$ 800 kpc from the Galaxy a second massive spiral galaxy is located, because this specific feature of our local Universe depends on the specific initial conditions of the matter distribution in the early Universe. We can at best hope to predict the *statistical properties* of the mass distribution, such as, for example, the average number density of clusters of galaxies above a given mass, or the probability of a massive galaxy being found within 800 kpc of another one. Likewise, cosmological simulations (see below) cannot predict *our* Universe; instead, they are at best able to generate cosmological mass distributions that have the same statistical properties as that in our Universe.

It is quite obvious that a very large number of statistical properties exist for the density field, all of which we can

examine and which we hope can be explained quantitatively by the correct model of cosmological structure formation. To make any progress at all, the statistical properties need to be sorted or classified. How can the statistical properties of a density field best be described?

Two universes are considered equivalent if their density fields δ have the same statistical properties. One may then imagine considering a large (statistical) ensemble of universes whose density fields all have the same statistical properties, but for which the individual functions $\delta(\boldsymbol{x})$ can all be different. This statistical ensemble is called a *random field*, and any individual distribution with the respective statistical properties is called a *realization of the random field*.

An example may clarify these concepts. We consider the waves on the surface of a large lake. The statistical properties of these waves—such as how many of them there are with a certain wavelength, and how their amplitudes are distributed—depend on the shape of the lake, its depth, and the strength and direction of the wind blowing over its surface. If we assume that the wind properties are not changing with time, the statistical properties of the water surface are constant over time. Of course, this does not mean that the amplitude of the surface height as a function of position is constant. Rather, it means that two photographs of the surface that are taken at different times are statistically indistinguishable: the distribution of the wave amplitudes will be the same, and there is no way of deciding which of the snapshots was taken first. Knowing the surface topography and the wind properties sufficiently well, one is able to compute the distribution of the wave amplitudes, but there is no way to predict the amplitude of the surface of the lake as a function of position at a particular time. Each snapshot of the lake is a realization of the random field, which in turn is characterized by the statistical properties of the waves.

7.3.1 Correlation functions

Galaxies are not randomly distributed in space, but rather they gather in groups, clusters, or even larger structures. Phrased differently, this means that the probability of finding a galaxy at location $\boldsymbol{x}$ is not independent of whether there is a galaxy at a neighboring point $\boldsymbol{y}$. It is more probable to find a galaxy in the vicinity of another one than at an arbitrary location.[3] This phenomenon is described such that one considers two points $\boldsymbol{x}$ and $\boldsymbol{y}$, and two volume elements $\mathrm{d}V$ around these points. If $\bar{n}$ is the average number density of galaxies, the probability of finding a galaxy in the volume element $\mathrm{d}V$ around $\boldsymbol{x}$ is then

$$P_1 = \bar{n}\,\mathrm{d}V\ ,$$

independent of $\boldsymbol{x}$ if we assume that the Universe is statistically homogeneous. We choose $\mathrm{d}V$ such that $P_1 \ll 1$, so that the probability of finding two or more galaxies in this volume element is negligible.

The probability of finding a galaxy in the volume element $\mathrm{d}V$ at location $\boldsymbol{x}$ and at the same time finding a galaxy in the volume element $\mathrm{d}V$ at location $\boldsymbol{y}$ is then

$$P_2 = (\bar{n}\,\mathrm{d}V)^2\left[1 + \xi_{\mathrm{g}}(\boldsymbol{x},\boldsymbol{y})\right]\ . \tag{7.26}$$

If the distribution of galaxies was uncorrelated, the probability P_2 would simply be the product of the probabilities of finding a galaxy at each of the locations $\boldsymbol{x}$ and $\boldsymbol{y}$ in a volume element $\mathrm{d}V$, so $P_2 = P_1^2$. But since the distribution is correlated, the relation does not apply in this simple form; rather, it needs to be modified, as was done in (7.26). Equation (7.26) defines the *two-point correlation function* (or simply 'correlation function') of galaxies $\xi_{\mathrm{g}}(\boldsymbol{x},\boldsymbol{y})$.

By analogy to this, the correlation function for the total matter density can be defined as

$$\begin{aligned}\langle\rho(\boldsymbol{x})\,\rho(\boldsymbol{y})\rangle &= \bar{\rho}^2\,\langle[1+\delta(\boldsymbol{x})]\,[1+\delta(\boldsymbol{y})]\rangle \\ &= \bar{\rho}^2\,(1+\langle\delta(\boldsymbol{x})\,\delta(\boldsymbol{y})\rangle) \\ &=: \bar{\rho}^2\,[1+\xi(\boldsymbol{x},\boldsymbol{y})]\ ,\end{aligned} \tag{7.27}$$

because the mean (or expectation) value $\langle\delta(\boldsymbol{x})\rangle = 0$ for all locations $\boldsymbol{x}$, as can be seen from the definition (7.1).

In the above equations, angular brackets denote averaging over an ensemble of distributions that all have identical statistical properties. In our example of the lake, the correlation function of the wave amplitudes at positions $\boldsymbol{x}$ and $\boldsymbol{y}$, for instance, would be determined by taking a large number of snapshots of its surface and then averaging the product of the amplitudes at these two locations over all these realizations.

Since the Universe is considered statistically homogeneous, ξ can only depend on the difference $\boldsymbol{x}-\boldsymbol{y}$ and not on $\boldsymbol{x}$ and $\boldsymbol{y}$ individually. Furthermore, ξ can only depend on the separation $r = |\boldsymbol{x}-\boldsymbol{y}|$, and not on the direction of the separation vector $\boldsymbol{x}-\boldsymbol{y}$ because of the assumed statistical isotropy of the Universe. Therefore, $\xi = \xi(r)$ is simply a function of the separation between two points.

For a homogeneous random field, the ensemble average can be replaced by spatial averaging, i.e., the correlation function can be determined by averaging over the products of densities at pairs of points, for a large number of pairs of points with given separation r. For determining the corre-

[3] An every-day life example of clustering is the following: the population density of many European countries is of order 100 people per km^2, and so the mean separation between two people is on the order of 100 m. For those of you who live in a town or a city, the first morning view from the window shows that typically, you find many people within that distance range, an experience strengthened once you get into your car or use public transport. Obviously, people are highly clustered.

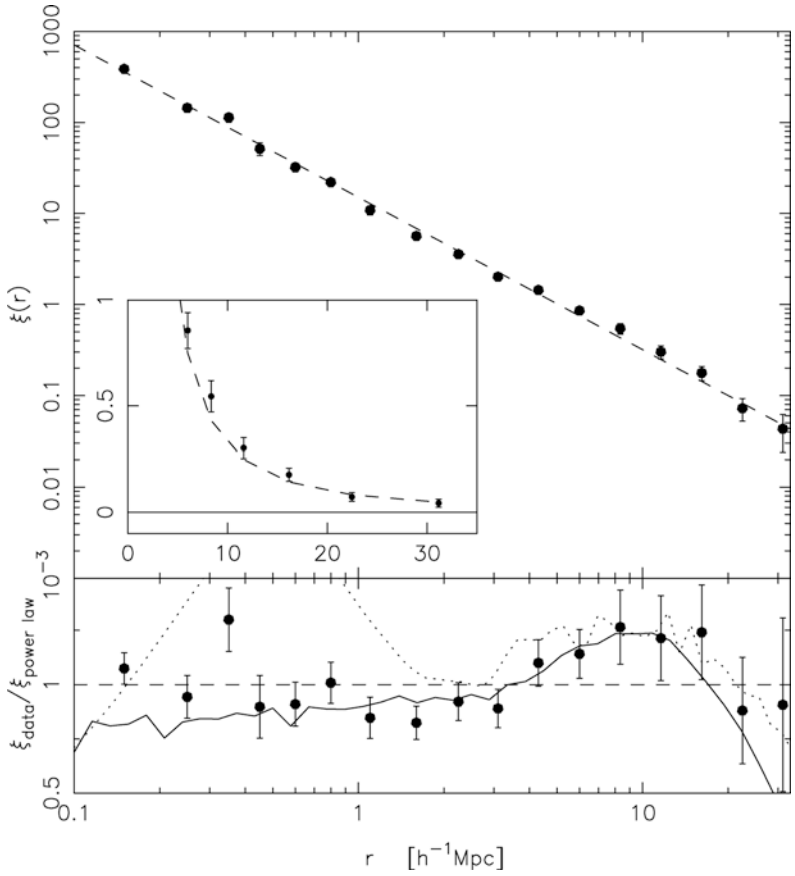


Fig. 7.4 The correlation function ξ_g of galaxies, as it was determined from the 2dF Galaxy Redshift Survey shown in Fig. 7.1. The *top panel* shows $\xi_g(r)$ as obtained from the survey (*points*), together with the best-fitting power law, $\xi_g(r) = (r/r_0)^{-\gamma}$, with correlation length $r_0 = 5.05h^{-1}$ Mpc and slope $\gamma = 1.67$. The *bottom panel* shows the ratio between the measured data points and the power-law fit (*points*), and the corresponding ratio for an earlier result obtained from a deprojection of the angular correlation function of a photometric survey (*solid curve*) and the correlation function obtained from the Hubble Volume simulation, an N-body simulation (see Fig. 7.12). Source: E. Hawkins et al. 2003, *The 2dF Galaxy Redshift Survey: correlation functions, peculiar velocities and the matter density of the Universe*, MNRAS 346, 78, p. 86, Fig. 11. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

lation function of galaxies, we note that $\xi_g(r)$ is the excess probability to find a galaxy at a separation r from another galaxy, relative to that of a random distribution. Therefore, $\xi_g(r)$ can be determined by first counting the number of galaxy pairs with separation in the interval Δr around r. Then one creates a random distribution of the same number of objects in the same volume, and again counts the pairs in the same distance interval. The ratio of these two pair counts then yields an estimate for $\xi_g(r)$.[4]

The equivalence of ensemble average and spatial average is called the ergodicity of the random field. Only by this can the correlation function (and all other statistical properties) in our Universe be measured at all, because we are able to observe only a single—namely our—realization of the hypothetical ensemble. From the measured correlations between galaxy positions, as determined from spectroscopic redshift surveys of galaxies (see Sect. 8.1.2), one finds the approximate relation

$$\xi_g(r) = \left(\frac{r}{r_0}\right)^{-\gamma} \tag{7.28}$$

for galaxies of luminosity $\sim L^*$ (see Fig. 7.4), where $r_0 \simeq 5h^{-1}$ Mpc denotes the correlation length, and where the slope is about $\gamma \simeq 1.7$. This relation is approximately valid over a range of separations $0.2h^{-1}$ Mpc $\lesssim r \lesssim 30h^{-1}$ Mpc.

Hence, the correlation function provides a means to characterize the structure of the cosmological matter distribution. Besides this two-point correlation function, correlations of higher order may also be defined, leading to general n-point correlation functions. These are more difficult to determine from observation, though. It can be shown that the statistical properties of a random field are fully specified by the set of all n-point correlations.

[4] This method is indeed used for estimating the galaxy correlation function, although with some important modifications to increase its accuracy and efficiency.

7.3.2 The power spectrum

An alternative (and equivalent) description of the statistical properties of a random field, and thus of the matter distribution in a universe, is the *power spectrum* $P(k)$. Roughly speaking, the power spectrum $P(k)$ describes the level of structure as a function of the length-scale $L \simeq 2\pi/k$; the larger $P(k)$, the larger the amplitude of the fluctuations on a length-scale $2\pi/k$. Here, k is a comoving *wave number*. Phrased differently, the density fluctuations are decomposed into a sum of plane waves of the form $\delta(\boldsymbol{x}) = \sum a_{\boldsymbol{k}} \cos(\boldsymbol{x} \cdot \boldsymbol{k})$, with a wave vector $\boldsymbol{k}$ and an amplitude $a_{\boldsymbol{k}}$. The power spectrum $P(k)$ then describes the mean of the squares, $|a_{\boldsymbol{k}}|^2$, of the amplitudes, averaged over all wave vectors with equal length $k = |\boldsymbol{k}|$. Technically speaking, this is a Fourier decomposition. Referring back to the example of waves on the surface of a lake, one finds that a characteristic wavelength L_c exists, which depends, among other factors, on the wind speed. In this case, the power spectrum will have a prominent maximum at $k = 2\pi/L_c$.

The power spectrum $P(k)$ and the correlation function are related through a Fourier transform; formally, one has[5]

$$P(k) = 2\pi \int_0^\infty \mathrm{d}x \; x^2 \, \frac{\sin(kx)}{kx} \, \xi(x) \; , \tag{7.29}$$

i.e., the integral over the correlation function with a weight factor depending on $k \sim 2\pi/L$. This relation can also be inverted, and thus $\xi(x)$ can be computed from $P(k)$.

In cosmology, one uses both concepts, correlation function and power spectrum, side-by-side. One of the two may be easier to determine from observational data, another one may be more straightforward to obtain from a model or from simulations. Also, our intuitive understanding of these two concepts may vary in different situations. Probably, for most non-cosmologists the concept of a correlation function is easier to grasp than that of a power spectrum. There are situations, however, where it is the opposite: The fluctuations of the pressure of air at a fixed point as a function of time, $\Delta P(t)$, is a function with a non-vanishing correlation $\xi(\tau) = \langle \Delta P(t) \, \Delta P(t+\tau) \rangle$, although we might have difficulties understanding the significance of this function. However, the corresponding power spectrum $P(\omega)$ is something well known, namely the frequency spectrum of sound.

Gaussian random fields. In general, knowing the power spectrum is not sufficient to unambiguously describe the statistical properties of any random field—in the same way as the correlation function $\xi(x)$ only provides an incomplete characterization. However, for certain classes of random fields, this is different. In particular, there are so-called *Gaussian random fields*, which are uniquely characterized by $P(k)$. Among the properties which characterize them, the probability distribution of the density fluctuations $\delta(\boldsymbol{x})$ at any point is a Gaussian. Such Gaussian random fields play an important role in cosmology because it is assumed that at very early epochs, the density field obeyed Gaussian statistics. This is a prediction of a large class of models of inflation which are supposed to generate the primordial density fluctuations in the Universe (see Sect. 7.9 below). Observational evidence for the Gaussian nature of the early density fluctuations comes from the observation of the anisotropy of the cosmic microwave background (see Sect. 8.6) which very strongly constrain any possible deviation from a Gaussian random field in the early Universe.

[5] This may not look like a 'standard' Fourier transform on first sight. However, the relation between $P(k)$ and $\xi(r)$ is given by a three-dimensional Fourier transform. Since the correlation function depends only on the separation $r = |\boldsymbol{r}|$, the two integrals over the angular coordinates can be performed explicitly, leading to the form of (7.29).

7.4 Evolution of density fluctuations

$P(k)$ and $\xi(x)$ both depend on cosmological time or redshift because the density field in the Universe evolves over time. Therefore, the dependence on t is explicitly written as $P(k,t)$ and $\xi(r,t)$. Note that $P(k,t)$ is linearly related to $\xi(x,t)$, according to (7.29), and ξ in turn depends quadratically on the density contrast δ. If $\boldsymbol{x}$ is the *comoving* separation vector, we then know from (7.16) the time dependence of the density fluctuations, $\delta(\boldsymbol{x},t) = D_+(t)\delta_0(\boldsymbol{x})$. Thus, within the scope of the validity of (7.16),

$$\xi(x,t) = D_+^2(t) \, \xi(x,t_0) \; , \tag{7.30}$$

and accordingly

$$P(k,t) = D_+^2(t) \, P(k,t_0) =: D_+^2(t) \, P_0(k) \; , \tag{7.31}$$

where k is a *comoving wave number*. We shall stress once again that these relations are valid only in the framework of Newtonian, linear perturbation theory in the matter dominated era of the Universe, to which we had restricted ourselves in Sect. 7.2.2. Equation (7.31) states that the knowledge of $P_0(k)$ is sufficient to obtain the power spectrum $P(k,t)$ at any time, again within the framework of linear perturbation theory.

7.4.1 The initial power spectrum

The Harrison–Zeldovich spectrum. Initially it may seem as if $P_0(k)$ is a function that can be chosen arbitrarily, but one objective of cosmology is to calculate this power spectrum and to compare it to observations. More than 30 years ago,

arguments were already developed to specify the functional form of the initial power spectrum.

At early times, the expansion of the Universe follows a power law, $a(t) \propto t^{1/2}$ in the radiation-dominated era. At that time, no natural length-scale existed in the Universe to which one might compare a wavelength. The only mathematical function that depends on a length but does not contain any characteristic scale is a power law[6]; hence for very early times one should expect

$$P(k) \propto k^{n_\mathrm{s}} \,. \tag{7.32}$$

Many years ago, Harrison, Zeldovich, Peebles and others argued that $n_\mathrm{s} = 1$, as for this slope, the amplitude of the fluctuations of the gravitational potential are constant, i.e., preferring neither small nor large scales. For this reason, the spectrum (7.32) with $n_\mathrm{s} = 1$ is called a scale-invariant spectrum, or *Harrison–Zeldovich spectrum*. With such a spectrum, we may choose a time t_i after the inflationary epoch and write

$$P(k, t_\mathrm{i}) = D_+^2(t_i)\, A\, k^{n_\mathrm{s}} \,, \tag{7.33}$$

where A is a normalization constant that cannot be determined from theory but has to be fixed by observations. As we will see in the following subsection, this is not the complete story: The result (7.33) needs to be modified to account for the different growth of the amplitude of density fluctuations in the radiation-dominated epoch of the Universe, compared to that in the later cosmic epochs from which our result (7.31) was derived.

Cold dark matter & hot dark matter. Furthermore, these modifications depend on the nature of the dark matter. One distinguishes between *cold dark matter (CDM)* and *hot dark matter (HDM)*. These two kinds of dark matter differ in the characteristic velocities of their constituents. Cold dark matter has a velocity dispersion that is negligible compared to astrophysically relevant velocities, e.g., the virial velocities of low-mass dark matter halos. Therefore, their initial velocity dispersion can well be approximated by zero, and all dark matter particles have the bulk velocity $\boldsymbol{u}$ of the cosmic 'fluid' (before the occurrence of multiple streams). In contrast, the velocity dispersion of hot dark matter is appreciable; as mentioned in Sect. 4.4.6 before, neutrinos are the best candidates for hot dark matter, in view of their known abundance, determined from the thermal history of the Universe (see Sect. 4.4), and their finite rest mass. The characteristic velocity of neutrinos is fully specified by their rest mass; despite their low temperature of $T_\nu \sim 1.9\,\mathrm{K}$ today, their thermal velocities of

$$v_\nu \sim 150\,(1+z)\left(\frac{m_\nu}{1\,\mathrm{eV}}\right)^{-1}\,\mathrm{km/s} \tag{7.34}$$

prevent them from forming matter concentrations at all mass scales except for the most massive ones, as their velocity is larger than the corresponding escape velocities. In other words, the finite velocity dispersion of hot dark matter is equivalent to assigning to it a pressure, which prevents them to fall into shallow gravitational potential wells. We will see below the dramatic differences between these two kinds of dark matter for the formation of structures in the Universe. In particular, this estimate shows that neutrinos cannot account for the dark matter on galaxy scales, and thus cannot explain the flat rotation curves of spiral galaxies.

If density fluctuations become too large on a certain scale, linear perturbation theory breaks down and (7.31) is no longer valid. Then the true current-day power spectrum $P(k, t_0)$ will deviate from $P_0(k)$. Nevertheless, in this case it is still useful to examine $P_0(k)$—it is then called the *linearly extrapolated power spectrum*.

7.4.2 Growth of density perturbations and the transfer function

Within the framework of linear Newtonian perturbation theory in the 'cosmic fluid', $\delta(\boldsymbol{x}, t) = D_+(t)\,\delta_0(\boldsymbol{x})$ applies. Modifications to this behavior are necessary for several reasons:

- If dark matter consists (partly) of hot dark matter, this may not be gravitationally bound to the potential well of a density concentration. In this case, the particles are able to move freely and to escape from the potential well, which in the end leads to its dissolution if these particles dominate the matter overdensity. From this argument, it follows immediately that for HDM small-scale density perturbations cannot form. For CDM this effect of *free streaming* does not occur. In our discussion of the evolution of density perturbations, we have not included the possible presence of hot dark matter, since we neglected any pressure term in the hydrodynamic equations.
- At redshifts $z \gtrsim z_\mathrm{eq}$, radiation dominates the density of the Universe. Since the expansion law $a(t)$ is then distinctly different from that in the matter-dominated phase, the growth rate for density fluctuations will also change.
- As discussed in Sect. 4.5.2, a cosmic horizon exists with comoving scale $r_\mathrm{H,com}(t)$. Physical interactions can take

[6] You can convince yourself of this by trying to find another type of function of a scale that does not involve a characteristic length; e.g., $\sin x$ does not work if x is a length, since the sine of a length is not defined; one thus needs something like $\sin(x/x_0)$, hence introducing a length-scale. The same arguments apply to other functions, such as the logarithm, the exponential etc. Also note that the sum of two power laws, e.g., $Ax^\alpha + Bx^\beta$ defines a characteristic scale, namely that value of x where the two terms become equal.

place only on scales smaller than $r_{\rm H,com}(t)$. For fluctuations of length-scales $L \sim 2\pi/k \gtrsim r_{\rm H,com}(t)$, Newtonian perturbation theory will cease to be valid, and one needs to apply linear perturbation theory in the framework of the General Relativity.

The transfer function. These effects together will lead to a modification of the shape of the power spectrum, relative to the relation (7.33); for example, the evolution of perturbations in the radiation-dominated cosmos proceeds differently from that in the matter-dominated era. The power spectrum $P(k)$ is affected by the combination of the above effects, and will be different from the primordial spectral shape, $P \propto k^{n_{\rm s}}$. The modification of the power spectrum is described in terms of the *transfer function* $T(k)$, in the form

$$P(k,t) = D_+^2(t)\, Ak^{n_{\rm s}}\, T^2(k)\ . \tag{7.35}$$

The transfer function can be computed for any cosmological model if the matter content of the universe is specified. In particular, $T(k)$ depends on the nature of dark matter.

CDM and HDM. The first of the above points immediately implies that a clear difference must exist between HDM and CDM models regarding structure formation and evolution. In HDM models, small-scale fluctuations are washed out by free-streaming of relativistic particles, i.e., the power spectrum is completely suppressed for large k, which is expressed by the transfer function $T(k)$ decreasing exponentially for large k. In the context of such a theory, very large structures will form first, and galaxies can form only later by fragmentation of large structures. However, this formation scenario is in clear contradiction with observations. For example, we observe galaxies and QSOs at $z > 6$ so that small-scale structure is already present at times when the Universe had less than 10 % of its current age. In addition, the observed correlation function of galaxies, both in the local Universe (see Fig. 7.4) and at higher redshift, is incompatible with cosmological models in which the dark matter is composed mainly of HDM.

Hot dark matter leads to structure formation that does not agree with observation. Therefore we can exclude HDM as the dominant constituent of dark matter. For this reason, it is now commonly assumed that the dark matter is 'cold'. The achievements of the CDM scenario in the comparison between model predictions and observations fully justify this assumption.

We shall elaborate on the last statement in quite some detail in Chap. 8. Anticipating these results, for most of the rest of this chapter we will neglect the possible presence of a pressure component of the dark matter, i.e., we will concentrate of cold dark matter.

Relevance of horizon size for structure growth. In linear perturbation theory, fluctuations grow at the same rate on all scales, or for all wave numbers, independent of each other. This applies not only in the Newtonian case, but also remains valid in the framework of General Relativity as long as the fluctuation amplitudes are small. Therefore, the behavior on any (comoving) length-scale can be investigated independently of the other scales. At very early times, perturbations with a comoving scale L are larger than the (comoving) horizon, and only for $z < z_{\rm enter}(L)$ does the horizon become larger than the considered scale L. Here, $z_{\rm enter}(L)$ is defined as the redshift at which the (comoving) horizon equals the (comoving) length-scale L,

$$r_{\rm H,com}(z_{\rm enter}(L)) = L\ . \tag{7.36}$$

It is common to say that at $z_{\rm enter}(L)$ the perturbation under consideration 'enters the horizon', whereas actually the process is the opposite—the horizon outgrows the perturbation. Relativistic perturbation theory shows that density fluctuations of scale L grow as long as $L > r_{\rm H,com}$, namely $\propto a^2$ if radiation dominates (thus, for $z > z_{\rm eq}$), or $\propto a$ if matter dominates (i.e., for $z < z_{\rm eq}$). Free-streaming particles or pressure gradients cannot impede the growth on scales larger than the horizon length because, according to the definition of the horizon, physical interactions—which pressure or free-streaming particles would be—cannot extend to scales larger than the horizon size.

Qualitative behavior of the transfer function. The behavior of the growth of a density perturbation on a scale L for $z < z_{\rm enter}(L)$ depends on $z_{\rm enter}$ itself. If a perturbation enters the horizon in the radiation-dominated phase, $z_{\rm eq} \lesssim z_{\rm enter}(L)$, it ceases to grow during the epoch $z_{\rm eq} \lesssim z \lesssim z_{\rm enter}(L)$. In this period, the energy density in the Universe is dominated by radiation, and the resulting expansion rate prevents an efficient perturbation growth. At later epochs, when $z \lesssim z_{\rm eq}$, the growth of density perturbations continues. If $z_{\rm enter}(L) \lesssim z_{\rm eq}$, thus if the perturbation enters the horizon during the matter-dominated epoch of the Universe, these perturbations will grow as described in Sect. 7.2.2, with $\delta \propto D_+(t)$. This implies that a length-scale L_0 is singled out, namely the one for which

$$z_{\rm eq} = z_{\rm enter}(L_0)\ , \tag{7.37}$$

so that L_0 is the comoving horizon size at matter-radiation equality. We can calculate this length-scale explicitly from (4.74),

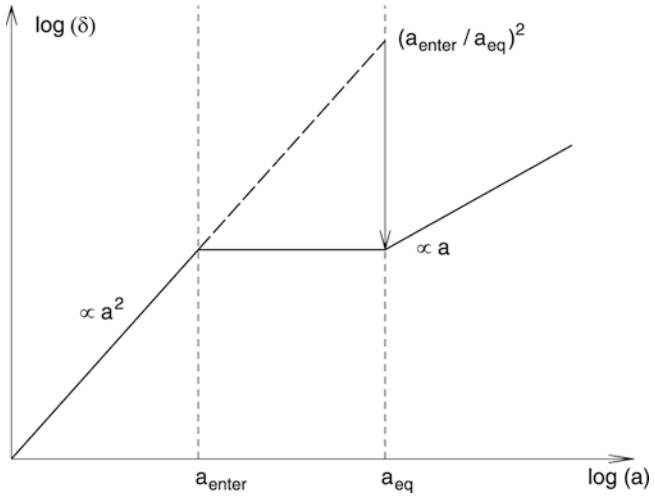


Fig. 7.5 A density perturbation that enters the horizon during the radiation-dominated epoch of the Universe ceases to grow until matter starts to dominate the energy content of the Universe. In comparison to a perturbation that enters the horizon later, during the matter-dominated epoch, the amplitude of the smaller perturbation is suppressed by a factor $(a_{\rm enter}/a_{\rm eq})^2$, which explains the qualitative behavior (7.40) of the transfer function. Adapted from: M. Bartelmann & P. Schneider 2001, *Weak Gravitational Lensing*, Phys. Rep. 340, 291

$$L_0 = r_{\rm H,com}(z_{\rm eq}) = \int_0^{a_{\rm eq}} \frac{c\,{\rm d}a}{a^2\,H(a)} = \frac{c}{H_0\sqrt{\Omega_{\rm m}}} \int_0^{a_{\rm eq}} \frac{{\rm d}a}{\sqrt{a+a_{\rm eq}}}$$
$$= \left(\sqrt{2}-1\right)\frac{2c}{H_0}\left(\frac{a_{\rm eq}}{\Omega_{\rm m}}\right)^{1/2}\,, \qquad (7.38)$$

where we made use of the Hubble function at early times where curvature and vacuum energy density are negligible, and the relation (4.30) for $a_{\rm eq}$ was used to write $H^2 = H_0^2\Omega_{\rm m}(a+a_{\rm eq})a^{-4}$. Again using (4.30), we finally obtain

$$\boxed{L_0 \approx 16(\Omega_{\rm m}h^2)^{-1}\ {\rm Mpc}}\,. \qquad (7.39)$$

Density fluctuations with $L > L_0$ enter the horizon after matter started to dominate the energy density of the Universe; hence their growth is not impeded by a phase of radiation-dominance. In contrast, density fluctuations with $L < L_0$ enter the horizon at a time when radiation was still dominating. These then cannot grow further as long as $z > z_{\rm eq}$, and only in the matter-dominated epoch will their amplitudes proceed to grow again. Their relative amplitude up to the present time has therefore grown by a smaller factor than that of fluctuations with $L > L_0$ (see Fig. 7.5): since fluctuations larger than the horizon grow $\propto a^2$ in the radiation-dominated era, a perturbation of scale $L < L_0$ has its amplitude suppressed by a factor $\left[a_{\rm enter}(L)/a_{\rm eq}\right]^2$ until matter-radiation equality, compared to a perturbation with scale $> L_0$. Since $r_{\rm H,com} \propto a$ in the radiation-dominated era [see (4.76)], these small-scale perturbations are suppressed by a factor $(L/L_0)^2$ relative to large-scale perturbations.

The quantitative consideration of these effects allows us to compute the transfer function. In general, this needs to be done numerically, but very good approximations exist. In particular, since all the physics involved at these early epochs concern small perturbations, and thus can safely be treated in a linear approximation, these numerical calculations pose no principal problems. Several codes are publicly available for calculating the transfer function for a specified set of cosmological parameters. For the limiting cases of $L \gg L_0$ and $L \ll L_0$, our previous discussion yields

$$T(k) \approx 1 \quad \text{for} \quad k \ll 1/L_0\,,$$
$$T(k) \approx (kL_0)^{-2} \quad \text{for} \quad k \gg 1/L_0\,. \qquad (7.40)$$

However, the important point is:

> In the framework of the CDM model, the transfer function can be computed, and thus, by means of (7.35), also the power spectrum of the density fluctuations as a function of length-scale and redshift. The amplitude of the power spectrum has to be obtained from observations.

The shape parameter. The transfer function depends on the combination kL_0, which is the inverse of the ratio of the length-scale under consideration ($\sim 2\pi/k$) and the horizon scale at the epoch of equality, and thus on $k(\Omega_{\rm m}h^2)^{-1}$. Since distances determined from redshift are measured in units of h^{-1} Mpc, the wave number is measured in units of $h\,{\rm Mpc}^{-1}$. Therefore, the shape of the transfer function, and thus also that of the power spectrum, depends on $\Gamma = \Omega_{\rm m}h$. Γ is called the *shape parameter* of the power spectrum. It is sometimes used as a free parameter instead of being identified with $\Omega_{\rm m}h$. A more detailed analysis shows that Γ depends also on $\Omega_{\rm b}$, but since $\Omega_{\rm b} \lesssim 0.05$ is small, according to primordial nucleosynthesis (see Sect. 4.4.5), this effect is relatively small—however, this small effect turns out to be of great importance for observational cosmology (see Sect. 7.4.3 below).

If the galaxy distribution follows the distribution of dark matter, the former can be used to determine the correlation function or the power spectrum, in particular the shape parameter Γ. Both from the distribution of galaxies projected onto the sphere (angular correlation function) and from the three-dimensional galaxy distribution (which is determined from redshift surveys), one finds that $\Gamma \sim 0.2$ (see Sect. 8.1.3). From $T(k) \approx 1$ for $kL_0 \ll 1$, and with (7.33), we find that $P(k) \propto k^{n_{\rm s}}$ for $kL_0 \ll 1$,

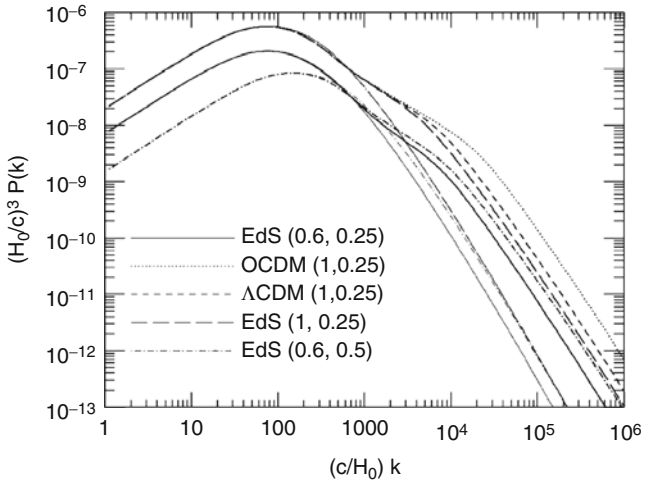


Fig. 7.6 The current power spectrum of density fluctuations for CDM models. The wave number k is given in units of H_0/c, and $(H_0/c)^3\,P(k)$ is dimensionless. The various *curves* have different cosmological parameters: EdS: $\Omega_\mathrm{m} = 1$, $\Omega_\Lambda = 0$; OCDM: $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0$; ΛCDM: $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0.7$. The values in *parentheses* specify (σ_8, Γ), where σ_8 is the normalization of the power spectrum (which will be discussed below), and where Γ is the shape parameter. The *thin curves* correspond to the power spectrum $P_0(k)$ linearly extrapolated to the present day, and the *bold curves* take the non-linear evolution into account. For this figure, the effects of baryons on the transfer function have been neglected

with $n_\mathrm{s} \approx 1$. This behavior is compatible with the CMB anisotropy measurements on large scales, as we will discuss in detail in Sect. 8.6.

In Fig. 7.6, the power spectrum is plotted for several cosmological models that have different density parameter, shape parameter, and normalization of the power spectrum. The thin curves show $P(k)$ as derived from linear perturbation theory, and the bold curves display the power spectrum with non-linear structure evolution taken approximately into account. The power spectra displayed all have a characteristic wave number at which the slope of $P(k)$ changes, or where the peak of $P(k)$ is located. It is specified by $k \sim 2\pi/L_0 \approx (\pi\Gamma/8)h\,\mathrm{Mpc}^{-1}$, with the characteristic length L_0 being defined in (7.39). The value of the shape parameter Γ determines the location of this peak.

Transfer function for other dark matter models. As mentioned before, the evolution of density fluctuations depends on the nature of dark matter. The asymptotic relation (7.40) is valid only for cold dark matter. From our previous discussion, we infer that the free streaming of hot dark matter would erase fluctuations on nearly all scales; in particular, if the hot dark matter particles are relativistic at the epoch of matter-radiation equality, then all fluctuations on scales smaller than L_0 would be destroyed (since the length scale over which a relativistic dark matter particle can freely stream equals the horizon scale). Accordingly, the transfer function for a hot dark matter universe will have an exponential decline for $kL_0 \ll 1$.

Besides pure CDM and HDM models (the latter being excluded by observation), one can consider models which are dominated by CDM, but which have a (small) contribution by HDM; these are called mixed dark matter (MDM) models. Such a contribution has indeed now become part of the standard model, due to the detected finite rest mass of neutrinos which implies $0 < \Omega_\nu \ll 1$. With this contribution, $T(k)$ is changed in such a way that small scales (i.e., large k) are slightly damped in the power spectrum. We will see later that by observing the power spectrum we can constrain the rest mass of neutrinos very well, and cosmological observations provide, in fact, by far the most stringent mass limits for neutrinos.

More exotic dark matter models suggest the existence of particles which are intermediate between cold and hot dark matter, in that their impact on the transfer function is a damping of small-scale fluctuations. The free-streaming of the warm dark matter particles has a similar effect as that of the neutrinos, except that their mass is higher, hence the velocity is lower, and thus the cut-off scale in the power spectrum is shifted to smaller length-scales. These warm dark matter models were introduced to reduce the apparent discrepancy between the abundance of satellite galaxies and the number of dark matter subclumps predicted by cold dark matter models (see Sect. 7.8). Despite lacking a natural candidate for the constituent of warm dark matter from particle physics, its properties can be strongly constrained with recent cosmological observations; we will come back to this issue further below.

The linear theory of the evolution of density fluctuations will break down at the latest when $|\delta| \sim 1$; the above equations for the power spectrum $P(k,t)$ are therefore valid only if the respective fluctuations are small. However, accurate fitting formulae now exist for $P(k,t)$ which are also valid in the non-linear regime. For some cosmological models, the non-linear power spectrum is displayed in Fig. 7.6.

7.4.3 The baryonic density fluctuations

The evolution of density fluctuations of baryons differs from that of dark matter. The reason for this is essentially the interaction of baryons with photons: although matter dominates the Universe for $z < z_\mathrm{eq}$, the energy density of baryons remains smaller than that of the photons for a longer time, until after recombination begins, as can be seen as follows: The baryon-to-photon density ratio is

$$\frac{\rho_\mathrm{b}}{\rho_\gamma} = \frac{\Omega_\mathrm{b} a^{-3}}{\Omega_\gamma a^{-4}} = a\frac{\Omega_\mathrm{b}}{\Omega_\mathrm{m}}\frac{\Omega_\mathrm{m}}{\Omega_\mathrm{r}}\frac{\Omega_\mathrm{r}}{\Omega_\gamma} = 1.68\frac{a}{a_\mathrm{eq}}\frac{\Omega_\mathrm{b}}{\Omega_\mathrm{m}} \sim 0.28\frac{a}{a_\mathrm{eq}}\,, \tag{7.41}$$

where we used the expression (4.30) for $a_{\rm eq}$, and (4.28) for the radiation-to-photon density—the neutrinos, which contribute to the radiation density, are not coupled to the baryons. In the final step, we used the estimate $\Omega_{\rm b} \sim \Omega_{\rm m}/6$ for our Universe. Hence, if radiation-matter equality happens at $z \sim 3000$, then the photon density is larger than that of the baryons for $z \gtrsim 800$.

Since photons and baryons interact with each other by photon scattering on free electrons, which again are tightly coupled electromagnetically to protons and helium nuclei, baryons and photons are strongly coupled before recombination, and form a single fluid. Due to the presence of photons, this fluid has a strong pressure, which prevents it from falling into potential wells formed by the dark matter. Thus, the pressure prevents strong inhomogeneities of the baryon-photon fluid.

Sound waves. To discuss the evolution of baryon perturbations in a bit more detail, we consider again a perturbation of comoving scale L. As long as the perturbation is larger than the horizon size, pressure effects can not affect the behavior of the fluid, and thus baryons and photons behave in the same way as the dark matter—the amplitude of their perturbations grow. As soon as the perturbation enters the horizon, the situation changes. Although the baryons are gravitationally pulled into the density maxima of the dark matter, pressure provides a restoring force which acts against a compression of the baryon-photon fluid. As a results, this fluid will develop sound waves.

Sound horizon. The maximum distance sound waves can travel up to a given epoch is called the *sound horizon*. Loosely speaking, it is given by the product of the sound speed and the cosmic time and has a very similar meaning as the (event) horizon that we discussed before. The sound speed in this photon-dominated fluid is given by $c_{\rm s} \approx c/\sqrt{3}$, as will be shown shortly. Thus, the sound horizon is about a factor of $\sqrt{3}$ smaller than the event horizon. As soon as a perturbation enters the sound horizon, the amplitude of the baryon-photon fluctuations can not grow anymore; instead, they undergo damped oscillations.

The adiabatic sound velocity $c_{\rm s}$ of a fluid is given in general by

$$c_{\rm s}^2 = \frac{\partial P}{\partial \rho} \; .$$

The pressure of the fluid is generated by the photons, $P = c^2\rho_\gamma/3 = c^2\rho_{\rm cr}\Omega_\gamma a^{-4}/3$, and the density is the sum of that of baryons and photons, $\rho = (\Omega_{\rm b}a^{-3} + \Omega_\gamma a^{-4})\rho_{\rm cr}$. Thus, the sound velocity is

$$c_{\rm s} = \sqrt{\frac{{\rm d}P/{\rm d}a}{{\rm d}\rho/{\rm d}a}} = \frac{c}{\sqrt{3}}\sqrt{\frac{4\Omega_\gamma a^{-5}}{3\Omega_{\rm b}a^{-4} + 4\Omega_\gamma a^{-5}}} = \frac{c}{\sqrt{3(1+\mathcal{R})}}, \tag{7.42}$$

where we have defined

$$\mathcal{R} = \frac{3}{4}\frac{\rho_{\rm b}}{\rho_\gamma} = \frac{3}{4}\frac{\Omega_{\rm b}}{\Omega_\gamma}\,a \; . \tag{7.43}$$

Note that $\mathcal{R}$ is smaller than unity until recombination, and thus $c_{\rm s} \approx c/\sqrt{3}$ provides a reasonable first approximation.

At recombination, the free electrons recombined with the hydrogen and helium nuclei, after which there are essentially no more free electrons which couple to the photon field. Hence, after recombination the baryon fluid lacks the pressure support of the photons, and the sound speed drops to zero—the sound waves do no longer propagate, but get frozen in. Now the baryons are free to react to the gravitational field created by the dark matter inhomogeneities, and they can fall into their potential wells. After some time, the spatial distribution of the baryons is essentially the same as that of the dark matter.

Hence, there is a maximum wavelength of the sound waves, namely the (comoving) sound horizon at recombination, $r_{\rm s}$, which can be calculated according to (4.74),

$$r_{\rm s} = \int_0^{a_{\rm rec}} \frac{c\,{\rm d}a}{\sqrt{3(1+\mathcal{R})}a^2 H(a)} \; , \tag{7.44}$$

except that we exchanged the speed of light by the speed of sound. Using the Hubble function for the early universe, where only matter and radiation are relevant, we find

$$\begin{aligned} r_{\rm s} &= \frac{c}{H_0\sqrt{\Omega_{\rm m}}\sqrt{3(1+\mathcal{R}_{\rm eff})}} \int_0^{a_{\rm rec}} \frac{{\rm d}a'}{\sqrt{a'+a_{\rm eq}}} \\ &= \frac{2c}{H_0\sqrt{\Omega_{\rm m}}\sqrt{3(1+\mathcal{R}_{\rm eff})}} \left(\sqrt{a_{\rm rec}+a_{\rm eq}} - \sqrt{a_{\rm eq}}\right) \\ &\sim \frac{120h^{-1}{\rm Mpc}}{\sqrt{\Omega_{\rm m}}\sqrt{1+\mathcal{R}_{\rm eff}}} \; , \end{aligned} \tag{7.45}$$

where $\mathcal{R}_{\rm eff}$ is a mean of $\mathcal{R}$ over the integration range, and the final expression is a rough estimate, obtained by assuming $a_{\rm eq} \ll a_{\rm rec} \sim 10^{-3}$.

Figure 7.7 illustrates the physical significance of this length scale, showing the time evolution of an initial density peak of all four components in the Universe. The length scale $r_{\rm s}$ is the distance the baryon-photon fluid propagates outwards from the initial density peak before baryons and photons decouple, after which the density perturbation of

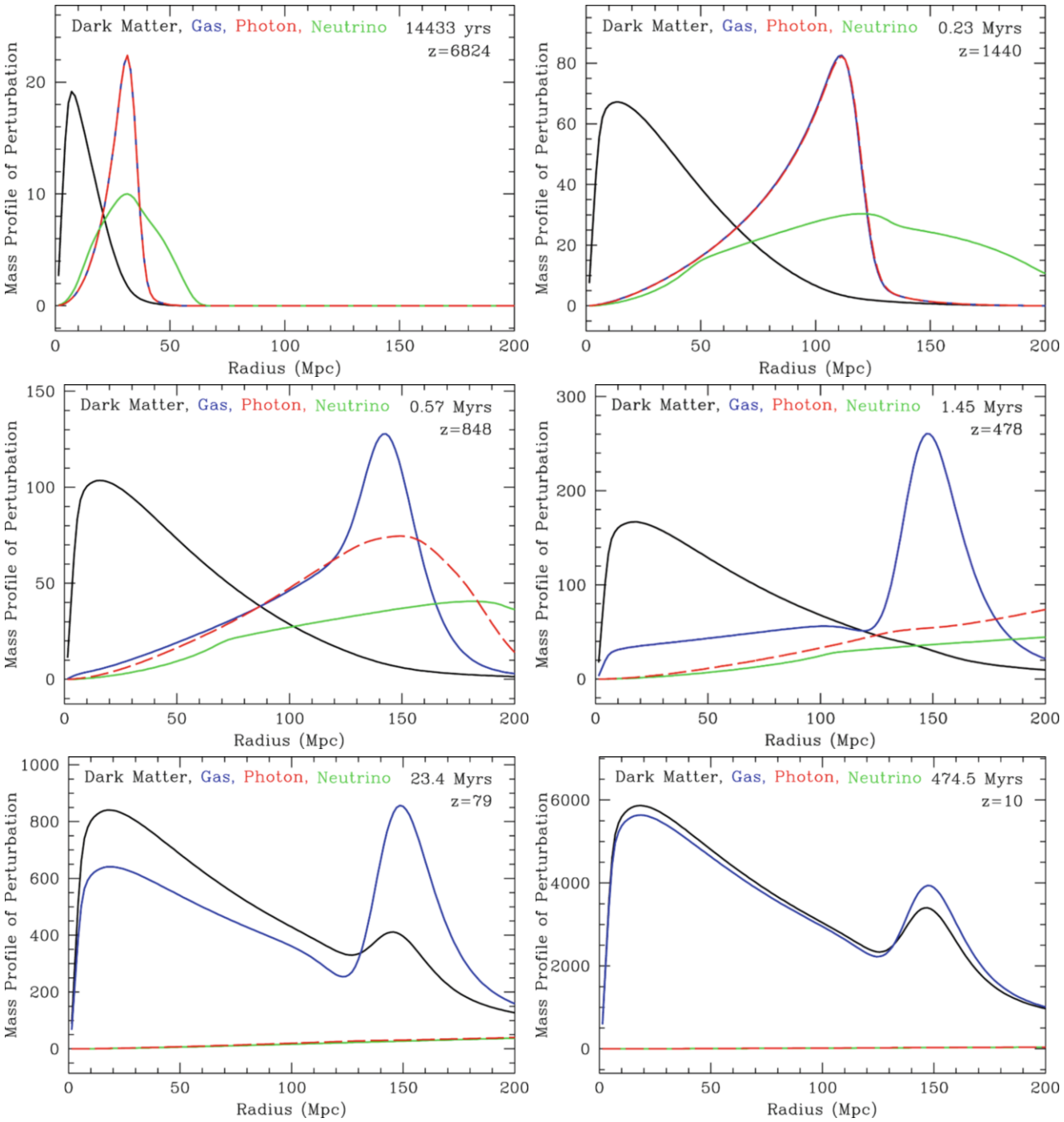


Fig. 7.7 Evolution in time of an initial density peak in all components of the cosmic matter. The x-axis shows the comoving radial coordinate, the y-axis displays the density, multiplied by (radius)^2. The different snapshots show the spatial distribution of the various species at later epochs. Neutrinos freely stream out of the perturbation at the speed of light. The photon and baryons are strongly coupled before recombination, and thus have the same spatial distribution. They move out from the initial density peak with the sound speed, $c_s \approx c/\sqrt{3}$. After recombination, the photons are no longer coupled to the baryons and freely stream away; correspondingly, the sound speed of the baryons drops to zero, and they stop propagating outwards. After that, the baryons are gravitationally attracted by the density peak of the dark matter and fall in; however, some of the matter also falls into the density peaks (in the example of this figure, it is an overdense spherical shell) created by baryons, whereas the density profile of neutrinos and photons becomes flat. At late time, the distributions of baryons and dark matter become identical (before the onset of non-linear processes such as halo formation). The central density peak, and the secondary peak have a well-defined separation, given by the distance a sound wave could travel before the baryons decoupled from the photons. Source: D.J. Eisenstein et al. 2007, *On the Robustness of the Acoustic Scale in the Low-Redshift Clustering of Matter*, ApJ 664, 660, p. 662, Fig. 1,

baryons gets frozen. Hence, this phenomenon is indeed characterized by the length-scale r_s.[7]

Baryonic acoustic oscillations. The sound waves in the baryon-photon fluid, the baryonic acoustic oscillations (BAOs), are observable today. Since at recombination, the photons interacted with matter for the last time, the cosmic microwave background radiation provides us with a picture of the density fluctuations at the epoch of recombination. Our observable cosmic microwave sky essentially is a picture of a two-dimensional cut at fixed time (the time of last scattering) through the density field of the baryons. A cut through an ensemble of sound waves shows an instantaneous picture of these waves. Hence, they are expected to be visible in the temperature distribution of the CMB. As we will see in Sect. 8.6, this is indeed the case: These BAOs imprint one of the most characteristic features on the CMB anisotropies. Since the sound waves are damped once they are inside the sound horizon, the largest amplitude waves are those whose wavelength equals the sound horizon at recombination.

BAOs in the low-redshift Universe. We have argued that the baryons, once they are no longer coupled to radiation and thus become pressureless, fall into the potential wells of the dark matter. This happens because the dark matter fluctuations can grow while the baryonic fluctuations could not due to the photon pressure, and because the mean density of dark matter is substantially larger than that of the baryons. This is almost the full story, but not entirely: Baryons make about 15 % of the total matter density, and are therefore not negligible. After recombination, the BAOs are frozen, like standing waves, and thus the total matter fluctuations are a superposition of the dark matter inhomogeneities and these standing waves. Whereas the dark matter dominates the density fluctuations, a small fraction of the matter also follows the inhomogeneities created by the standing waves. Since these waves have a characteristic length scale—the sound horizon at recombination—this characteristic length scale should be visible in the properties of the matter distribution even today. As we shall see in Sect. 8.1, the correlation function of galaxies contains a characteristic feature at the length scale r_s. Hence, relics of the sound waves in the pre-recombination era are even visible in the current Universe. The effects of the BAOs are included in the transfer function $T(k)$, which thus shows some low-amplitude oscillations, often called 'wiggles' (they not seen in Fig. 7.6, since there the transfer function of a dark matter-only models was used).

[7] One may wonder why the neutrinos in Fig. 7.7 have a broad distribution in space, and not simply a sharp shell—they all stream out with velocity c. The reason is that the initial conditions were chosen such that they correspond to a growing mode. For those, we argued that any density perturbation is associated with a velocity perturbation. As can be seen from (7.25), the relation between density and peculiar velocity is non-local, i.e., the velocity field associated with a density peak at the center is non-zero at all radii. This velocity field also causes the dark matter distribution to expand—once the neutrinos and the photon-baryon fluid have moved out, the gravitational field inside the dark matter peak is weaker than it would have been predicted from a pure dark matter distribution, yielding an expanding peculiar velocity field near the center.

7.5 Non-linear structure evolution

Linear perturbation theory has a limited range of applicability; in particular, the evolution of structures like clusters of galaxies cannot be treated within the framework of linear perturbation theory. One might imagine that one can evolve the system of (7.8), (7.10) and (7.11) to higher order in the small variables δ and $|\boldsymbol{u}|$, and thus consider non-linear perturbation theory. In fact, a quite extensive literature exists on this topic in which such calculations are performed. However, while this higher-order perturbation theory indeed allows us to follow density fluctuations to somewhat larger values of $|\delta|$, the mathematical effort required is substantial. In addition, the fluid approximation is no longer valid if gravitationally bound systems form because, as mentioned earlier, multiple steams of matter will occur in this case.

However, for some interesting limiting cases, analytical descriptions exist which are able to represent the characteristics of the non-linear evolution of the mass distribution in the Universe. We shall now investigate a special and very important case of such a non-linear model. In general, studying the non-linear structure evolution requires the use of numerical methods. Therefore, we will also discuss some aspects of such numerical simulations.

7.5.1 Model of spherical collapse

Assumptions. We consider a spherical region in an expanding universe, with its density $\rho(t)$ enhanced compared to the mean cosmic density $\bar{\rho}(t)$,

$$\rho(t) = [1 + \delta(t)]\,\bar{\rho}(t)\ , \tag{7.46}$$

where we use the density contrast δ as defined in (7.1). For reasons of simplicity we assume that the density within the sphere is homogeneous although, as we will later see, this is not really a restriction. The density perturbation is assumed to be small for small t, so that it will grow linearly at first, $\delta(t) \propto D_+(t)$, as long as $\delta \ll 1$. If we consider a time t_i which is sufficiently early such that $\delta(t_i) \ll 1$, then according to the definition of the growth factor D_+, $\delta(t_i) = \delta_0\, D_+(t_i)$, where δ_0 is the density contrast linearly extrapolated to the present day. It should be mentioned once again that $\delta_0 \neq$

$\delta(t_0)$, because the latter is in general affected by the non-linear evolution.

Let R be the initial *comoving* radius of the overdense sphere; as long as $\delta \ll 1$, the *comoving* radius will change only marginally. The mass within this sphere is

$$M = \frac{4\pi}{3} R^3 \rho_0 (1 + \delta_i) \approx \frac{4\pi}{3} R^3 \rho_0 , \qquad (7.47)$$

because the physical (or proper) radius is $R_{\rm phys} = aR$, and $\bar{\rho} = \rho_0/a^3$. This means that a unique relation exists between the initial *comoving* radius and the mass of this sphere, independent of the choice of t_i and δ_0, if only we choose $\delta(t_i) = \delta_0 D_+(t_i) \ll 1$.

Evolution. Due to the enhanced gravitational force, the sphere will expand slightly more slowly than the universe as a whole, which will lead to an increase in its density contrast. This then decelerates the expansion rate even further, relative to the cosmic expansion rate. Indeed, the equations of motion for the radius of the sphere are identical to the Friedmann equations for the cosmic expansion, only with the sphere having an effective Ω_m different from that of the mean universe. If the initial density is sufficiently large, the expansion of the sphere will come to a halt, i.e., its proper radius $R_{\rm phys}(t)$ will reach a maximum; after this, the sphere will recollapse.

If $t_{\rm max}$ is the time of maximum expansion, then the sphere will, theoretically, collapse to a single point at time $t_{\rm coll} = 2t_{\rm max}$. The relation $t_{\rm coll} = 2t_{\rm max}$ follows from the time reversal symmetry of the equation of motion: the time to the maximum expansion is equal to the time from that point back to complete collapse.[8] The question of whether the expansion of the sphere will come to a halt depends on the density contrast $\delta(t_i)$ or δ_0—compare the discussion of the expansion of the Universe in Sect. 4.3.1—and on the model for the cosmic background.

Special case: Einstein–de Sitter model. In the special case of $\Omega_m = 1$ and $\Omega_\Lambda = 0$, this behavior can easily be quantified analytically; we thus treat this case separately. In this cosmological model, any sphere with $\delta_0 > 0$ is a "closed universe" and will therefore recollapse at some time. For the collapse to take place before t_1, $\delta(t_i)$ or δ_0 needs to exceed a threshold value. For instance, for a collapse at $t_{\rm coll} \le t_0$, a linearly extrapolated overdensity of

$$\delta_0 \ge \delta_c = \frac{3}{20} (12\pi)^{2/3} \simeq 1.69 \qquad (7.48)$$

is required. More generally, one finds that $\delta_0 \ge \delta_c (1 + z)$ is needed for the collapse to occur before redshift z.

One can calculate δ_c also for other values of the density parameters. It turns out that the modifications are relatively small, and thus the value (7.48) is a useful approximation also for other cosmological models.

Violent relaxation and virial equilibrium. Of course, the sphere will not really collapse to a single point. This would only be the case if the sphere was perfectly homogeneous and if the particles in the sphere moved along perfectly radial orbits. In reality, small-scale density and gravitational fluctuations will exist within such a sphere. These then lead to deviations of the particles' tracks from perfectly radial orbits, an effect that becomes more important as the density contrast of the sphere increases. The particles will scatter on these fluctuations in the gravitational field and will virialize; this process of *violent relaxation* has already been described in Sect. 6.3.3 and occurs on short time-scales—roughly the dynamical time-scale, i.e., the time it takes the particles to fully cross the sphere. In this case, the virialization is essentially complete at $t_{\rm coll}$. After that, the sphere will be in virial equilibrium, and its average density will be[9]

$$\langle\rho\rangle = (1 + \delta_{\rm vir})\, \bar{\rho}(t_{\rm coll}), \quad \text{where} \quad (1 + \delta_{\rm vir}) \simeq 18\pi^2 \approx 178 , \qquad (7.49)$$

where the final expression is obtained for an EdS model. With the same assumption, corresponding expressions can be derived for other models as well. However, since the numerical factor for the overdensity of a virialized halo depends on the idealized assumptions made in the spherical collapse model, its exact value is of little importance. Nevertheless, the foregoing relation forms the basis for the statement that the virialized region, e.g., of a cluster, is a sphere with an average density $\sim$ 200 times the critical density $\rho_{\rm cr}$ of the Universe at the epoch of collapse. Another conclusion from this consideration is that a massive galaxy cluster with a virial radius of $1.5\,h^{-1}$Mpc must have formed from the collapse of a region that originally had a comoving radius larger by about an order of magnitude (see problem 7.3). Such a virialized mass concentration of dark matter is called a *dark matter halo*.

[8] This occurs for the same reason that it takes a stone thrown up into the air the same time to reach its peak altitude as to fall back to the ground from there.

[9] This result is obtained from conservation of energy and from the virial theorem. The total energy $E_{\rm tot}$ of the sphere is a constant. At the time of maximum expansion, it is given solely by the gravitational binding energy of the system since then the expansion velocity, and thus the kinetic energy, vanishes. On the other hand, the virial theorem implies that in virial equilibrium $E_{\rm kin} = -E_{\rm pot}/2$, and by combining this with the conservation of energy $E_{\rm tot} = E_{\rm kin} + E_{\rm pot}$ one is then able to compute $E_{\rm pot}$ in equilibrium and hence the radius and density of the collapsed sphere. For an EdS model, $r_{\rm vir} = r_{\rm max}/2$.

Up to now, we have considered the collapse of a homogeneous sphere. From the above arguments one can easily convince oneself that the model is still valid if the sphere has a radial density profile with a density that decreases outwards. In this case, the initial density contrast will also decrease as a function of radius. The inner regions of such a sphere will then collapse faster than the outer ones; a halo of lower mass will form first, and only later, when the outer regions have also collapsed, will a halo with higher mass form. From this it follows that halos of low initial mass will grow in mass by further accretion of matter.

The spherical collapse model is a simple model for the non-linear evolution of a density perturbation in the Universe. Despite being simplistic, it represents the fundamental principles of gravitational collapse and yields approximate relations, e.g., for the collapse time and mean density inside the virialized region, as they are found from numerical simulations.

7.5.2 Number density of dark matter halos

Press–Schechter model. The model of spherical collapse allows us to approximately compute the number density of dark matter halos as a function of their mass and redshift; this model is called the *Press–Schechter model*.

We consider a field of density fluctuations $\delta_0(\boldsymbol{x})$, featuring fluctuations on all scales according to the power spectrum $P_0(k)$. Assume that we smooth this field with a *comoving* smoothing length R, by convolving it with a filter function of this scale. In our example of the waves on a lake, we could examine a picture of its surface taken through a pane of milk-glass, by which all the contours on small scales would be blurred. Then, let $\delta_R(\boldsymbol{x})$ be the smoothed density field, linearly extrapolated to the present day. This field does not contain any fluctuations on scales $\lesssim R$, because these have been smoothed out. Each maximum in $\delta_R(\boldsymbol{x})$ corresponds to a peak with characteristic scale $\gtrsim R$ and, according to (7.47), each of these maxima corresponds to a mass peak of mass $M \sim (4\pi R^3/3)\rho_0$. If the amplitude δ_R of the density peak is sufficiently large, a sphere of (comoving) radius R around the peak will decouple from the linear growth of density fluctuations and will begin to grow non-linearly. Its expansion will come to a halt, and then it will recollapse. This process is similar to that in the spherical collapse model and can be described approximately by this model. The density contrast required for the collapse, $\delta_R \geq \delta_{\min}$, can be computed for any cosmological model and for any redshift.

If the statistical properties of $\delta_0(\boldsymbol{x})$ are Gaussian—which is expected for a variety of reasons—the statistical properties of the fluctuation field δ_0 are completely defined by the power spectrum $P(k)$. Then the abundance of density maxima with $\delta_R \geq \delta_{\min}$ can be computed, and hence the (comoving) number density $n(M, z)$ of relaxed dark matter halos in the Universe as a function of mass M and redshift z can be determined.

The mass spectrum. The most important results of the Press–Schechter model are easily explained (see Fig. 7.8). The number density of halos of mass M depends of course on the amplitude of the density fluctuation δ_0—i.e., on the normalization of the power spectrum $P_0(k)$. Hence, the normalization of $P_0(k)$ can be determined by comparing the prediction of the Press–Schechter model with the observed number density of galaxy clusters, as we will discuss further in Sect. 8.2.1 below. The corresponding result is called the "cluster-normalized power spectrum".

Furthermore, we find that $n(M, z)$ is a decreasing function of halo mass M. This follows immediately from the previous argument, since a larger M requires a larger smoothing length $R \propto M^{1/3}$, together with the fact that the number density of mass peaks of a given amplitude $\delta_{\min}$ decreases with increasing smoothing length. For large M, $n(M, z)$ decreases exponentially because sufficiently high peaks become very rare for large smoothing lengths. Therefore, *very* few clusters with mass $\gtrsim 2 \times 10^{15} M_\odot$ exist today. At higher redshift, the cut-off in the abundance is at smaller masses, so that massive clusters are expected to be increasingly rare at higher z. From Fig. 7.8, we can see that the number density of clusters with $M \gtrsim 10^{15} M_\odot$ today is about $10^{-7}\,\mathrm{Mpc}^{-3}$, so the average separation between two such clusters is larger than 100 Mpc, which is compatible with the observation that the most nearby massive cluster (Coma) is about 90 Mpc away from us.

The density contrast $\delta_{\min}$ required for a collapse before redshift z is a function of z, as we have seen above. In particular, for the Einstein–de Sitter model we have $\delta_{\min} \simeq 1.69(1+z)$. In general, $\delta_{\min} \approx 1.69/D_+(z)$. This means that the redshift dependence of $\delta_{\min}$ depends on the cosmological model and is basically described by the growth factor $D_+(z)$. Since $D_+(z)$ is, at fixed z [we recall that, by definition, $D_+(0) = 1$], larger for smaller Ω_m (see Fig. 7.3), the ratio of the number density of halos at redshift z to the one in the current Universe, $n(M, z)/n(M, 0)$, is larger the smaller Ω_m is. For cluster masses ($M \sim 10^{15} M_\odot$), the evolution of this ratio in the Einstein–de Sitter model is dramatic, whereas it is less strong in open and in flat, Λ-dominated universes (see Fig. 7.8).

Density fluctuations on a given mass scale. We considered above the smoothed density field $\delta_R(\boldsymbol{x})$, which corresponds to a mass $M = (4\pi/3)R^3\rho_0$. Like before, we here interpret δ and δ_R as the linear density field. As was true for the original density field, the expectation value—or the spatial average—of the smoothed density field vanishes, $\langle \delta_R(\boldsymbol{x}) \rangle =$

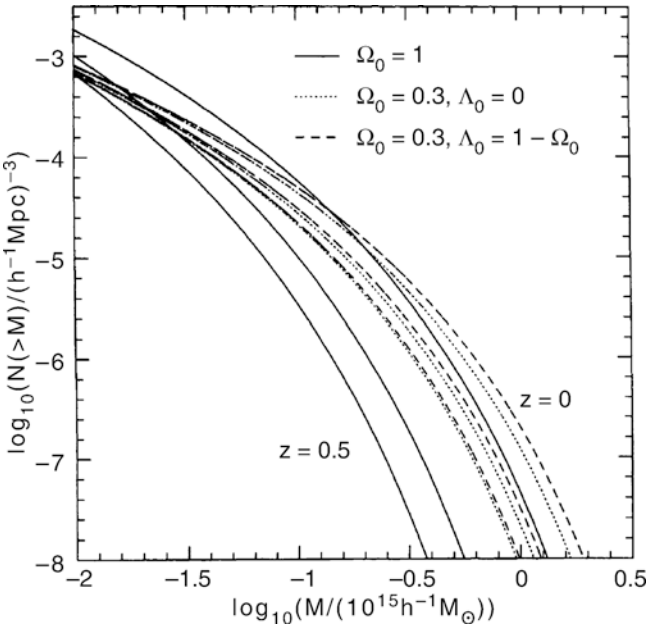


Fig. 7.8 Number density of dark matter halos with mass $> M$, computed from the Press–Schechter model. The comoving number density is shown for three different redshifts, $z = 0$ (*upper curves*), $z = 0.33$, and $z = 0.5$ (*lower curves*), for three different cosmological models: an Einstein–de Sitter model (*solid curves*), a low-density open model with $\Omega_m = 0.3$ and $\Omega_\Lambda = 0$ (*dotted curves*), and a flat universe of low density with $\Omega_m = 1 - \Omega_\Lambda = 0.3$ (*dashed curves*). The normalization of the density fluctuation field has been chosen such that the number density of halos with $M > 10^{14}h^{-1}M_\odot$ at $z = 0$ in all models agrees with the local number density of galaxy clusters. Note in particular the dramatic redshift evolution in the EdS model. Source: V.R. Eke et al. 1996, *Cluster evolution as a diagnostic for* Ω, MNRAS 282, 263, p. 269, Fig. 4. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

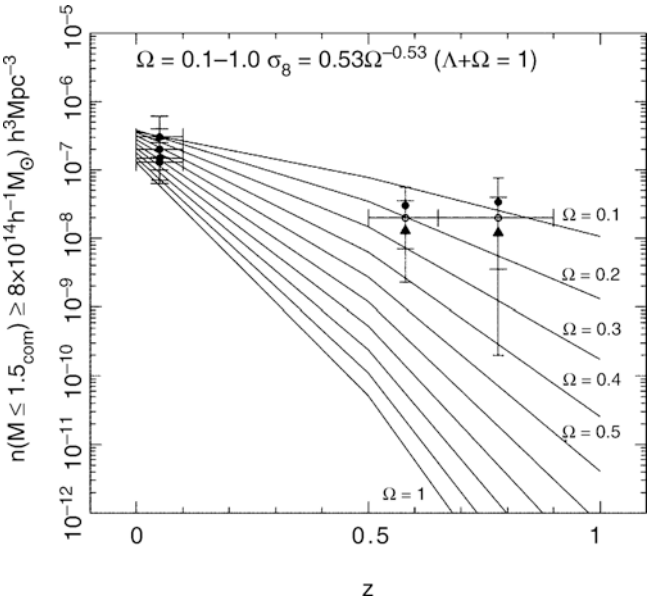


Fig. 7.9 Expected (comoving) number density of galaxy clusters with mass $> 8 \times 10^{14}h^{-1}M_\odot$ within a (comoving) radius of $R < 1.5h^{-1}$ Mpc, for flat cosmological models and different values of the density parameter Ω_m. The normalization of the power spectrum in the models has been chosen such that the current cluster number density is approximately reproduced. The *points* with error bars show results from observations of galaxy clusters at different redshift—although the error bars at high redshift are very large, a high density Universe is seen to be excluded. Source: N.A. Bahcall & X. Fan 1998, *The Most Massive Distant Clusters: Determining* Ω *and* σ_8, ApJ 504, 1, p. 2, Fig. 1. ©AAS. Reproduced with permission

0. The amplitude of fluctuations of the smoothed field can be characterized by the dispersion of the field,

$$\sigma^2(M) := \left\langle |\delta_R(\boldsymbol{x})|^2 \right\rangle , \tag{7.50}$$

which depends on the smoothing scale, and thus on the mass (these two variables are therefore interchangeable). The larger the smoothing scale, the smaller are the relative fluctuations of the resulting smoothed field. Hence, $\sigma(M)$ is a monotonically decreasing function of the mass. The larger $\sigma(M)$, the more abundant are peaks with an amplitude above some threshold at this mass scale. This is particularly true for the density threshold δ_c, required for collapse until today. Indeed, the halo abundance as predicted by Press-Schechter theory depends only on the ratio ν between the density threshold required for collapse and the dispersion of fluctuations on a given mass scale,

$$\nu := \frac{\delta_c}{D_+(z)\,\sigma(M)} , \tag{7.51}$$

where we accounted for the fact that collapse before redshift z requires a fluctuation amplitude of $\delta_c / D_+(z)$.

A first application. By comparing the number density of galaxy clusters at high redshift with the current abundance, we can thus obtain constraints on Ω_m, and in some sense also on Ω_Λ. Even a few very massive clusters at $z \gtrsim 0.5$ are sufficient to rule out the Einstein–de Sitter model by this argument. As a matter of fact, the existence of the cluster MS 1054−03 (Fig. 6.21) alone, the mass of which was determined by dynamical methods, from its X-ray emission, and by the gravitational lensing effect, is already sufficient to falsify the Einstein–de Sitter model (see Fig. 7.9).[10] However, at least one problem exists in the application of this method, namely making a sufficiently accurate mass determination for distant clusters and, in addition, determining whether they are virialized and thus are described by the Press–Schechter model. Also the completeness of the local cluster sample is a potential, though smaller problem.

[10] Until about 2000, this cluster was the highest-redshift massive cluster known.

A special case. To get a more specific impression of the Press–Schechter mass spectrum, we consider the special case where the power spectrum $P_0(k)$ is described by a power law, $P_0(k) \propto k^n$. From Fig. 7.6, we can see that this provides quite a good description over a large range of k if one concentrates on scales either clearly above or far below the maximum of P_0. The length-scale at which P_0 has its maximum is specified roughly by (7.39). As we can also see from Fig. 7.6, the non-linear evolution that the Press–Schechter model refers to is relevant only for scales considerably smaller than this maximum, rendering the power law a useful first approximation, with $n \sim -1.5$. In this case, the mass function can be written in closed form,

$$\frac{\mathrm{d}n}{\mathrm{d}M}(M,z) = \frac{\rho_{\mathrm{cr}}\Omega_{\mathrm{m}}}{\sqrt{\pi}}\,\frac{\gamma}{M^2}\left(\frac{M}{M^*(z)}\right)^{\gamma/2} \times \exp\left[-\left(\frac{M}{M^*(z)}\right)^{\gamma}\right] , \qquad (7.52)$$

where $(\mathrm{d}n/\mathrm{d}M)\mathrm{d}M$ is the comoving number density of halos with mass in the interval between M and $M + \mathrm{d}M$, $\gamma = 1 + n/3 \sim 0.5$, and $M^*(z)$ is the z-dependent mass-scale above which the mass spectrum is exponentially cut off. More specifically, $M^*(z)$ is defined as the mass where the parameter ν [see (7.51)] is unity, i.e., it is given implicitly by

$$\sigma(M^*(z)) = \delta_{\mathrm{c}}/D_+(z) . \qquad (7.53)$$

For masses considerably smaller than $M^*(z)$, the Press–Schechter mass spectrum is basically a power law in M. The characteristic mass-scale $M^*(z)$ for this particular model can be derived explicitly,

$$M^*(z) = M_0^*\,[D_+(z)]^{2/\gamma} = M_0^*\,(1+z)^{-2/\gamma} , \qquad (7.54)$$

where the final expression applies to an Einstein–de Sitter universe only. Hence, the characteristic mass-scale grows over time, and it describes the mass-scale on which the mass distribution in the universe is just becoming non-linear for a particular redshift. This mass-scale at the current epoch, M_0^*, depends on the normalization of the power spectrum; it approximately separates groups from clusters of galaxies, and explains the fact that clusters are (exponentially) less abundant than groups.

Hierarchical structure formation. Furthermore, the Press–Schechter model describes a very general property of structure formation in a CDM model, namely that low-mass structures—like galaxy-mass dark halos—form at early times, whereas large mass accumulations evolve only later. The explanation for this is found in the shape of the power spectrum $P(k)$ as described in (7.35) together with the asymptotic form (7.40) of the transfer function $T(k)$. A model like this is also called a *hierarchical structure formation* or a 'bottom-up'-scenario. In such a model, small structures that form early later merge to form large structures.

Comparison with numerical simulations. The Press–Schechter model is a very simple model, based on assumptions that are not really justified in detail. Nevertheless, its predictions are in astounding agreement with the number density of halos determined from simulations, and this model, published in 1974, has for nearly 25 years predicted the halo density with an accuracy that was difficult to achieve in numerical simulations. Only since the mid-1990s have the precision and statistics of numerical simulations of structure formation reached a level on which significant discrepancies with the Press–Schechter model became clearly noticeable. However, the analytical description was also improved; generalizing a spherical collapse, the more realistic ellipsoidal collapse has been investigated, by which the number density of halos is modified relative to the Press–Schechter model. This advanced model is found to be in better agreement with numerical results. Furthermore, simple fit functions have been found which fit the dark matter halo abundance from numerical simulations very well, as demonstrated in Fig. 7.10, so that today we have a good description of $n(M,z)$ that very accurately resembles the results from numerical simulations. These mass functions share the property that the halo abundance depends only on the 'peak height' ν, defined in (7.51).

7.5.3 Numerical simulations of structure formation

Analytical considerations—such as, for instance, linear perturbation theory or the spherical collapse model—are only capable of describing limiting cases of structure formation. In general, gravitational dynamics is too complicated to be analytically examined in detail. For this reason, experiments to simulate structure formation by means of numerical methods have been performed for some time already. The results of these simulations, when compared to observations, have contributed very substantially to establishing the standard model of cosmology, because only through them did it become possible to quantitatively distinguish the predictions of this model from those of other models. Of course, the enormous development in computer hardware rendered corresponding progress in simulations possible; in addition, the continuous improvement of numerical algorithms has allowed a steadily improved spatial resolution of the simulations.

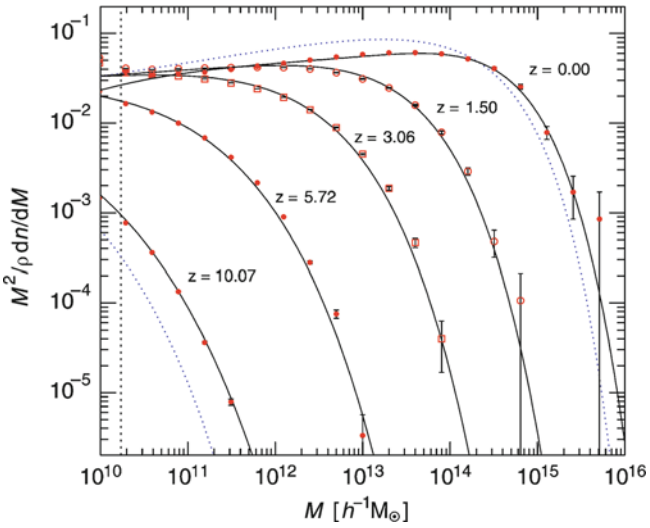


Fig. 7.10 The mass spectrum of dark matter halos is plotted for five different redshifts (data points with error bars), as measured in the Millennium Simulation (which we will discuss more extensively below—see Fig. 7.13). The *solid curves* describe an approximation for the mass spectrum, which was obtained from *different* simulations, and which obviously provides an excellent description of the simulation results. For $z = 0$ and $z = 10$, the prediction of the Press–Schechter model is indicated by the *dotted curves*, underestimating the abundance of very massive halos and overestimating the density of lower-mass halos. The *vertical dotted line* indicates the lowest halo mass which can still be resolved in these simulations. Source: V. Springel 2005, *Simulating the joint evolution of quasars, galaxies and their large-scale distribution*, Nature 435, 629, Fig. 2. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2005

Since the Universe is dominated by dark matter, for many purposes it is sufficient to compute the behavior of this matter component and thus to consider solely gravitational interactions. Only in recent years has computing power increased to a level where hydrodynamical processes can also approximately be taken into account, so that the baryonic component of the Universe can be traced as well. In addition, radiative transfer can be included in such simulations, hence the influence of radiation on the heating and cooling of the baryonic component can also be examined. We will describe such simulations in Sect. 10.6.1, when we discuss the cosmic evolution of the galaxy population.

The principle of simulations. Representative dark matter particles. We will now give a brief description of the principle of such simulations, where we confine ourselves to dark matter. Of course, no individual particles of dark matter are traced in the simulations: since it presumably consists of elementary particles, which therefore have a high number density, one would only be able to simulate an extremely small, microscopic section of the Universe. Rather, one examines the behavior of dark matter in the expanding Universe by representing its particles by bodies of mass M, and by then assuming that these "macroscopic particles" behave like the dark matter particles in a volume $V = M/\rho$. Effectively, this corresponds to the assumption that dark matter consists of particles of mass M. Since this assumption cannot be valid in detail, we will later need to modify the resulting equations of motion.

Choice of simulation volume. The next point one needs to realize right from the start is that one cannot simulate the full spatial volume of the Universe (which may be infinite) but only a representative section of it. Typically, a *comoving* cube with side length L is chosen. For this section to be representative, the linear extent L should be larger than the largest observed structures in the Universe. Otherwise, the effects of the large-scale structure would be neglected. For example, hardly any structure is found in the Universe on scales $\gtrsim 200\,h^{-1}$ Mpc, so that $L \gtrsim 200\,h^{-1}$ Mpc is a reasonable value for the comoving size of the cube. However, in a box of that size one cannot expect to get a representative result for the largest structures in the Universe (such as 'Great Walls') or for the abundance of very massive clusters, as their space density is of order $\sim 10^{-6}\left(h^{-1}\,\mathrm{Mpc}\right)^{-3}$, i.e., one expects to find at most a handful of very massive clusters in such a volume.

Since the numerical effort scales with the number of grid points at which the gravitational force is computed, $N_{\mathrm{grid}} = (L/\Delta x)^3$, and which is limited by the computer's speed and memory, the choice of L also immediately implies the length-scale of the numerical resolution. Furthermore, the total mass within the numerical volume is $\propto \Omega_{\mathrm{m}} L^3$, so that for a given maximum number of particles, the minimum mass that can be resolved in the simulation is also known.

Periodic boundary conditions. Since particles close to the boundaries of the cube also feel gravitational forces from matter outside the cube, one cannot simply assume the region outside the cube to be empty. We need to make assumptions about the matter distribution outside the numerical volume. Since one assumes that the Universe is essentially homogeneous on scales $> L$, the cube is extended periodically—for instance, a particle leaving the cube at its upper boundary will immediately re-enter the cube from the lower side, with the same velocity vector. The mass distribution (and with it also the force field) is periodic in these simulations, with a period of L. This assumption of periodicity has an effect on the results for the mass distribution on scales comparable to L; the quantitative analysis of the results from these simulations should therefore be confined to scales $\lesssim L/2$.

Computation of the force field. With the above assumptions, the equation of motion for all particles can now be set up. The force on the i-th particle is

$$\boldsymbol{F}_i = \sum_{j \neq i} \frac{M^2\,(\boldsymbol{r}_j - \boldsymbol{r}_i)}{\left|\boldsymbol{r}_j - \boldsymbol{r}_i\right|^3}\,, \tag{7.55}$$

thus the sum of forces exerted by all the other particles, where these are periodically extended. This aspect may appear at first sight more difficult than it actually is, as we will see next.

The computation of the force acting on individual particles by the summation (7.55) is not feasible in practice. For example, assume the simulation to trace 10^{10} particles, then in total 10^{20} terms need to be calculated using (7.55)—for each time step. Even on the most powerful computers this is not feasible today. To handle this problem, one evaluates the force in an approximate way. One first notes that the force experienced by the i-th particle, exerted by the j-th particle, is not very sensitive to small variations in the separation vector $\boldsymbol{r}_i - \boldsymbol{r}_j$, as long as these variations are much smaller than the separation itself. Except for the nearest particles, the force on the i-th particle can then be computed by introducing a grid into the cube and shifting the particles

in the simulation to the closest grid point.[11] With this, a discrete mass distribution on a regular grid is obtained.

The force field of this mass distribution can then be computed by means of a Fast Fourier Transform (FFT), a fast and very efficient algorithm. However, the introduction of the grid establishes a lower limit to the spatial force resolution. Because the size of the grid cells also defines the spatial resolution of the force field, it is chosen to be roughly the mean separation between two particles, so that the number of grid points is typically of the same order as the number of particles. This is called the PM (particle-mesh) method.

To achieve better spatial resolution, the interaction of closely neighboring particles can considered separately. This is done by splitting the gravitational potential $\Phi(r) = -GM/r$ of a particle into a short- and long-range part, $\Phi(r) = \Phi_\mathrm{s}(r) + \Phi_\mathrm{l}$. For example, one can choose $\Phi_\mathrm{s}(r) = \Phi(r)\, f(r/r_\mathrm{s})$, where the function f smoothly declines from $f(0) = 1$ to $f(1) = 0$, and $f(x) = 0$ for $x > 1$. Thus, the short-range gravitational potential $\Phi_\mathrm{s}(r)$ vanishes for $r > r_\mathrm{s}$. The long-range potential then is $\Phi_\mathrm{l}(r) = \Phi(r)\,[1 - f(r/r_\mathrm{s})]$, and hence vanishes at $r = 0$, whereas for $r > r_\mathrm{s}$, $\Phi_\mathrm{l} = \Phi$. The force on a particle is then given by the sum of the gradients of the short- and long-range potential. For the former, only those particles with separation $\le r_\mathrm{s}$ contribute, and this can be calculated by a sum of pairwise forces. On the other hand, the force field corresponding to Φ_l is smooth and is calculated by the grid method, as explained before. This kind of calculation of the force is called the $\mathrm{P^3M}$ (particle-particle particle-mesh) method.

Softening length. The force law (7.55) also describes strong collisions of particles, e.g., where a particle changes its velocity direction by $\sim 90°$ in a collision if it comes close enough to another particle. Of course, this effect is a consequence of replacing the dark matter constituents by macroscopic 'particles' of mass M. As we have seen in Sect. 3.2.4, the typical relaxation time-scale for a system is $\propto N/\ln N$, and since the mass in the numerical volume is defined by L, one has $N \propto 1/M$. Reducing the particles' mass and increasing N accordingly, the abundance of strong collisions would decrease, but computer power and memory is then a limiting factor. Thus to correct for the artefact of strong collisions, the force law is modified for small separations such that strong collisions no longer occur. The length-scale below which the force equation is modified ('softened') and deviates from $\propto 1/r^2$ is called *softening length*. Its choice depends on the method with which the force field is evaluated. If the force is calculated on a grid, as in the PM method, where the force resolution is limited by the size of grid cells, the softening length is typically chosen to be of similar size. On the other hand, for $\mathrm{P^3M}$-like methods, the softening length can be chosen substantially smaller, however at the expense of requiring smaller time steps. Of course, the softening length defines the smallest length scale on which the results of the simulation can be trusted: scales below or comparable to the softening length are not resolved, and the behavior on these small scales is affected by numerical artefacts.

Initial conditions and evolution. The initial conditions for the simulation are set at very high redshift. The particles are then distributed such that the power spectrum of the resulting mass distribution resembles a Gaussian random field with the theoretical (linear) power spectrum $P(k, z)$ of the cosmological model. The equations of motion for the particles with the force field described above are then integrated in time. The choice of the time step is a critical issue in this integration, as can be seen from the fact that the force on particles with relatively close neighbors will change more quickly than that on rather isolated particles. Hence, the time step is either chosen such that it is short enough for the former particles—which requires substantial computation time—or the time step is varied for different particles individually, which is clearly the more efficient strategy. For different times in the evolution, the particle positions and velocities are stored; these results are then available for subsequent analysis.

Examples of simulations. The size of the simulations, measured by the number of particles considered, has increased enormously in recent years with the corresponding increase in computing capacities and the development of efficient algorithms. In modern simulations, 1024^3 or even more particles are traced. One example of such a simulation is presented in Fig. 7.11, where the structure evolution was computed for four different cosmological models. The parameters for these simulations and the initial conditions (i.e., the initial realization of the random field) were chosen such that the resulting density distributions for the current epoch (at $z = 0$) are as similar as possible; by this, the dependence of the redshift evolution of the density field on the cosmological parameters can be recognized clearly. Comparing simulations like these with observations has contributed substantially to our realizing that the matter density in our Universe is considerably smaller than the critical density.

Massive clusters of galaxies have a very low number density, which can be seen from the fact that the massive cluster closest to us (Coma) is about 90 Mpc away. This is directly related to the exponential decrease of the abundance of dark matter halos with mass, as described by the Press–Schechter model (see Sect. 7.5.2). In simulations such as that shown in Fig. 7.11, the simulated volume is still too small to derive statistically meaningful results on such sparse mass concentrations. This difficulty has been one of the reasons for simulating considerably larger volumes. The Hubble Volume Simulations (see Fig. 7.12) use a cube with a side length of $3000h^{-1}$ Mpc, not much less than the currently visible Universe. This simulation is particularly well-suited to studying the statistical properties of very massive structures, like, e.g., the distribution of galaxy clusters. On the other hand, this large volume, together with the limited total number of particles that can be followed, means that the mass and spatial resolution of this simulation are insufficient for studying smaller-scale objects like galaxies.

The Millennium Simulation (MS) was performed in 2004, assuming a cosmological model with $\Omega_\mathrm{m} = 0.25$, $\Omega_\Lambda = 0.75$, a power spectrum normalization of $\sigma_8 = 0.9$, and a Hubble constant of $h = 0.73$. A cube of side length $500h^{-1}$ Mpc was considered, in which $(2160)^3 \approx 10^{10}$ particles with a mass of $8.6 \times 10^8 h^{-1}\, M_\odot$ each were traced. With this choice of parameters, one can spatially resolve the halos of galaxies. At the same time, the volume is large enough for the simulation to contain a large number of massive clusters whose evolutionary history can be followed.

[11] In practice, the mass of a particle is distributed to all eight neighboring grid points, with the relative proportion of the mass depending on the distance of the particle to each of these grid points.

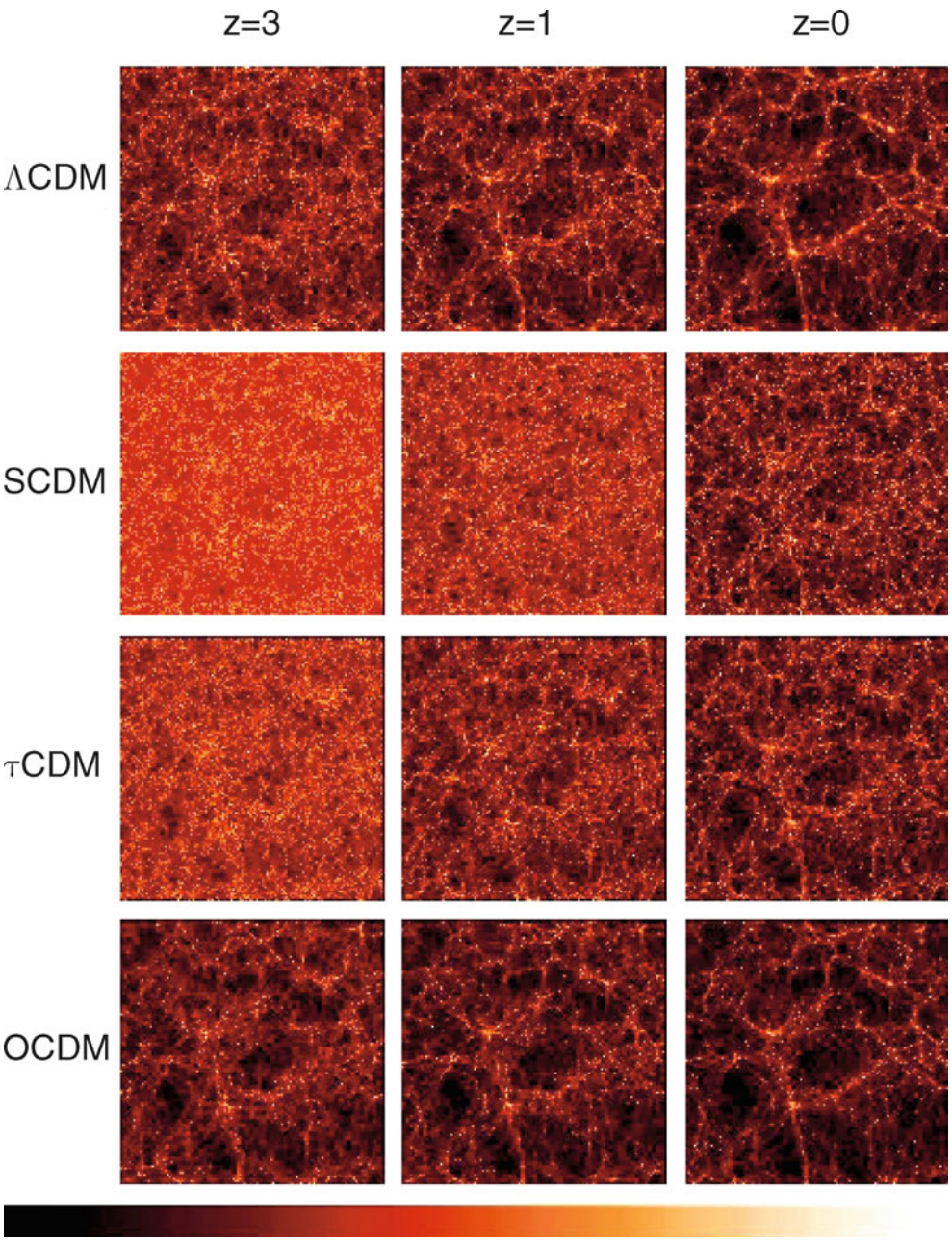


Fig. 7.11 Simulations of the dark matter distribution in the Universe for four different cosmological models: $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0.7$ (ΛCDM), $\Omega_\mathrm{m} = 1.0$, $\Omega_\Lambda = 0.0$ (SCDM and τCDM), and $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0$ (OCDM). The two Einstein–de Sitter models differ in their shape parameter Γ which specifies the shape of the power spectrum $P(k)$ through the location of its peak. For each of the models, the mass distribution is presented for three different redshifts, $z = 3$, $z = 1$, and today, $z = 0$. Whereas the current mass distribution is quite similar in all four models (the model parameters were chosen as such), they clearly differ at high redshift. We can see, for instance, that significantly less structure has formed at high redshift in the SCDM model compared to the other models. From the analysis of the matter distribution at high redshift, one can therefore distinguish between the different models. In these simulations by the VIRGO Consortium, 256^3 particles were traced; the side length of the simulated volume is $\sim 240h^{-1}$ Mpc. Credit: VIRGO Collaboration, J. Colberg/MPA Garching. The simulations were carried out by the Virgo Supercomputing Consortium using computers based at the Computing Centre of the Max-Planck Society in Garching and at the Edinburgh parallel Computing Centre. Research article: A. Jenkins et al. 1998, *Evolution of Structure in Cold Dark Matter Universes*, ApJ 499, 20

The spatial resolution of the simulation is $\sim 5h^{-1}$ kpc, yielding a linear dynamic range of $\sim 10^5$. The resulting mass distribution at $z = 0$ is displayed in Fig. 7.13 in slices of $15h^{-1}$ Mpc thickness each, where the linear scale changes by a factor of four from one slice to the next. The images zoom in to a region around a massive cluster that becomes visible with its rich substructure in the uppermost slice, as well as filaments of the matter distribution, at the intersections of which massive halos form. The mass distribution in the Millennium Simulation is of great interest for numerous different investigations. We will discuss some of its results further in Chap. 10.

The MS was complemented by two related simulations, the Millennium-II (MS-II) and the Millennium-XXL (MXXL). Both assume the same cosmological parameters as the original Millennium Simulation, but differ in the volume considered. MS-II has the same number of particles as the original MS, but a five times smaller box size, i.e., $L = 100h^{-1}$ Mpc, yielding 125 times better mass resolution. MXXL treats a considerably larger box with $L = 3000h^{-1}$ Mpc, yielding the same comoving volume as the whole observable Universe within a redshift of $z = 0.72$, and $6720^3 \approx 3 \times 10^{11}$ particles, which makes it one of the largest N-body simulations up to now (2014). Together,

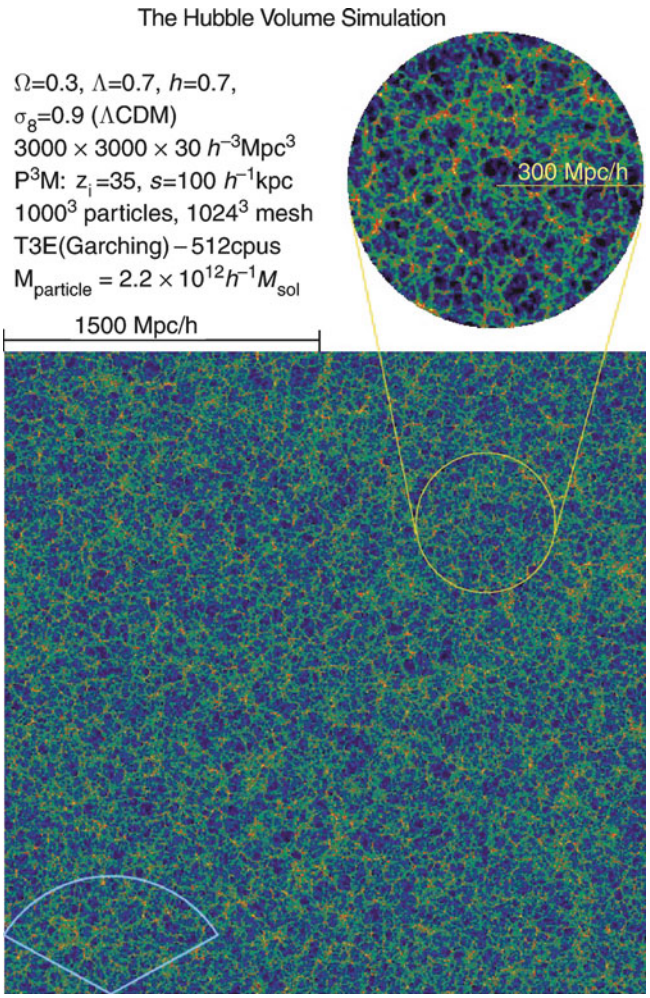


Fig. 7.12 The Hubble Volume Simulations: simulated is a box of volume $(3000h^{-1}\,\mathrm{Mpc})^3$, containing 10^9 particles, where a ΛCDM model with $\Omega_\mathrm{m} = 0.3$ and $\Omega_\Lambda = 0.7$ was chosen. Displayed is the projection of the density distribution of a $30h^{-1}$ Mpc thick slice of the cube. Simulations like this can be used to analyze the statistical properties of the mass distribution in the Universe on large scales. The sector in the lower left corner represents roughly the size of the CfA redshift survey (see Fig. 7.2). Credit: VIRGO Collaboration. The simulations were carried out by the Virgo Supercomputing Consortium using computers based at the Computing Centre of the Max-Planck Society in Garching and at the Edinburgh parallel Computing Centre. Research article: J.M. Colberg et al. 2000, *Clustering of galaxy clusters in CDM universes*, MNRAS, 319, 209

these three simulations can be used to study the effects of numerical resolution. For example, the combination of the simulations can measure the abundance of dark matter halos over nearly eight orders of magnitude in mass (see Fig. 7.14).

Analysis of numerical results. The analysis of the numerical results is nearly as intricate as the simulation itself because the positions and velocities of some 10^{10} particles alone do not provide any new insights. However, properly displaying the resulting mass distribution offers a first important insight into the large-scale structure in a CDM universe. In Fig. 7.15, we can see strong concentrations of the matter distribution, interlinked by large filaments. The overall structure resembles that of a web—often called the 'Cosmic Web'. The mass is assembled in sheets; where two sheets intersect, filaments form, and at the intersection of filaments are the more massive dark matter halos. These halos themselves are not smooth, but contain a rich substructure, as one recognizes in the smaller-scale zooms of Fig. 7.15. More insight is provided by looking at the time evolution of the density field; Fig. 7.16 shows four snapshots of the same region in the Millennium-II simulation. First, the density contrast increases with time, as expected from gravitational instability. At high redshift, only small mass concentrations have formed; more massive ones show up only at smaller redshift, in accordance to what we discussed in Sect. 7.5.2: Low-mass halos form first, massive ones only later. This is the hierarchical built-up of structure in a CDM universe. The halos grow in mass due to two different processes, by merging with other halos and by accretion of matter. The first process leads to the substructure (subhalos) of halos, which are the relics of earlier mergers. The accretion of matter is due to the gravitational pull of the mass concentrations on the surrounding matter. This accretion process happens predominantly through the filaments which are connected to the halo—it is therefore not a spherically symmetric effect.

The output of the simulation needs to be analyzed with respect to specific questions. Obviously, the (non-linear) power spectrum $P(k, z)$ of the matter distribution can be computed from the spatial distribution of particles, i.e., the density field; the corresponding results have led to the construction of the analytic fit formulae presented in Fig. 7.6. Furthermore, one can search for voids in the resulting particle distribution, which can then be compared to the observed abundance and typical size of voids.

Identification of dark matter halos. One of the main applications is the identification of collapsed mass concentrations (i.e., dark matter halos); their number density can be compared to predictions from the Press–Schechter model and to observations. For this, one needs to specify what a dark matter halo is and how this specification can be applied to the output of simulations. The spherical collapse model suggests that a halo is a spherical region inside of which the mean density is $\sim$ 200 times the critical density, but we recall that this particular value of 200 was based on a number of idealized assumptions. Furthermore, the dark matter concentrations usually deviate quite strongly from spherical symmetry.

Different methods for the identification of halos are used in simulations; for example, based on the position and velocities of particles, one can consider a halo to consist of

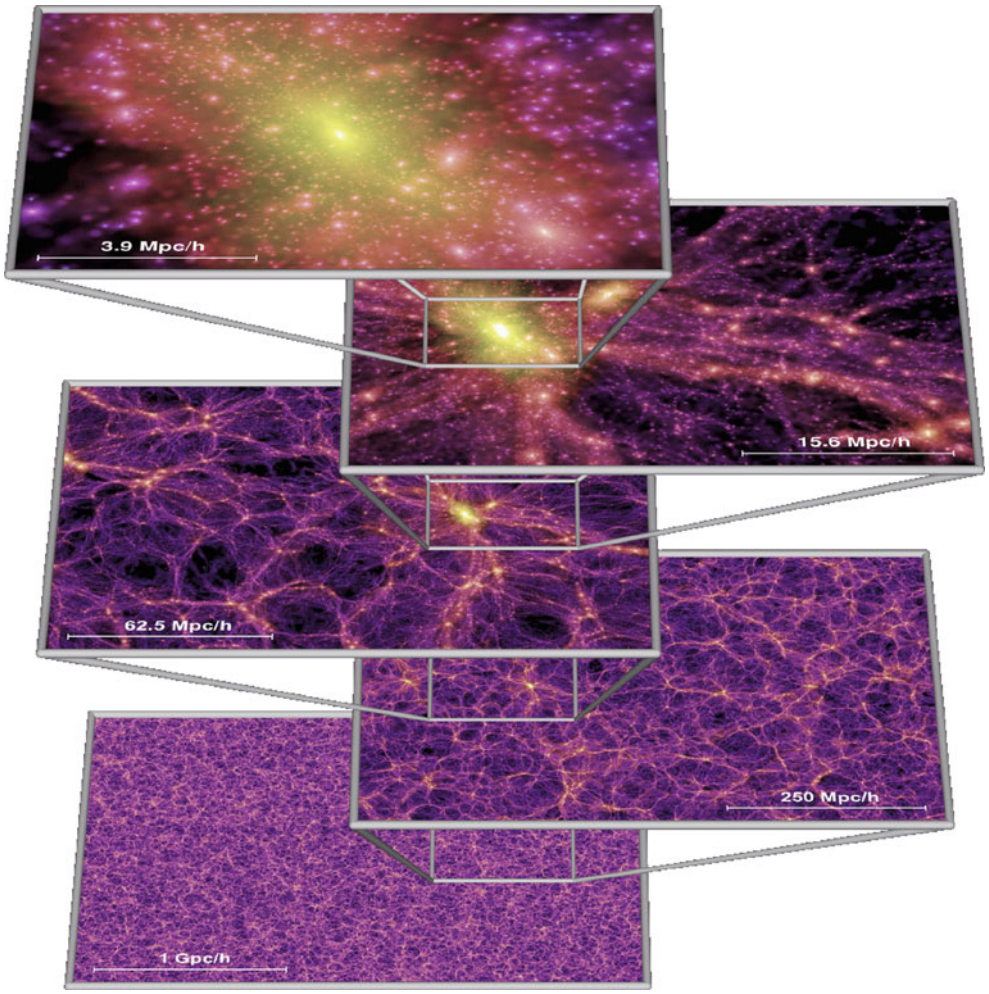


Fig. 7.13 Distribution of matter in slices of thickness $15h^{-1}$ Mpc each, computed in the Millennium Simulation. This simulation took about a month in 2004, running on 512 CPU processors. The output of the simulation, i.e., the position and velocities of all 10^{10} particles at 64 times steps, has a data volume of ~ 27 TB. The region shown in the two lower slices is larger than the simulated box which has a sidelength of $500h^{-1}$ Mpc; nevertheless, the matter distribution shows no periodicity in the figure as the slice was cut at a skewed angle to the box axes. Source: V. Springel 2005, *Simulating the joint evolution of quasars, galaxies and their large-scale distribution*, Nature 435, 629, Fig. 1. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2005

all particles which are gravitationally bound to it. Perhaps the most frequently used method consists of linking all particles whose separation is smaller than a fraction b of the mean particle separation $\sqrt{1/n}$, where $n = N/L^3$ is the number density of particles. Then those particles which are connected by a link are considered to be members of a halo; this is called the friends-of-friends algorithm. One finds from numerical experiments that by choosing $b = 0.2$, the characteristic density of these halos is about 200 times the critical density. Obviously, the way a halo is defined will affect the resulting mass spectrum, as can be seen in Fig. 7.14, where results from the 'gravitationally bound' and the friends-of-friends method are compared. The same ambiguity arises when the abundance of dark matter halos from simulations are compared to observed abundances of astronomical objects, such as galaxies or clusters.

It has been found that the Press–Schechter mass function represents the basic aspects of the mass spectrum astonishingly well, but a comparison with more recent simulations has shown significant deviations. More accurate formulae for the mass spectrum of halos have been constructed from simulations (see Figs. 7.10 and 7.14).

The direct link between the results from dark matter simulations and the observed properties of the Universe requires an understanding of the relation between dark matter and luminous matter. Dark matter halos in simulations cannot be compared to the observed galaxy distribution without further assumptions, e.g., on the mass-to-light ratio. We will return to these aspects later.

7.6 Properties of dark matter halos

The bound and virialized structures formed in the evolution of the cosmic density field, i.e., the dark matter halos, are of particular interest since they are believed to host the luminous objects in the Universe, galaxies and clusters. Therefore, we shall describe their properties in more detail here.

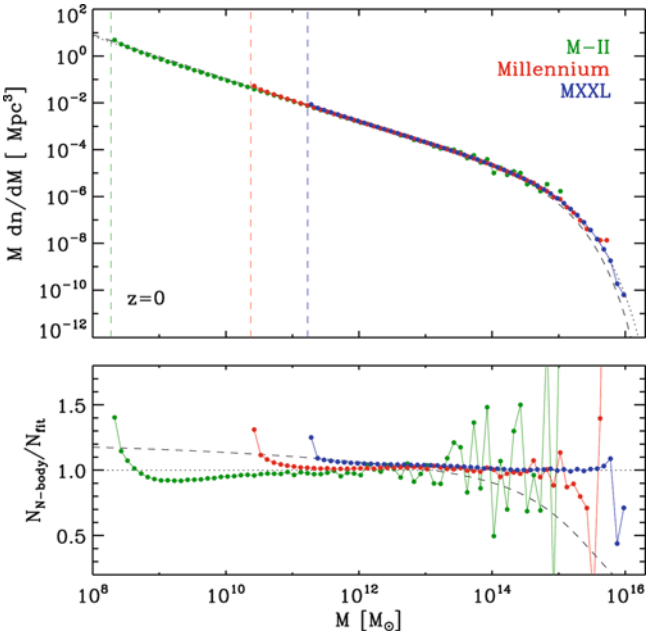


Fig. 7.14 The *upper panel* shows the comoving abundance of dark matter halos at $z = 0$, as obtained from the Millennium Simulation, the Millennium-II and the Millennium-XXL, shown in *red*, *green* and *blue*, respectively. Due to the different size and mass resolution, each of these simulations can determine the abundance best in a limited range of masses; together, they cover almost a factor 10^8 in mass. The *vertical dashed lines* indicates the minimum mass for each simulation, taken to be the mass of 20 simulation particles. The *dotted curve* shows a fit to the mass spectrum. In the *lower panel*, the ratio of the halo abundance and the fit is shown. The agreement between the results of these three simulations is clearly seen. Here, halos are identified by the friends-of-friends method. The *dashed curve* shows a fit to the halo abundance when a halo is identified by the set of all self-bound particles. Obviously, the mass spectrum depends on the details of the characterization of a halo. Source: R.E. Angulo et al. 2012, *Scaling relations for galaxy clusters in the Millennium-XXL simulation*, MNRAS 426, 2046, Fig. 2. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

7.6.1 Profile of dark matter halos

As already mentioned above, dark matter halos can be identified in mass distributions generated by numerical simulations. Besides the abundance of halos as a function of their mass and redshift, their radial mass profile can also be analyzed if individual halos are represented by a sufficient number of dark matter particles. The ability to obtain halo mass profiles depends on the mass resolution of a simulation. A surprising result has been obtained from these studies, namely that halos seem to show a universal density profile. We will briefly discuss this result in the following.

If we define a halo as described above, i.e., as a spherical region within which the average density is $\sim$ 200 times the critical density at the respective redshift, the mass M of the halo is related to its (virial) radius r_{200} by

$$M = \frac{4\pi}{3} r_{200}^3 \, 200 \, \rho_{\rm cr}(z) \; .$$

Since the critical density at redshift z is specified by $\rho_{\rm cr}(z) = 3H^2(z)/(8\pi G)$, we can write this as

$$M = \frac{100 r_{200}^3 H^2(z)}{G} \; , \tag{7.56}$$

so that at each redshift, a unique relation exists between the halo mass and its radius. We can also define the virial velocity V_{200} of a halo as the circular velocity at the virial radius,

$$V_{200}^2 = \frac{GM}{r_{200}} \; . \tag{7.57}$$

Combining (7.56) and (7.57), we can express the halo mass and virial radius as a function of the virial velocity,

$$M = \frac{V_{200}^3}{10GH(z)} \; , \quad r_{200} = \frac{V_{200}}{10H(z)} \; . \tag{7.58}$$

Since the Hubble function $H(z)$ increases with redshift, the virial radius at fixed virial velocity decreases with redshift. From (7.56) we also see that r_{200} decreases with redshift at fixed halo mass. Hence, halos at a given mass (or given virial velocity) are more compact at higher redshift than they are today, because the critical density was higher in the past.

The NFW profile. The *density profile of halos* averaged over spherical shells seems to have a universal functional form, which was first reported by Julio Navarro, Carlos Frenk & Simon White in a series of articles in the mid-1990s. This *NFW-profile* is described by

$$\boxed{\rho(r) = \frac{\rho_{\rm s}}{(r/r_{\rm s})(1 + r/r_{\rm s})^2}} \; , \tag{7.59}$$

where $r_{\rm s}$ specifies a characteristic radius, and $\rho_{\rm s} = 4\rho(r_{\rm s})$ determines the amplitude of the density profile. For $r \ll r_{\rm s}$ we find $\rho \propto r^{-1}$, whereas for $r \gg r_{\rm s}$, the profile follows $\rho \propto r^{-3}$. Therefore, $r_{\rm s}$ is the radius at which the slope of the density profile changes (see Fig. 7.17). $\rho_{\rm s}$ can be expressed in terms of $r_{\rm s}$, since, according to the definition of r_{200},

$$\begin{aligned} \bar{\rho} = 200\rho_{\rm cr}(z) &= \frac{3}{4\pi r_{200}^3} \int_0^{r_{200}} 4\pi r^2 \, {\rm d}r \, \rho(r) \\ &= 3\rho_{\rm s} \int_0^1 \frac{{\rm d}x \; x^2}{c \; x \; (1 + cx)^2} \; , \end{aligned} \tag{7.60}$$

where in the last step the integration variable was changed to $x = r/r_{200}$, and the *concentration index*

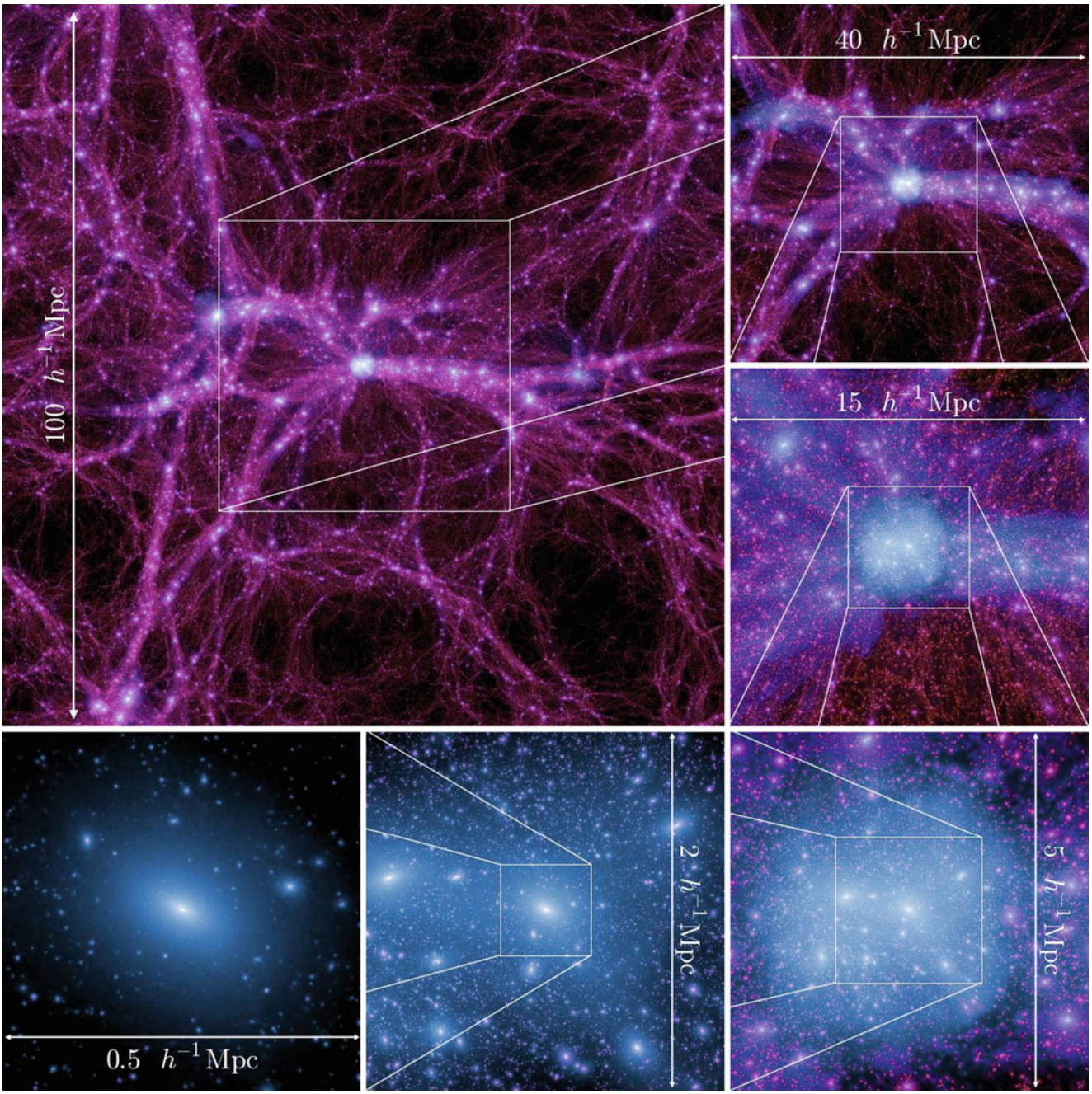


Fig. 7.15 The Millennium-II simulation. The *large upper-left panel* shows a $15h^{-1}$ Mpc slice through the full simulation, while the other *panels* display subsequent zooms of the central region, where the most massive halo in the simulation is located, with decreasing thickness of the slices. Source: M. Boylan-Kolchin et al. 2009, *Resolving cosmic structure formation with the Millennium-II Simulation*, MNRAS 398, 1150, p. 1153, Fig. 1. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

$$\boxed{c := \frac{r_{200}}{r_s}} \tag{7.61}$$

was defined. The larger the value of c, the more strongly the mass is concentrated towards the inner regions. Equation (7.60) implies that ρ_s can be expressed in terms of $\rho_{cr}(z)$ and c, and performing the integration in (7.60) yields

$$\rho_s = \frac{200}{3}\rho_{cr}(z)\,\frac{c^3}{\ln(1+c) - c/(1+c)}\,.$$

Since M is determined by r_{200}, the NFW profile is parametrized by r_{200} (or by the mass of the halo) and by the concentration c that describes the shape of the distribution. Simulations show that the concentration index c is strongly

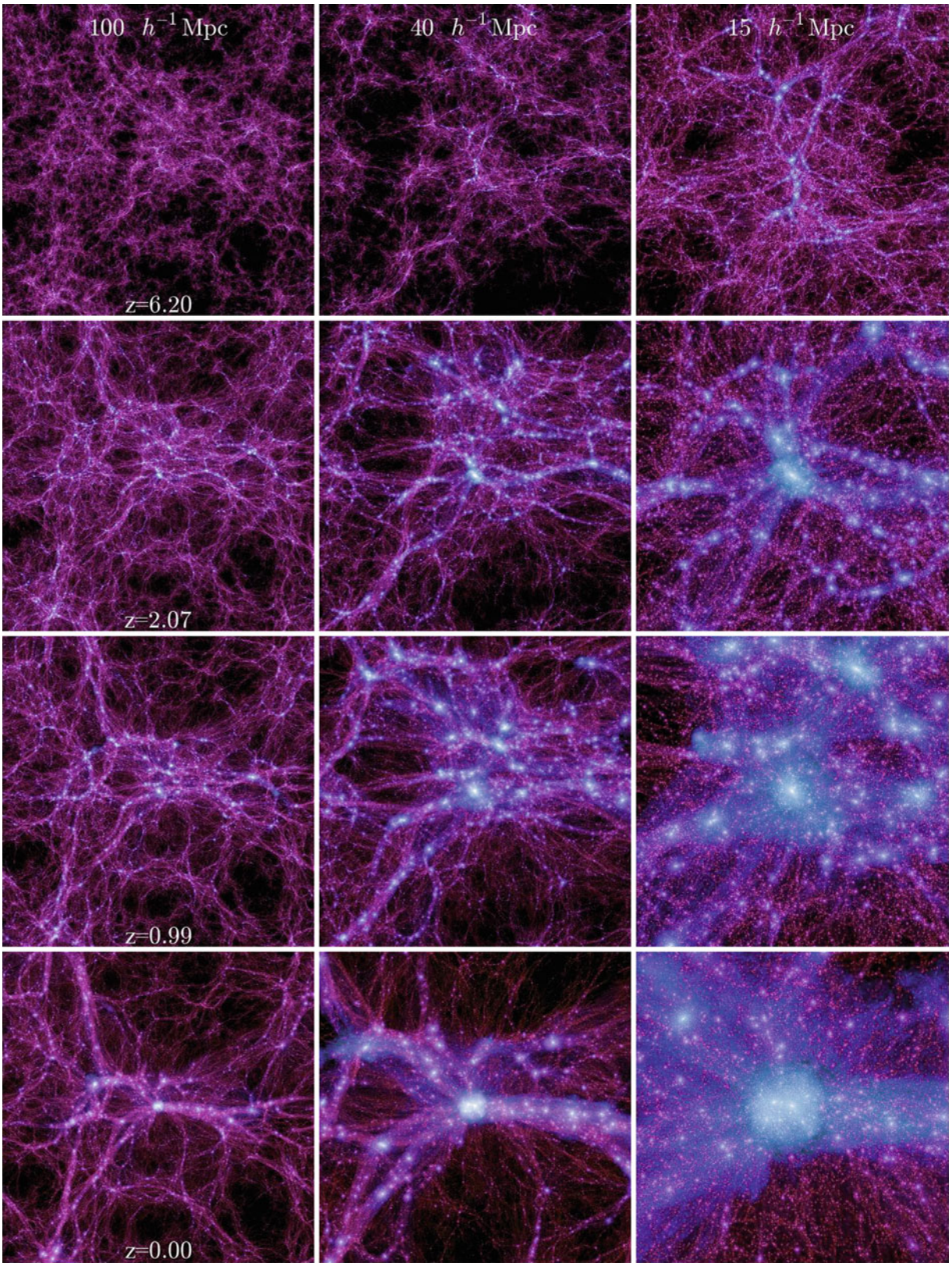


Fig. 7.16 Time evolution in the Millennium-II simulation. The most massive halo in the simulation is shown at four different redshifts, and three spatial resolutions. The thickness of the slices from left to right is 15, 10, and $6h^{-1}$ Mpc, respectively. Source: M. Boylan-Kolchin et al. 2009, *Resolving cosmic structure formation with the Millennium-II Simulation*, MNRAS 398, 1150, p. 1155, Fig. 2. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

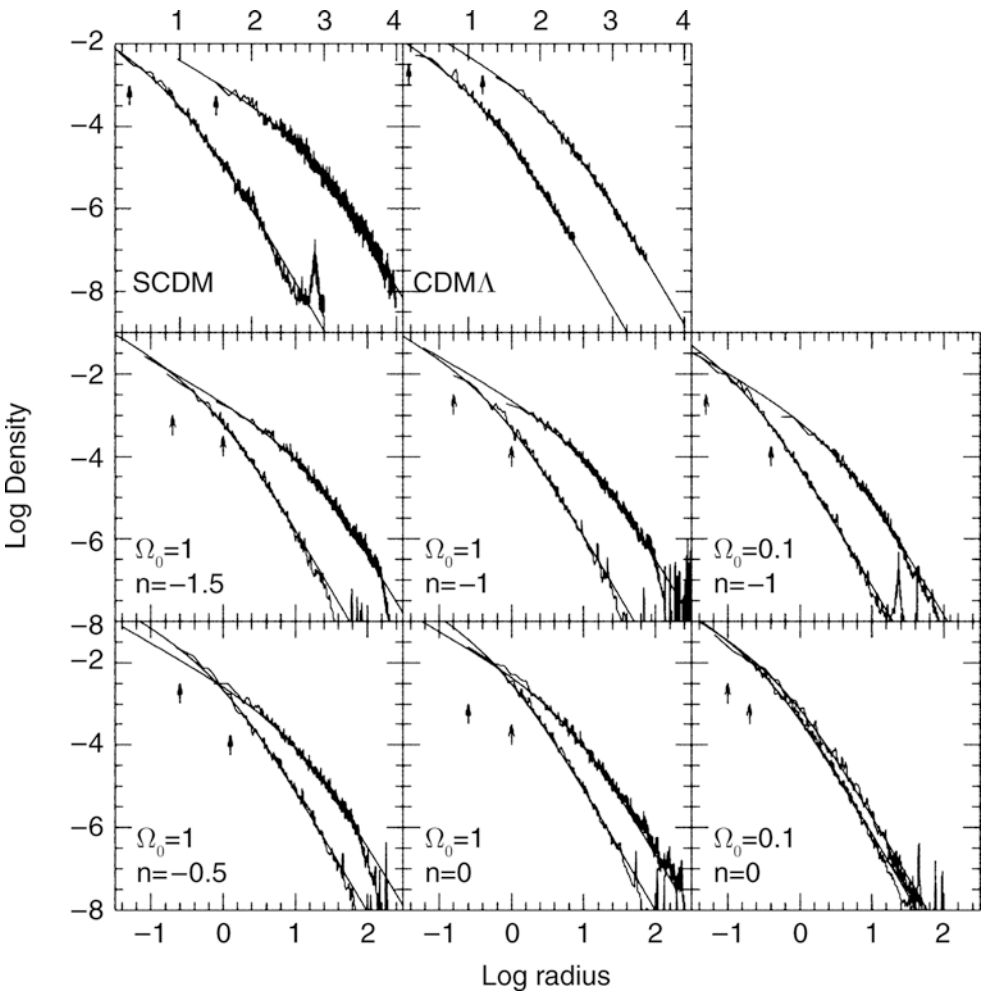


Fig. 7.17 For eight different cosmological simulations, the density profile is shown for the most massive and the least massive halo, each as a function of the radius, together with the best fitting density profile (7.59). The cosmological models represent an EdS model (here denoted by SCDM, *top left*), a ΛCDM model (*top right*), and different models with power spectra that are assumed to be power laws locally, $P(k) \propto k^n$. The *arrows* indicate the softening length in the gravitational force for the respective halos; thus, the major part of the profiles is numerically well resolved. Source: J.F. Navarro et al. 1997, *A Universal Density Profile from Hierarchical Clustering*, ApJ 490, 493, p. 496, Fig. 2. ©AAS. Reproduced with permission

correlated with the mass and the redshift of the halo; one finds approximately

$$c \approx 6.7 \left(\frac{M}{2 \times 10^{12} h^{-1} M_\odot} \right)^{-0.1} (1+z)^{-0.5} \qquad (7.62)$$

for relaxed halos. A similar result can also be obtained from analytical scaling arguments, under the assumption of the existence of a universal density profile. In Fig. 7.18, the density profile of dark matter halos is plotted as a function of the scaled radius r/r_{200}, where the similarity in the profile shapes for the different simulations becomes clearly visible, as well as the dependence of the concentration index on the halo mass. The range over which the density distribution of numerically simulated halos is described by the profile (7.59) is bounded above by the virial radius r_{200}, whereas in the central region of halos the numerical resolution of the simulations is too low to test (7.59) for very small r. The latter comment concerns the inner $\sim 1\,\%$ of the halo mass.

Generalization. No good analytical argument has yet been found for the existence of such a universal density profile, in particular not for the specific functional form of the NFW profile. As a matter of fact, other numerical simulations found slightly different density profiles, in particular towards the center. The reason for the differences between different simulations may be related to resolution issues. More recent numerical results have established a slight deviations of the mean halo profile from the NFW law, showing that a better fit is provided by the so-called Einasto profile,

$$\rho(r) = \rho_s \exp\left(\frac{-2}{\alpha} \left[\left(\frac{r}{r_s} \right)^\alpha - 1 \right] \right) , \qquad (7.63)$$

where r_s is a scale radius, ρ_s is the density at the scale radius, and α determines the overall shape; to good accuracy, $\alpha \sim 0.17$, though its value depends somewhat on the halo mass. The slope of an Einasto profile is a power law in radius, since

$$\frac{\mathrm{d} \ln \rho}{\mathrm{d} \ln r} = -2 \left(\frac{r}{r_s} \right)^\alpha . \qquad (7.64)$$

Thus, at the scale radius, the slope is -2, and it gradually decreases towards the center. This profile is not truly cuspy, since the slope approaches zero as $r \to 0$, however, due to the smallness of α, it does so very slowly.

Comparison with observations. The comparison of these theoretical profiles with an observed density distribution is

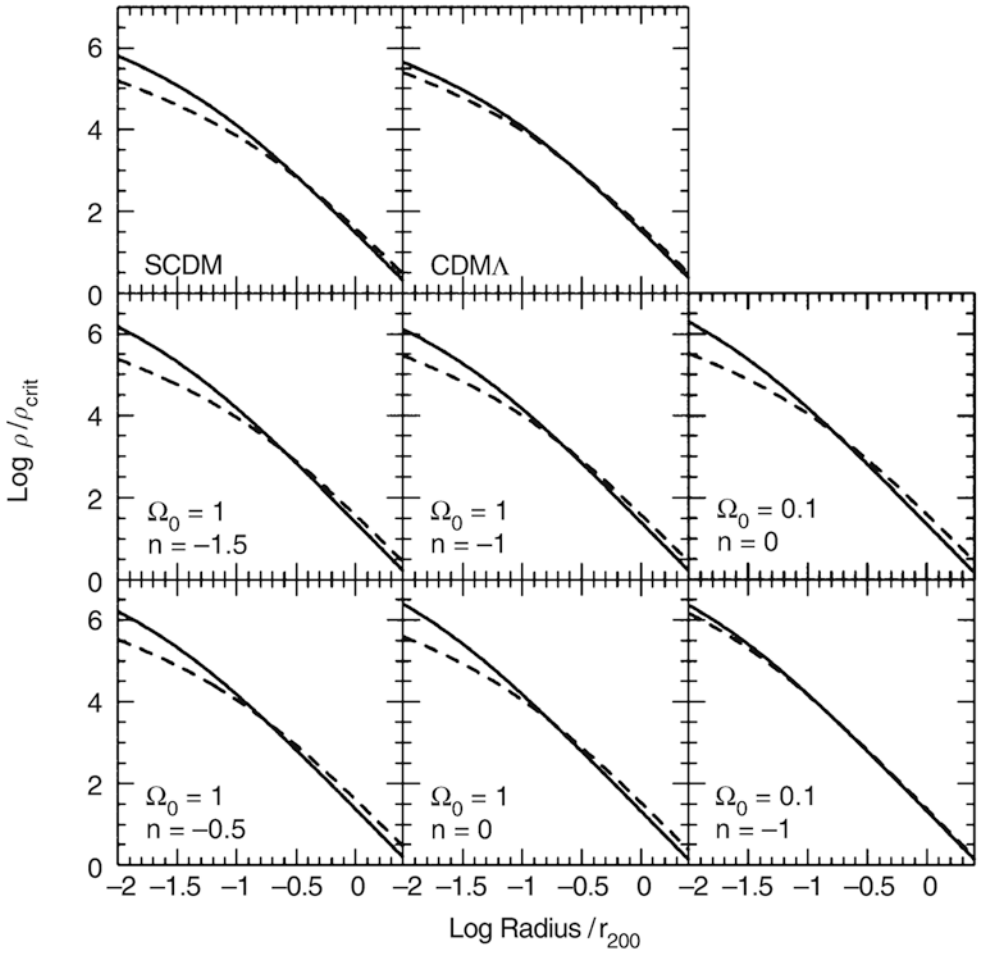


Fig. 7.18 The density profiles from Fig. 7.17, but now the density is scaled by the critical density, and the radius scaled by r_{200}. *Solid* (*dashed*) *curves* correspond to halos of low (high) mass—thus, halos of low mass are relatively denser close to the center, and they have a higher concentration index c. Source: J.F. Navarro et al. 1997, *A Universal Density Profile from Hierarchical Clustering*, ApJ 490, 493, p. 497, Fig. 3. ©AAS. Reproduced with permission

by no means simple because the density profile of dark matter is of course not directly observable. For instance, in normal spiral galaxies, $\rho(r)$ is dominated by baryonic matter at small radii. In the Milky Way, roughly half of the matter within R_0 consists of stars and gas, so that only little information is provided on $\rho_{\rm DM}$ in the central region. It is often assumed that galaxies with very low surface brightness (LSBs) are dominated by dark matter well into the center. The rotation curves of LSB galaxies are apparently *not* in agreement with the expectations from the NFW model shown in Fig. 7.19; in particular, they provide no evidence of a cusp in the central density distribution ($\rho \to \infty$ for $r \to 0$).

Part of this discrepancy may perhaps be explained by the finite angular resolution of the 21 cm line measurements of the rotation curves; however, the discrepancy remains if higher-resolution rotation curves are measured using optical long-slit and integral-field spectroscopy. As an additional point, the kinematics of these galaxies may be more complicated, and in some cases their dynamical center is difficult to determine. The orbits of stars and gas in these galaxies may show a more complex behavior than expected from a smooth density profile. The mass distribution in (the inner parts of) a dark matter halo is neither smooth nor axially symmetric, and stars and gas do not move on circular orbits in a thin plane of symmetry. Instead, simulations show that the pressure support of the gas, together with non-circular motions and projection effects systematically underestimate the rotational velocity in the center of dark matter halos, thereby creating the impression of a constant density core. Nevertheless, the observed rotation curves of LSB galaxies may prove to be a major problem for the CDM model—hence, this potential discrepancy must be resolved.

An additional complication is the fact that not only is baryonic matter present in the inner regions of galaxies (and clusters), thus contributing to the density, but also these baryons have modified the density profile of dark matter halos in the course of cosmic evolution. Baryons are dissipative, they can cool, form a disk, and accrete inwards. Also the opposite can happen: the explosion of supernovae can push some of the gas to larger radii, or even drive it out of the halo, in particular for low-mass ones. The changes in the resulting density distribution of baryons by dissipative processes cause a change of the gravitational potential over time, to which dark matter also reacts. The dark matter profile in real galaxies is thus modified compared to pure dark matter simulations.

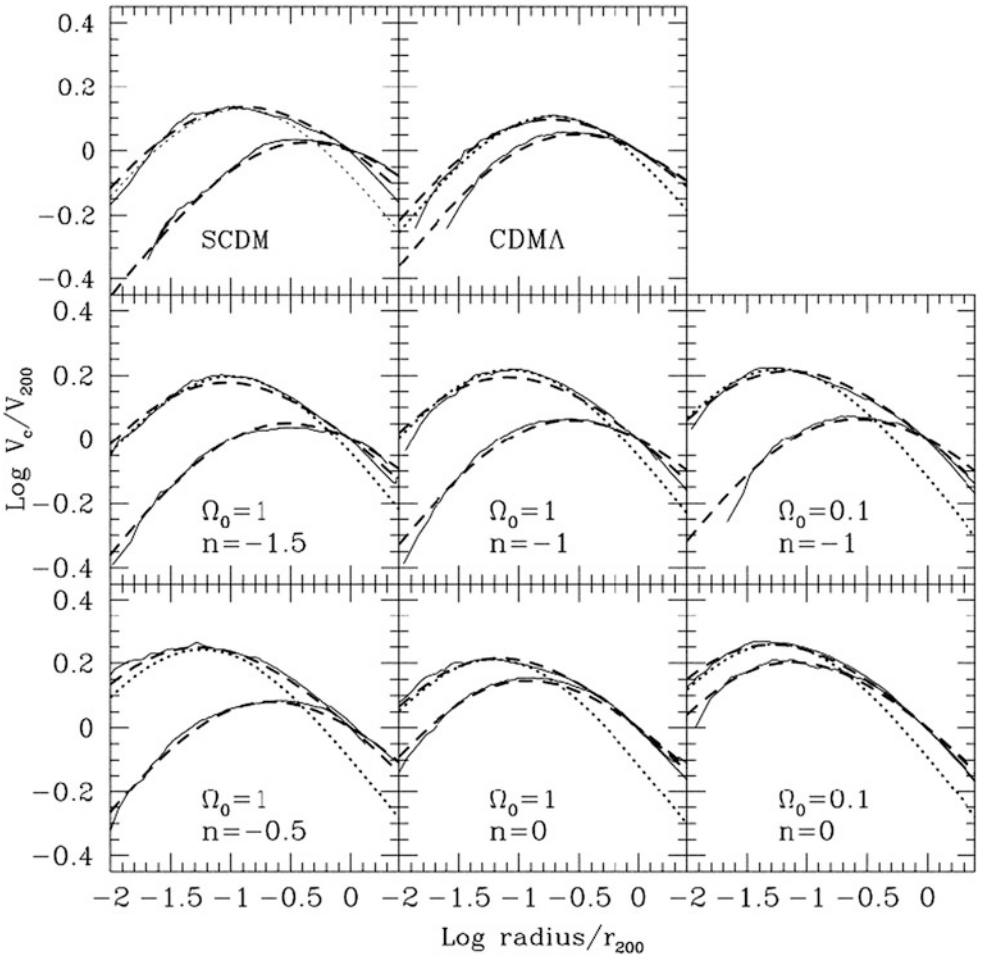


Fig. 7.19 The rotation curves in the NFW density profiles from Fig. 7.17, in units of the rotational velocity at r_{200}. All curves initially increase, reach a maximum, and then decrease again; over a fairly wide range in radius, the rotation curves are approximately flat. The *solid curves* are taken directly from the simulation, while *dashed curves* indicate the rotation curves expected from the NFW profile. The *dotted curve* in each panel presents a fit to the low-mass halo data with the so-called Hernquist profile, a mass distribution frequently used in modeling—it fits the rotation curve very well in the inner part of the halo, but fails beyond $\sim 0.1 r_{200}$. In these scaled units, halos of low mass have a relatively higher maximum rotational velocity. Source: J.F. Navarro et al. 1997, *A Universal Density Profile from Hierarchical Clustering*, ApJ 490, 493, p. 498, Fig. 4. ©AAS. Reproduced with permission

For galaxy clusters, the situation is more favorable, since the mass fraction of stars in them is smaller than in galaxies, and the stars are less concentrated towards the center. Indeed, it has been found that the X-ray data of many clusters are compatible with an NFW profile. Analyses based on the gravitational lensing effect also show that an NFW mass profile provides a very good description for the strong and weak lensing data; we will elaborate on this in Sect. 7.7 below. Additionally, Fig. 7.20 shows that the radial profile of the galaxy density in clusters on average follows an NFW profile, where the mean concentration index is $c \approx 3$, i.e., smaller than expected for the *mass* profile of clusters. One interpretation of this result is that the galaxy distribution in clusters is less strongly concentrated than the density of dark matter.

7.6.2 The shape and spin of halos

Halo shapes. Whereas the spherical collapse model made the simplifying assumption that the overdense regions are spherical, there is no reason for halos to have that symmetry. In fact, if one considers the peaks of a random density field, these maxima in general have different curvature along different directions. Correspondingly, the resulting halos are expected to deviate from sphericity.

Approximating the surfaces of constant density in a halo by an ellipsoid of semi-axes $a_1 \leq a_2 \leq a_3$, the shape of a halo is characterized by the ratios $s = a_1/a_3$ and $q = a_2/a_3$. If $s = q < 1$, the shape is similar to that of a cigar, and the ellipsoid is called 'prolate'. A halo with $s < 1$ and $q = 1$ has the shape of a hamburger, termed 'oblate'. In general, all three axes are different; such objects are called triaxial.

Numerical simulations show that the shape of dark matter halos depends strongly on their formation time and their merger history. If two halos of comparable mass collide and merge, the shape of the resulting halo will be strongly prolate. On the other hand, halos that form early and experience no such strong mergers tend to be more spherical.

Angular momentum of halos. During their evolution, dark matter halos can obtain a finite angular momentum, which can be measured from the output of simulations. The origin of this angular momentum can be traced back to tidal torques.

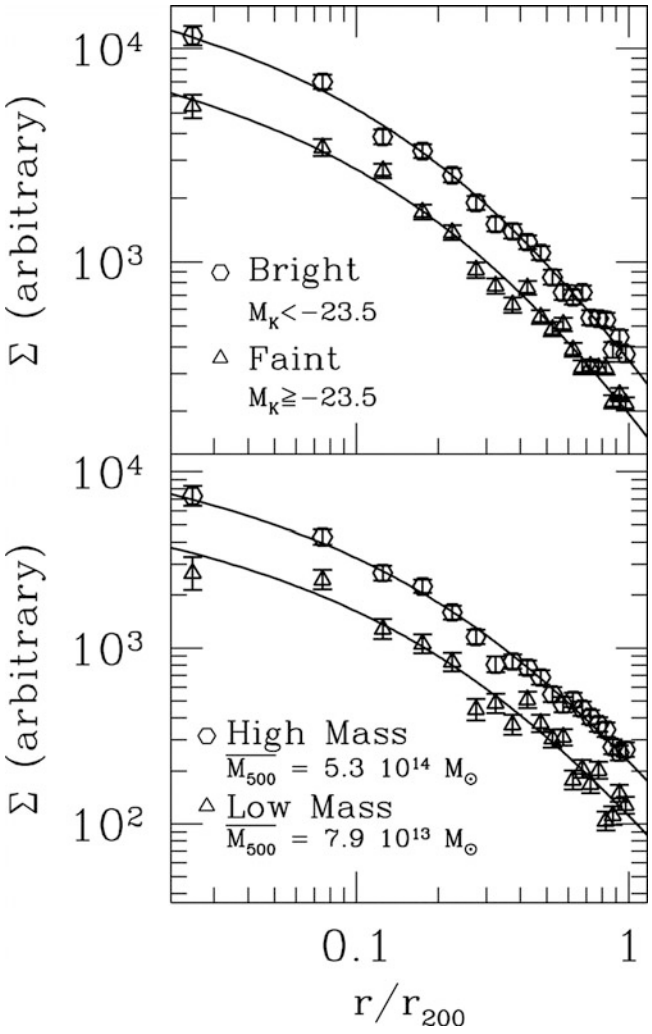


Fig. 7.20 The galaxy distribution averaged over 93 nearby clusters of galaxies, as a function of the projected distance to the cluster center. Galaxies were selected in the NIR, and cluster masses, and thus r_{200}, were determined from X-ray data. Plotted is the projected number density of cluster galaxies, averaged over the various clusters, versus the scaled radius r/r_{200}. In the *top panel* the galaxy sample is split into luminous and less luminous galaxies, while in the *bottom panel* the cluster sample is split according to the cluster mass. The *solid curves* show a fit of the projected NFW profile, which turns out to be an excellent description in all cases. The concentration index is, with $c \approx 3$, roughly the same in all cases, and somewhat smaller than expected for the mass profile of clusters. Source: Y.-T. Lin et al. 2004, *K-Band Properties of Galaxy Clusters and Groups: Luminosity Function, Radial Distribution, and Halo Occupation Number*, ApJ 610, 745, p. 756, Fig. 8. ©AAS. Reproduced with permission

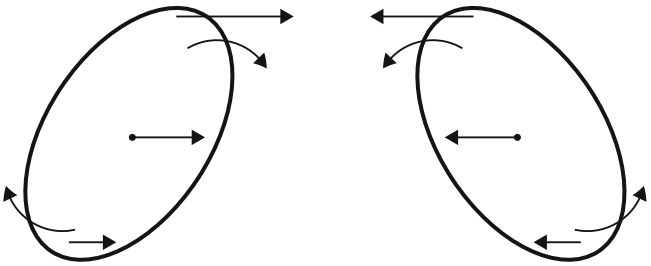


Fig. 7.21 Sketch of two non-spherical density concentrations and their mutual gravitational attraction. Points located closer to the neighboring overdensity feel a stronger force than the center of mass, leading to the tidal torque and a rotation of the body. This is how mass concentrations attain their angular momentum

As we have seen, halos are not spherical in general. If an ellipsoidal body is located in a tidal gravitational field, a torque acts on it, trying to align the body with the direction of the tidal field (see Fig. 7.21). This causes the body to rotate, yielding an angular momentum.

Usually, the angular momentum of a halo is quantified by the so-called spin parameter λ. To motivate its definition, consider a rigidly rotating sphere of radius R, angular velocity ω, and mass M. The sphere has a gravitational binding energy of

$$|E| \sim \frac{G\,M^2}{R}$$

and an angular momentum of

$$J \sim M\,R^2\,\omega\,,$$

with the constants of proportionality dependent on the density distribution inside the sphere. In order for the sphere to be rotationally supported, the gravitational acceleration on its surface should be balanced by the centrifugal acceleration, so that

$$\omega^2\,R \sim \frac{G\,M}{R^2}\,, \quad \text{or}$$

$$J \sim M\,R^2\,\omega \sim M\,R^2\sqrt{\frac{G\,M}{R^3}}$$

$$\sim M^{5/2}\,G\,|E|^{-1/2}\,,$$

where we have expressed R in terms of $|E|$. Hence, one defines the dimensionless *spin parameter*

$$\boxed{\lambda := \frac{J\,|E|^{1/2}}{G\,M^{5/2}}}\,. \tag{7.65}$$

For plausible density profiles, one finds that $\lambda \sim 0.4$ corresponds to rotational support.

The spin parameter of halos measured in simulations is typically an order of magnitude smaller than required for rotational support. This shows that the deviation from sphericity is *not* due to their rotation, but by the distribution of orbits of the dark matter particles. More quantitatively, one finds that the spin parameter of halos has a probability distribution of the form

$$p(\lambda)\,\mathrm{d}\lambda = \frac{1}{\sigma_\lambda\sqrt{2\pi}}\exp\left(-\frac{\ln^2(\lambda/\bar{\lambda})}{2\sigma_\lambda^2}\right)\frac{\mathrm{d}\lambda}{\lambda}\,,$$

with $\bar{\lambda} \sim 0.04$ and $\sigma_\lambda \sim 0.5$. Furthermore, there is the tendency that halos in denser environments have larger than average values of λ, as is expected from the above argument: on average, the tidal gravitational field is stronger in overdense regions.

In the early stages of halo formation, the baryons have about the same spatial distribution as that of the dark matter. Therefore, the gas in dark matter halos will attain a similar specific angular momentum as the halo itself. This angular momentum will turn out to be a key element for the formation of galaxies, as will be discussed in Chap. 10.

7.6.3 The bias of dark matter halos

Since dark matter halos host the observable tracers of the Universe, i.e., galaxies and groups and clusters of galaxies, it is interesting to study the properties of their spatial distribution and to relate this to the observable spatial distribution of galaxies and clusters. We shall consider this latter aspect in Sect. 8.1, but describe here the clustering properties of halos.

Halo number density contrast. We start by considering the density field in the universe, smoothed over spheres of radius R. As before, we call the smoothed density field $\delta_R(\boldsymbol{x})$. If R is large, any such sphere will contain many halos. Furthermore, we consider dark matter halos with mass between M and $M + \mathrm{d}M$. Their mean number density is $\bar{n} = \frac{\mathrm{d}n}{\mathrm{d}M}\mathrm{d}M$, as given by the halo abundance discussed before. Let the number of halos within a sphere of radius R centered on $\boldsymbol{x}$ be $N(\boldsymbol{x};M)\mathrm{d}M$; then we can define the local density of halos as $n(\boldsymbol{x};M)\mathrm{d}M = N(\boldsymbol{x};M)\mathrm{d}M/V$, where $V = (4\pi/3)R^3$ is the volume of the sphere. In analogy to the definition of the density contrast of matter, we can consider the relative density contrast of halos,

$$\delta_\mathrm{h}(\boldsymbol{x};M) = \frac{n(\boldsymbol{x};M) - \bar{n}(M)}{\bar{n}(M)}\,. \qquad (7.66)$$

We can now ask whether the two density fields, $\delta_R(\boldsymbol{x})$ and $\delta_\mathrm{h}(\boldsymbol{x};M)$ are similar. For example, suppose that $\delta_R(\boldsymbol{x}) = 1$ at one point, i.e., the matter density there is twice the cosmic mean. Does one expect to find also twice as many halos in the sphere surrounding $\boldsymbol{x}$ as one finds on average in a sphere of this radius?

Halo biasing. In general, one expects the number density of halos to be large in those regions of space where the matter density is also high. However, we have no good reason to assume that the number density of halos follows exactly the matter distribution. In fact, we can argue that in general, these two distributions should be different: Consider the density fluctuations of the matter to be divided into large- and small-scale fluctuations (see Fig. 7.22). The spherical collapse model predicts that a halo forms when the linear density contrast exceeds $\delta_\mathrm{c} = 1.69$. If we now use the decomposition $\delta(\boldsymbol{x}) = \delta_\mathrm{s}(\boldsymbol{x}) + \delta_\mathrm{l}(\boldsymbol{x})$, where the 's' and 'l' stand for small- and large-scale fluctuations, then in regions of positive δ_l, a halo can collapse even if δ_s is less than the threshold value δ_c, whereas in underdense regions with $\delta_\mathrm{l} < 0$, the small-scale fluctuations must reach a higher value than δ_c for the collapse to occur, namely $\delta_\mathrm{s} > \delta_c - \delta_\mathrm{l}$. In other words, in regions of overdense large-scale fluctuations, the smaller-scale fluctuations get a head-start for their gravitational collapse. As a matter of fact, this so-called peak-background split of the density fluctuation field can be used to analytically predict the relation between δ_h and δ.

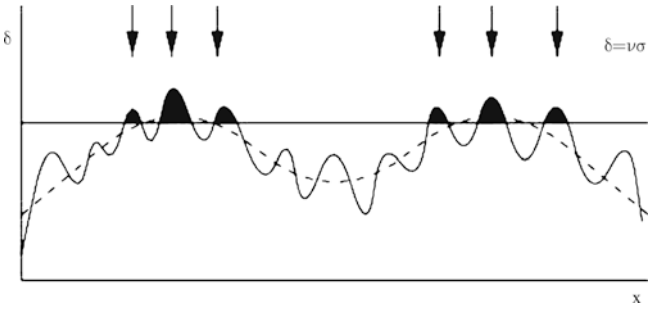


Fig. 7.22 The sketch represents a particular model of biasing. Let the one-dimensional density profile of matter be specified by the *solid curve*, which results from a superposition of a large-scale (represented by the *dashed curve*) and a small-scale fluctuation. Assuming that halos can form only at locations where the density field exceeds a certain threshold—plotted as a *straight line*—the halos in this density profile will be localized at the positions indicated by the *arrows*. Obviously, the locations of the halos are highly correlated; they only form near the peaks of the large-scale fluctuation. In this picture, the correlation of halos on small scales is much stronger than the correlation of the underlying density field. Source: J.A. Peacock 2003, *Large-scale surveys and cosmic structure*, astro-ph/0309240, Fig. 8

The fact that in general $\delta_\mathrm{h}(\boldsymbol{x};M) \neq \delta(\boldsymbol{x})$ is called halo biasing. The most simple way this biasing could be modeled is by assuming that these two density fluctuation fields are simply proportional to each other, i.e.,

$$\delta_\mathrm{h}(\boldsymbol{x};M) = b_\mathrm{h}(M)\,\delta(\boldsymbol{x})\,, \qquad (7.67)$$

where $b_\mathrm{h}(M)$ is called the *halo bias factor*. The ansatz (7.67) is called linear deterministic biasing. Whereas it cannot be valid in detail, this ansatz provides a useful description for large scales, i.e., when R is chosen sufficiently large. In particular, averaged over scales larger than $\sim 10h^{-1}\,\mathrm{Mpc}$, (7.67) is rather well satisfied, whereas on smaller scales, it ceases to be valid.

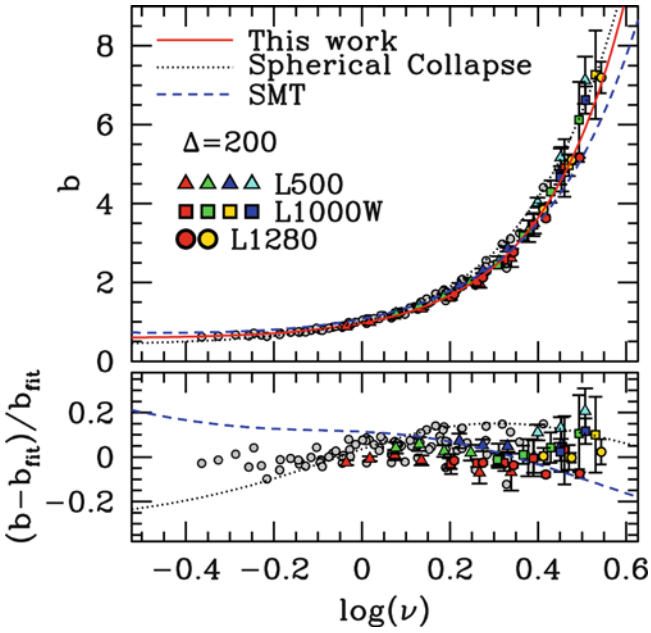


Fig. 7.23 *Top panel*: The large-scale bias of dark matter halos, as determined from a set of three different simulations (indicated by the three different types of symbols), as a function of the 'peak height' ν [see (7.51)]. For each type of symbol, different *colors* indicate different redshifts at which the bias was determined. The three *curves* show halo bias as obtained in the framework of Press–Schechter theory (*black dotted*) and the models of spheroidal collapse (*blue dashed*), as well as a fitting function to the data (*red solid*). The *bottom panel* shows the relative deviation of the measured halo bias from the fit, indicating its accuracy. Source: J.L. Tinker et al. 2010, *The Large-scale Bias of Dark Matter Halos: Numerical Calibration and Model Tests*, ApJ 724, 878, p. 880, Fig. 1. ©AAS. Reproduced with permission

Dependence on M and z. As mentioned before, using the peak-background split, the bias factor can be estimated analytically, and can also be determined from N-body simulations; these two estimates mutually agree rather well (see Fig. 7.23). One finds that $b_\mathrm{h}(M)$ is a monotonically increasing function of halo mass, with $b_\mathrm{h} < 1$ for $M \lesssim M^*(z)$ [or $\nu \lesssim 1$, where the 'peak height' ν is given in (7.51)] with $M^*(z)$ given by (7.53), and $b_\mathrm{h}(M) > 1$ for $M \gtrsim M^*(z)$ ($\nu \gtrsim 1$). In fact, as was the case for the halo abundance, the halo bias essentially depends solely on the value of ν. Towards low masses, the halo bias approaches a constant value ~ 0.7, and it rather steeply increases beyond M^*, reaching values of a few on the largest mass scales. Furthermore, at fixed mass, the halo bias increases with redshift, since according to (7.51), ν increases with redshift for fixed M.

These qualitative properties can be easily understood in terms of the peak-background split picture: the little head-start for the collapse of a peak in a region of large-scale overdensity is the more important, the rarer peaks of amplitude $> \delta_\mathrm{c}$ are. Hence, at rather high masses, this little additional push may actually be necessary for such halos to form at all. Conversely, for smaller masses, the abundance of peaks is high, and their number density with amplitude $\delta_\mathrm{c} + \epsilon$ is rather the same as that with $\delta_\mathrm{c} - \epsilon$. Hence, for them the impact of the large-scale fluctuation is negligible. As a general rule, the rarer density peaks are, the larger is the bias.

As an immediate consequence, (7.67) implies that the correlation function of halos is different from the correlation function of the matter distribution. Since the correlation function is quadratic in the density field, we readily find

$$\xi_\mathrm{h}(\boldsymbol{y}; M) := \langle \delta_\mathrm{h}(\boldsymbol{x})\, \delta_\mathrm{h}(\boldsymbol{x}+\boldsymbol{y})\rangle = b_\mathrm{h}^2(M)\, \xi(\boldsymbol{y})\,. \quad (7.68)$$

Thus, massive halos with $b_\mathrm{h} > 1$ are more strongly clustered than the underlying matter distribution, whereas low-mass halos are clustered less. One therefore expects that galaxies are less clustered than galaxy clusters; we will see later that this is indeed the case. A similar expression applies of course also to the power spectrum. Indeed, the halo bias in the simulations (Fig. 7.23) was measured from the ratio of the power spectra of halos and the overall matter distribution, restricted to the largest scales.

7.7 Weak gravitational lensing studies of dark matter halos

Whereas rotation curves probe the inner part of galaxy halos—typically out to a radius not larger than 1/10 of the virial radius—obtaining information for the mass profile at larger radii is difficult, due to the lack of luminous tracers. Weak gravitational lensing (see Sect. 6.6.2) offers the possibility to study the mass profile out to very large radii. In fact, it is by far the most powerful method for probing the outer parts of (galaxy-mass) halos. Combined with strong lensing at smaller radii, the halo mass profile of massive clusters can be studied over a broad range in radii (Sect. 7.7.1).

Apart from very massive clusters, the weak lensing signal of individual objects is not strong enough to be studied individually. The ellipticity dispersion of the faint galaxies which act as background sources for the lensing effect provides a noise component which is too large compared to the lensing signal. However, one can combine the weak lensing signal of a large number of lensing galaxies, and thus study their mean mass profiles. Provided the area of the imaging survey is large enough, the lens galaxies can be grouped into samples of similar properties (like redshift, color, luminosity, etc.), and thus the average mass profiles of galaxies can be investigated in dependence on these parameters.

In this section, we describe the basic method of this *galaxy-galaxy lensing (GGL)* technique and present some of the recent results (Sect. 7.7.2). We then discuss a method

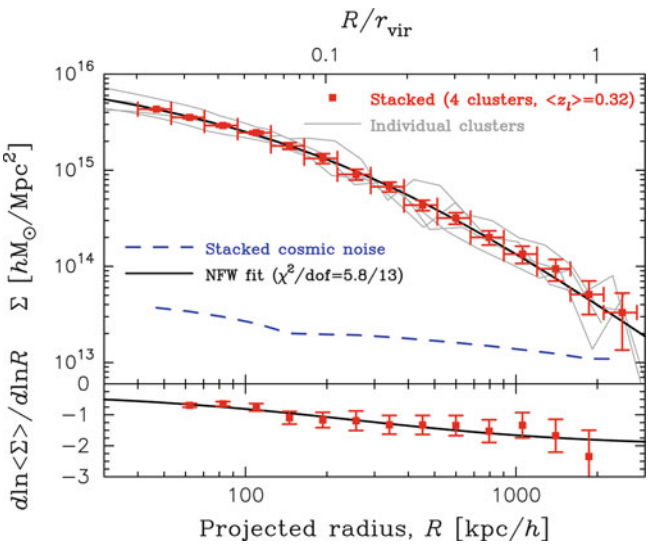


Fig. 7.24 The average surface mass density profile (*red points* with error bars) obtained from a strong and weak lensing analysis of four massive clusters of galaxies for which very high quality data are available. The four individual mass profiles are shown as *thin grey curves*. The *thick solid curve* is the best fitting line-of-sight projected NFW-profile to the mean mass profile; it provides an exceptionally good fit. The *blue dashed curve* shows the noise of the mean mass profile that is expected from the lensing effects of the large-scale structure between us and the clusters, and between the clusters and the source population. The *bottom panel* shows the slope of the surface mass density as a function of radius. There is a continuous steepening towards larger radii, again compatible with the NFW-profile, shown as solid curve. Source: K. Umetsu et al. 2011, *A Precise Cluster Mass Profile Averaged from the Highest-quality Lensing Data*, ApJ 738, 41, p. 6, Fig. 1. ©AAS. Reproduced with permission

to interpret them, introducing the so-called *halo model* in Sect. 7.7.3. Finally, we generalize this technique to the statistical study of the mass distribution of galaxy groups in Sect. 7.7.4.

7.7.1 Massive clusters

Very massive clusters show a sufficiently strong lensing signal for their mass distribution to be studied individually. In order to get information from a wide range in radii, the best results are obtained from combining strong lensing (multiple images and arcs) in the inner region with weak lensing for large radii.

Figure 7.24 shows results for four strong lensing clusters, supplemented by weak lensing information for large R. It is seen that the resulting mass profile is very well fit with an NFW-profile, a result also obtained by other studies. Therefore, it appears that lensing studies of clusters support the prediction of the CDM model for the existence of a universal mass profile.

Whereas the lensing data are well fit by the functional form of the NFW-profile, the resulting concentration parameters are found to be larger than the CDM prediction (7.62), even if the predicted spread of the c-M-relation is taken into account. It thus appears that strong lensing clusters are 'over-concentrated'. However, there are severe selection effects at work. First, the more concentrated a mass distribution is, the more likely it is that it produces giant arcs and multiple images (because the area over which the surface mass density exceeds the critical surface mass density for lensing is then larger). Thus, a strong lensing selection favors halos which have a higher-than-average concentration. Second, there are projection effects. The c-M-relation is obtained by considering the spherically-averaged mass distribution of halos. Since halos are triaxial in general, the concentration fitted to the projected mass profile will depend on the projection direction. Geometrically it is obvious that the largest concentration is obtained if the projection occurs along the direction of the largest axis, which then also maximizes the projected mass density, and hence the strong lensing strength. Thus, again, this leads to a selection effect for strong lensing clusters which biases the concentration to high values. Finally, the strong lensing probability increases substantially when a cluster is in the process of a merger; the resulting asymmetry of the mass distribution renders the occurrence of spectacular strong lensing features particularly likely. Indeed, we have seen in Sect. 6.6.1 that many of the famous giant arc clusters show a bimodal distribution. These clusters may therefore be extreme outliers in the c-M-relation. For these reasons, the 'over-concentration' is not regarded to be a serious issue for CDM models.

7.7.2 Galaxy-galaxy lensing

As we discussed in Sect. 6.6.2, the tidal component of the gravitational field causes a distortion of the observed shape of distant galaxies. This distortion is such that for axi-symmetric matter distributions, images are stretched in a direction tangent to the center of mass (see Fig. 6.53).

The shear. In the language of gravitation lensing, this distortion—or the tidal components of the deflection—is quantified by the *shear*. It is symbolized by the sticks in Fig. 6.53. The shear is linearly related to the surface mass density κ of the lens.

From the equations of gravitational lensing, one can show for an axi-symmetric mass distribution with dimensionless surface mass density $\kappa(\theta)$ that the shear $\gamma(\theta)$ is

$$\gamma(\theta) = \bar{\kappa}(\theta) - \kappa(\theta) \,, \tag{7.69}$$

where $\bar{\kappa}(\theta)$ is the mean surface mass density inside a circle of radius θ; it is related to the dimensionless mass $m(\theta)$ [see (3.70)] by $\bar{\kappa}(\theta) = m(\theta)/\theta^2$. Remarkably, the relation (7.69) is also valid for arbitrary mass distributions, if we interpret $\gamma(\theta)$ as the tangential component of the shear averaged over the circle of radius θ, and $\kappa(\theta)$ to be the mean surface mass density on that circle.

The principle of galaxy-galaxy lensing. The observed ellipticity of the image of a background source is the sum of the intrinsic ellipticity and the shear. Since the intrinsic orientation of galaxies has no preferred direction, the intrinsic ellipticity will have a mean of zero if we average over enough background galaxies.

Thus, consider a set of (foreground) galaxies, together with a population of (background) galaxies for which their ellipticities have been measured. If one then considers all foreground-background pairs within a small angular separation interval $\Delta\theta$ around θ, and measures the tangential component (relative to the center of the foreground galaxies) of the background ellipticities, the mean of this ellipticity provides an estimate of the mean shear of the foreground galaxies, since the intrinsic ellipticities of the background galaxies will average out to almost zero. In this way, the shear profile $\gamma(\theta)$ can be estimated. Since the shear is directly related to the mass density by (7.69), this shear profile determines the mass profile, up to an overall additive constant (which is related to the mass-sheet degeneracy; see problem 3.5).

If the redshifts of the foreground galaxies are individually known, the angular separation between foreground and background galaxy can be translated into a transverse separation, $R = D_{\rm d}\theta$. Furthermore, if the redshift distribution of the background galaxies are known as well, then the mean of the distance ratio $D_{\rm ds}/D_{\rm s}$ can be calculated. Therefore, with these two pieces of information, the critical surface mass density $\Sigma_{\rm cr}$ [see (3.67)] can be determined. Multiplying (7.69) by $\Sigma_{\rm cr}$ then yields

$$\Sigma_{\rm cr}\,\gamma(\theta) \equiv \Delta\Sigma(R) = \bar{\Sigma}(R) - \Sigma(R)\,, \qquad (7.70)$$

i.e., the observed shear profile can be directly related to the mean physical surface mass density inside the circle of radius R, minus the average surface mass density at radius R.

On small angular scales, the shear profile corresponds to the mass profile of the galaxy halos, and is well-fitted by either an SIS profile, or that of the universal halo density profile as described by the NFW model. Hence, GGL allows us to study the mean mass profiles of galaxy halos.

Selected results. In Fig. 7.25, we show the GGL signal $\gamma(\theta)$ as obtained from the RCS2 (the second Red Cluster Sequence) survey, where the bright foreground galaxies were identified by their spectroscopic redshifts as determined by SDSS. A clear signal is seen out to scales of ~ 1 deg. For the upper axis in this figure, the angular scale was converted to a transverse separation, using the mean distance of the lens galaxy sample. Thus we see that the measured shear signal probes the mean mass profiles of the foreground galaxies out to ~ 20 Mpc, i.e., far larger than the virial radius of galaxy-scale dark matter halos. For such large radii, we do not expect that the mass profile of galaxies is described by the universal mass profile (7.59) whose validity is restricted to within the virial radius.

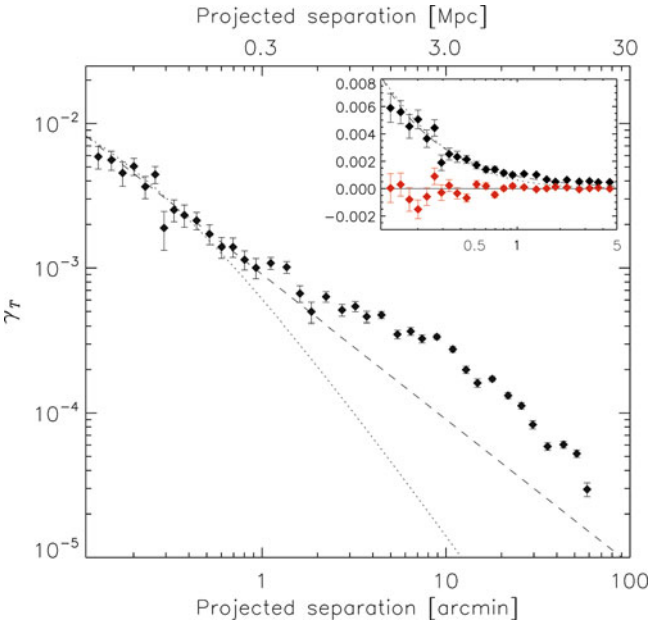


Fig. 7.25 The GGL signal, i.e., the mean tangential shear around foreground galaxies, as measured in the RCS2 survey, is shown by the *symbols* with error bars. The upper axis translates the angular scale into a transverse separation, for the mean distance of the foreground galaxy population. The *insert* displays a zoom of the central region, where the *red points* show the so-called cross component of the shear, whose average should be compatible with zero—which is seen to be the case. The two *curves* show the shear profile of the best-fitting SIS and NFW mass models. Source: E. van Uitert et al. 2011, *Galaxy-galaxy lensing constraints on the relation between baryons and dark matter in galaxies in the Red Sequence Cluster Survey 2*, A&A 534, A14, p. 6, Fig. 5. ©ESO. Reproduced with permission

The inner part of the GGL signal is fitted with an NFW profile (dotted curve) and an SIS model (dashed curve) in Fig. 7.25. In the inner ~ 200 kpc, both mass models yield good fits, but fall short of explaining the shear signal at larger radii. We will see below how the extended shear profile can be interpreted.

Nevertheless, on small scales, the mean shear profile of galaxies probes the mass profile of the galaxy and its dark matter halo. Selecting galaxy samples with different luminosity or stellar mass, the dependence of the parameters describing the NFW profile—in particular the virial mass—on the stellar mass can be studied.

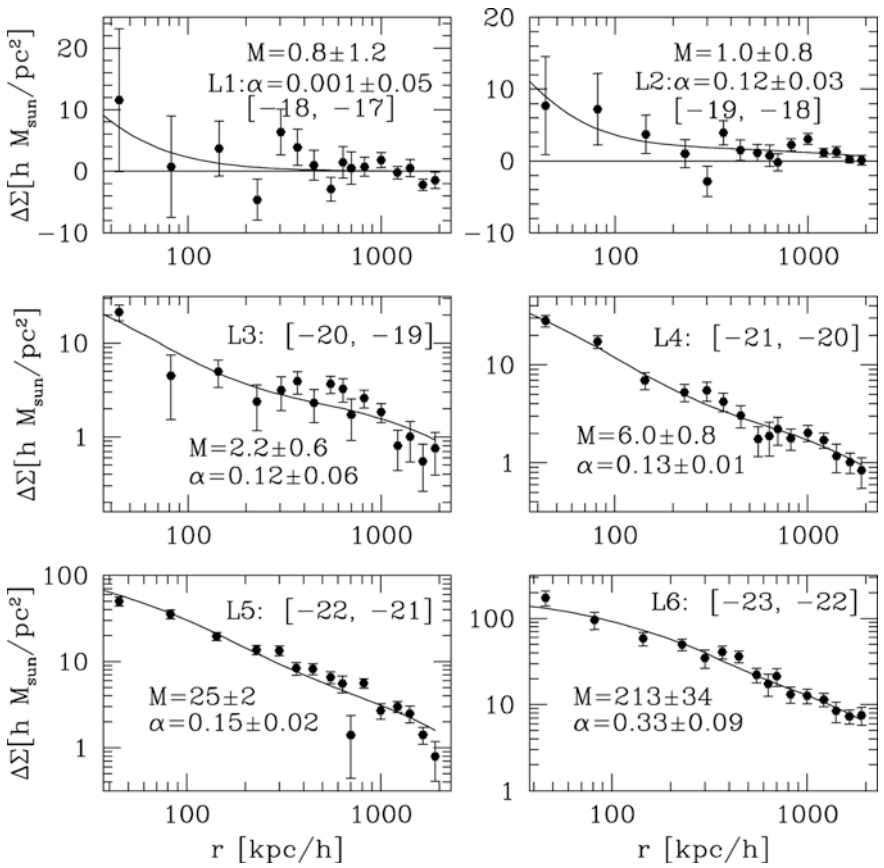


Fig. 7.26 The galaxy-galaxy lensing signal for six luminosity bins of foreground galaxies, as indicated by the absolute magnitude interval in each panel. More than 2.7×10^5 galaxies with spectroscopic redshifts were used as foreground galaxies in this analysis. The *curves* show a two-parameter model fitted to the data, based on the halo model, and the fit parameters are indicated: M is the virial mass of the halo (in units of $10^{11}h^{-1}M_\odot$) in which the galaxies reside, and α is the fraction of the galaxies which are not central inside the halo, but satellite galaxies. Source: U. Seljak et al. 2005, *Cosmological parameter analysis including SDSS Lyα forest and galaxy bias: Constraints on the primordial spectrum of fluctuations, neutrino mass, and dark energy*, Phys. Rev. D 71, 043511, Fig. 1. http://journals.aps.org./prd/abstract/10.1103/PhysRevD.71.043511. Published with kind permission ©APS 2005. All Rights Reserved

The combination of imaging and spectroscopy in the Sloan Digital Sky Survey makes this an ideal data set for studying GGL, since one can select foreground galaxies with known redshifts. Figure 7.26 shows the GGL signal $\Delta\Sigma(R)$ for six different bins of absolute magnitude. The shear signal is clearly detected out to large radii for the more luminous galaxy samples. The signal increases with luminosity, showing that the galaxy+halo mass is a monotonic function of galaxy luminosity, as expected. As was the case for the RCS2 results shown in Fig. 7.25, the GGL signal from the SDSS shown in Fig. 7.26 extends to separations much larger than the expected virial radius for galaxy-mass halos.

7.7.3 Interpretation: The halo model

The shortfall of the NFW mass profile to explain the observed GGL signal on scales larger than the virial radius can be understood from noting that galaxies (and their halos) are not isolated. We have seen in Sect. 7.6.3 that dark matter halos are correlated. Furthermore, many galaxies are members of galaxy groups or clusters.

Therefore, the mean mass profile around galaxies will be a superposition of several components. On small scales, it is dominated by the stellar mass of the galaxy and the dark matter halo in which it is embedded. On intermediate scales, one then starts to see the impact of the groups and clusters in which a certain fraction of the galaxies are embedded. On even larger scales, say $\gtrsim$1 Mpc, which exceed the size of most galaxy clusters, the signal becomes increasingly dominated by the mass from dark matter halos which are correlated with the host halo of the galaxy. Disentangling the various contributions of the GGL signal has to be done in the framework of a model.

Ingredients of the halo model. A very successful framework for the interpretation of the GGL signal is the halo model, which shall be briefly sketched here. It assumes that the all mass in the universe is contained in dark matter halos, so that the density distribution can be written as a sum of the density profiles of these halos,

$$\rho(\boldsymbol{x}) = \sum_i \rho_{\rm h}(|\boldsymbol{x} - \boldsymbol{x}_i|; M_i) \ , \qquad (7.71)$$

where the sum extends over all dark matter halos (in a given volume of space), $\boldsymbol{x}_i$ and M_i is the position and mass of the i-th halo, and $\rho_{\rm h}$ is the halo mass profile, which is assumed to be spherically symmetric (so that the mass contribution of the i-th halo at the location $\boldsymbol{x}$ depends only on the separation $|\boldsymbol{x} - \boldsymbol{x}_i|$ between the point $\boldsymbol{x}$ and the halo center $\boldsymbol{x}_i$). Fur-

thermore, by writing (7.71) we have assumed that the density profile of a halo is fully characterized by its mass M_i—e.g., that the density profile ρ_h is given by the NFW-profile with a concentration determined by its mass [see (7.62)]. One can account for the dispersion of the concentration parameter about its mean (7.62) by using c as a further argument of ρ_h. The halo population is characterized by the halo mass spectrum (Sect. 7.5.2) which, together with the halo correlation function (7.68) is obtained from numerical simulations.

The basic assumption of the halo model, namely that all the mass in the Universe is contained in halos, is fairly well supported by cosmological simulations. Down to a mass scale of $2 \times 10^8\, M_\odot$ (the scale that is resolved by the Millennium-II simulation), about 60 % of the total mass of the present Universe is contained in halos. Extrapolating to even lower halo mass, using analytic models (see Sect. 7.5.2), suggests that this fraction rises to some 80 % down to the smallest halo masses. Whereas the mass fraction contained in halos is lower at higher redshifts, these results provide a good motivation for the halo model.

The mass correlation function. Combining these ingredients, the correlation function of matter can be derived. If we want to calculate the correlator $\langle \rho(\boldsymbol{x})\, \rho(\boldsymbol{x}')\rangle$, we obtain a double sum,

$$\left\langle \rho(\boldsymbol{x})\, \rho(\boldsymbol{x}')\right\rangle = \sum_{ij} \left\langle \rho_\mathrm{h}(|\boldsymbol{x} - \boldsymbol{x}_i|; M_i)\, \rho_\mathrm{h}(|\boldsymbol{x}' - \boldsymbol{x}_j|; M_j)\right\rangle \,. \tag{7.72}$$

This double sum can then be split into a diagonal term (i.e., where $i = j$) and a non-diagonal one. For the former, the density of the halo is correlated with itself—we call this the one-halo term of the mass correlation function. The non-diagonal term correlates the mass in one halo with that of another halo, giving rise to the two-halo term in the correlation function. This latter term arises from the fact that the halo centers are correlated, according to (7.68).

Averaging the result (7.72) over the mass spectrum of halos, as well as over the probability distribution of halo positions $\boldsymbol{x}_i$, accounting for their mass-dependent correlation, one obtains the two-point correlation function of the matter distribution as predicted by the halo model. Indeed, the halo model yields a description of the matter distribution which appears to be an astonishingly good approximation to the more accurate results from simulations. As argued before, the matter correlation function is a sum of two terms, $\xi_\mathrm{m} = \xi_\mathrm{m}^\mathrm{1h} + \xi_\mathrm{m}^\mathrm{2h}$, where the first term describes matter correlations within the same halo, whereas the second is the correlation between two different halos.

Inclusion of galaxies. Galaxies are next introduced into this halo model. This is done by first noting that galaxies form at the center of dark matter halos. Satellite galaxies in halos, such as galaxies in clusters or the dwarf galaxies in the Milky Way, presumably also formed at the center of dark matter halos which subsequently merged with the larger halo.

Hence, one considers a population of galaxies with luminosities (or stellar masses) in a given interval. Let $\langle N|M\rangle$ be the mean number of galaxies (with the prescribed properties) that live in a halo of mass M. This function cannot be obtained from first principles, but can be constrained by observations of galaxies in groups and clusters whose masses are estimated by X-ray or weak lensing observations; additional constraints come from the luminosity function of galaxies. It is usually assumed that the number N of galaxies is Poisson distributed, with mean $\langle N|M\rangle$. Frequently, the mean number of galaxies is prescribed by a power law in mass, $\langle N|M\rangle \propto M^\epsilon$, above some mass threshold which depends on the luminosity (or stellar mass) of the galaxies under consideration. For lower-mass halos, $\langle N|M\rangle$ is assumed to decline rapidly—low-mass halos do not host luminous galaxies (e.g., one does not expect to find any L^* galaxy in a halo of mass $< 10^{11}\, M_\odot$).

If there is only one galaxy in a halo, it is assumed to have formed at the center, whereas if $N > 1$, one assumes that one of them lies at the center (the central galaxy), whereas the rest (satellite galaxies) are distributed according to a specific radial distribution function; typically, they are assumed to also follow an NFW profile, as motivated by the radial distribution of galaxies in clusters (Fig. 7.20).

The halo model as sketched above predicts the two-point correlation functions of matter, galaxies, and the cross-correlation between matter and galaxies. Whereas this prescription of the distribution of matter and galaxies contains a number of free functions—such as $\langle N|M\rangle$—these are well constrained by comparing the predicted abundance of galaxies with observations. Most of the properties of the halo model can be summarized in the fraction of galaxies α which are satellite galaxies, i.e., a fraction $(1-\alpha)$ of all galaxies are central galaxies.

Application to galaxy-galaxy lensing. Since galaxy-galaxy lensing measures the mean shear around galaxies, and since the shear is directly related to the mass distribution [see (7.70)], GGL measures the correlation between galaxy positions and the matter distribution. Therefore, the expected GGL signal can be obtained from the halo model as described above.

In Fig. 7.27, the GGL signal as predicted by the halo model is illustrated, for a halo of fixed virial mass M_{200}. The signal is a superposition of various terms. First, the baryons of the galaxies under consideration produce a lensing signal on small scales, due to their compactness. Second, central galaxies and satellites both contribute to the signal, with their respective abundance fractions $(1-\alpha)$ and α, respectively.

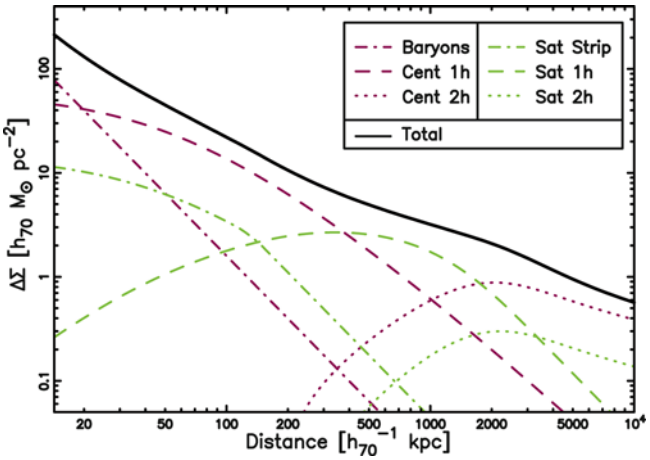


Fig. 7.27 Illustration of the GGL signal in terms of the halo model. Here, a halo of virial mass $M_{200} = 10^{12} M_\odot$ is chosen, with a stellar mass of $M_* = 5 \times 10^{10} M_\odot$. Furthermore, it is assumed that a fraction $\alpha = 0.2$ of all galaxies are satellite galaxies. The GGL signal is then the sum $\Delta\Sigma = \Delta\Sigma_{\rm bar} + (1-\alpha)\Delta\Sigma_{\rm cent} + \alpha\Delta\Sigma_{\rm sat}$ of three terms: That from the stellar mass of the galaxies ($\Delta\Sigma_{\rm bar}$), the signal around central galaxies ($\Delta\Sigma_{\rm cent}$), weighted by their fraction $(1-\alpha)$, and the signal around satellite galaxies ($\Delta\Sigma_{\rm sat}$), weighted by their fraction α. Each of these two latter terms is composed of the signal generated by the matter in the same halo where the galaxies are located (the one-halo term), and that of the neighboring halos (the two-halo term). Finally, one assumes that the satellite galaxies have their own dark matter subhalo, which is a (tidally) stripped version of its original host halo. Hence one writes $\Delta\Sigma_{\rm cent} = \Delta\Sigma^{\rm 1h}_{\rm cent} + \Delta\Sigma^{\rm 2h}_{\rm cent}$ and $\Delta\Sigma_{\rm sat} = \Delta\Sigma^{\rm strip}_{\rm sat} + \Delta\Sigma^{\rm 1h}_{\rm sat} + \Delta\Sigma^{\rm 2h}_{\rm sat}$. The various *curves* in the figure represent these different contributions, as labeled. Source: M. Velander et al. 2013, *CFHTLenS: The relation between galaxy dark matter haloes and baryons from weak gravitational lensing*, arXiv:1304.4265, p. 6, Fig. 3. Reproduced by permission of the author

For both of them, the mass correlated with the galaxy can reside in the same halo as the galaxy, giving rise to the 1-halo term, or in a different halo. Finally, it was assumed for Fig. 7.27 that the satellite galaxies have retained their own dark matter (sub-)halo, though with only half the mass an isolated galaxy of the same luminosity would have; this reduction of halo mass is a natural consequence of tidal stripping.

We see that on small scales, the baryons of the galaxies and the 1-halo term of central galaxies totally dominate the lensing signal. Hence, on scales smaller than about the virial radius of the halo (which is about 200 kpc for the halo mass assumed in the figure), the GGL signal indeed probes the radial mass profile of the galaxy+dark matter halo. Beyond the virial radius, the 1-halo term of satellite galaxies becomes stronger and then starts to dominate the signal. For scales beyond ~ 1 Mpc, the signal becomes increasingly dominated by other halos which are correlated with the host halo, i.e., the 2-halo terms of central galaxies and satellites becomes the dominant signal. The transition between these various regimes gives the total GGL signal its characteristic shape, as seen by the black solid curve in Fig. 7.27, which is also seen in Figs. 7.26 and 7.25.

Studying the GGL results as a function of luminosity or stellar mass, and separately for red and blue galaxies, one finds a number of interesting results. First, for fixed luminosity, the signal is considerably larger for red galaxies than for blue ones. This is particularly true at large separations which reflects the clustering properties of the halos in which the different galaxy types are embedded. The halo model yields a satisfactory fit to the data, as shown in Fig. 7.26. From these fits, the halo mass M_{200} as a function of luminosity or stellar mass M_* can be obtained. It is found that the $M_{200}(M_*)$-relation is steeper for red galaxies. When fitted with a power law, one finds $M_{200} \propto M_*^{\sim 1.5}$ for red galaxies, whereas the slope is flatter than unity for blue galaxies. Furthermore, the satellite fraction is higher for red galaxies, approaching unity for $M_* \lesssim 5 \times 10^9 M_\odot$ (i.e., low-mass red galaxies are almost never central galaxies in halos), and decreasing to $\alpha \sim 0.2$ for higher-mass red galaxies. In contrast, almost all blue galaxies are centrals.

Studying the relation between halo and stellar mass further, one finds that the two are not simply related by a power law: The ratio M_{200}/M_* is not a monotonic function of M_*, but reaches a minimum at $M_* \sim 5 \times 10^{10} M_\odot$, as shown in Fig. 7.28. This result implies that the efficiency with which gas is transformed into stars in a halo is a function of halo mass, and that there is a preferred mass scale for maximum star-formation efficiency. We will discuss this result and its implications in much more detail in Chap. 10.

7.7.4 Masses of groups and clusters

In a similar manner as done for galaxies, one can also superpose the weak lensing signal of galaxy groups and clusters in order to determine their mean mass profiles, as a function of some observable, such as the optical luminosity of the group or its richness (i.e., number of bright group members). The large number of redshifts obtained with the SDSS allowed the construction of group catalogs, based on the spatial (i.e., 3D) overdensity of galaxies; in particular, we described the properties of the maxBCG group catalog in Sect. 6.2.4.

For this group sample, the galaxy-group signal $\Delta\Sigma$ as a function of separation is shown in Fig. 7.29, separately for different richness bins of these groups. The first point to note is that the lensing signal, and thus the mass, increases with increasing richness—more massive groups contain more luminous galaxies on average.

The interpretation of the lensing signal is again performed in the framework of the halo model, and the differently colored curves in Fig. 7.29 correspond to the various contributions. The NFW-profile of the dark matter halo

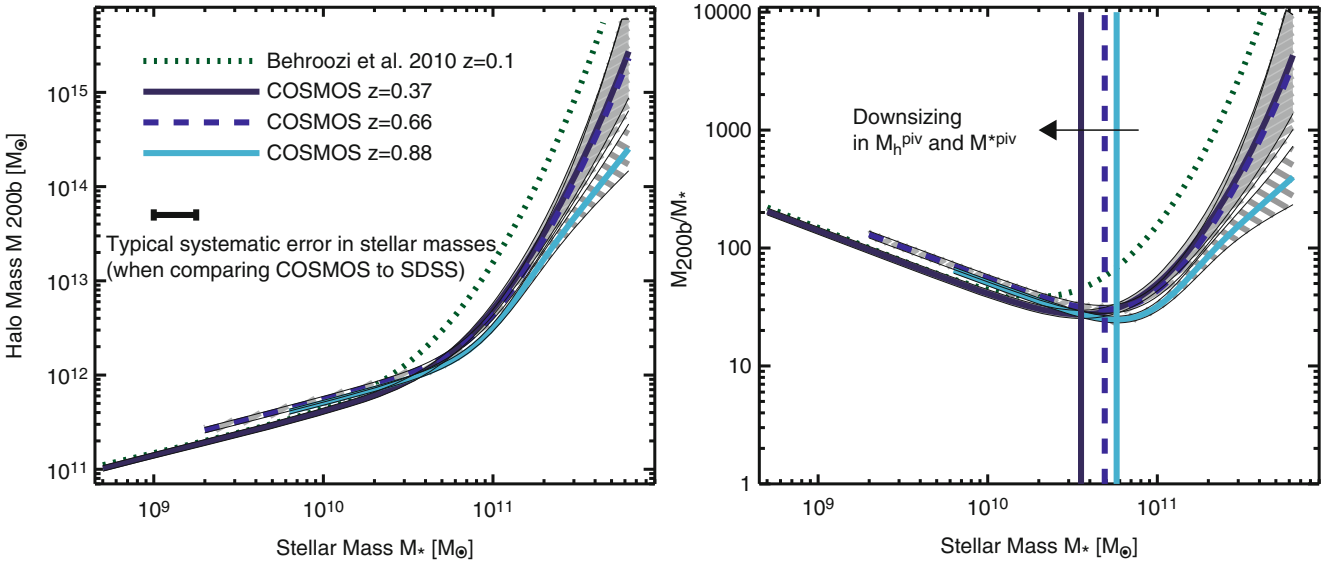


Fig. 7.28 From the analysis of the GGL signal in the COSMOS data field, the relation between stellar and halo mass can be studied over a wide range of galaxy masses and redshifts. In the *left panel*, the halo mass is plotted as a function of the stellar mass, as obtained by fitting the GGL signal with the halo model, for three different redshift intervals (where the mean redshift is indicated by line-type), and the *shaded area* around the curves indicates the estimated uncertainty. For comparison, the *dotted curve* shows the same relation obtained by matching the abundance of low-redshift galaxies with the halo mass function. The relative calibration between the two different methods is uncertain by an amount indicated by the error bar. One finds that the functional form of the $M_{200}(M_*)$-relation exhibits a characteristic change of slope, steepening above $\sim 5 \times 10^{10} M_\odot$. In the *right panel*, the same result is shown, except that now the ratio M_{200}/M_* is plotted. This ratio has a minimum at the mass scale where the $M_{200}(M_*)$-relation steepens. Hence, there is a characteristic mass scale at which the stellar contents of a halo is maximized. As we will discuss in Chap. 10, this corresponds to a mass scale of halos where the conversion of baryons into stars is maximally efficient. This characteristic mass scale seems to decrease with redshift, an effect sometimes called 'downsizing'. Source: A. Leauthau et al. 2012, *New Constraints on the Evolution of the Stellar-to-dark Matter Connection: A Combined Analysis of Galaxy-Galaxy Lensing, Clustering, and Stellar Mass Functions from* $z = 0.2$ *to* $z = 1$, ApJ 744, 159, p. 17, Fig. 11. ©AAS. Reproduced with permission

of the groups is shown as green curves, whereas the red curves show the lensing signal from the baryonic component of the brightest cluster galaxy (BCG), assumed to represent the center of the group halo. However, not in every case is the BCG correctly identified as the group center. For a fraction of groups, the BCG is displaced from the center (or may even be misidentified). This fraction was estimated from simulations, and amounts to some 40 % for groups with small richness, decreasing to $\sim$ 20 % for more massive groups and clusters. The corresponding correction to the lensing signal is shown as orange curves. Finally, the 2-halo term dominates the signal on the largest scales. The sum of these contributions is shown by the violet curves, which provide a very good fit to the lensing data.

From the model fit of the stacked group/cluster lensing signal shown in Fig. 7.29, one can derive the halo mass M_{200} as a function of the richness or the total optical luminosity of the groups. The results are shown in Fig. 7.30. The halo mass increases monotonically with richness and luminosity, approximately as $M_{200} \propto N_{200}^{1.28}$ and $M_{200} \propto L_{200}^{1.22}$. In particular we note that the latter relation implies that the mass-to-light ratio of clusters increases with mass, in agreement with the results obtained from GGL in the COSMOS field (see Fig. 7.28). Furthermore, it is found that the mass of the BCG increases with the halo mass, for low-mass groups, but then saturates for large group masses.

Summarizing this section, the statistical (or stacked) weak lensing signal of galaxies and groups provides the most direct way to study the relation between their observed properties and their mass properties. The NFW mass profile yields satisfactory fits to the lensing data when used in the context of the halo model, which provides a convenient framework of parametrizing the distribution of mass and galaxies in the Universe.

7.8 The substructure of halos

Sub-halos of galaxies and clusters of galaxies. Numerical simulations of structure formation in the CDM model show that the mass density in halos is not smooth; instead, they reveal that halos contain numerous halos of much lower mass, so-called sub-halos. For instance, a halo with the mass of a galaxy cluster contains hundreds or even thousands of halos with masses that are orders of magnitude lower. These sub-halos are indeed observed, since the substructure in clusters is visible—in the form of the cluster galaxies. In

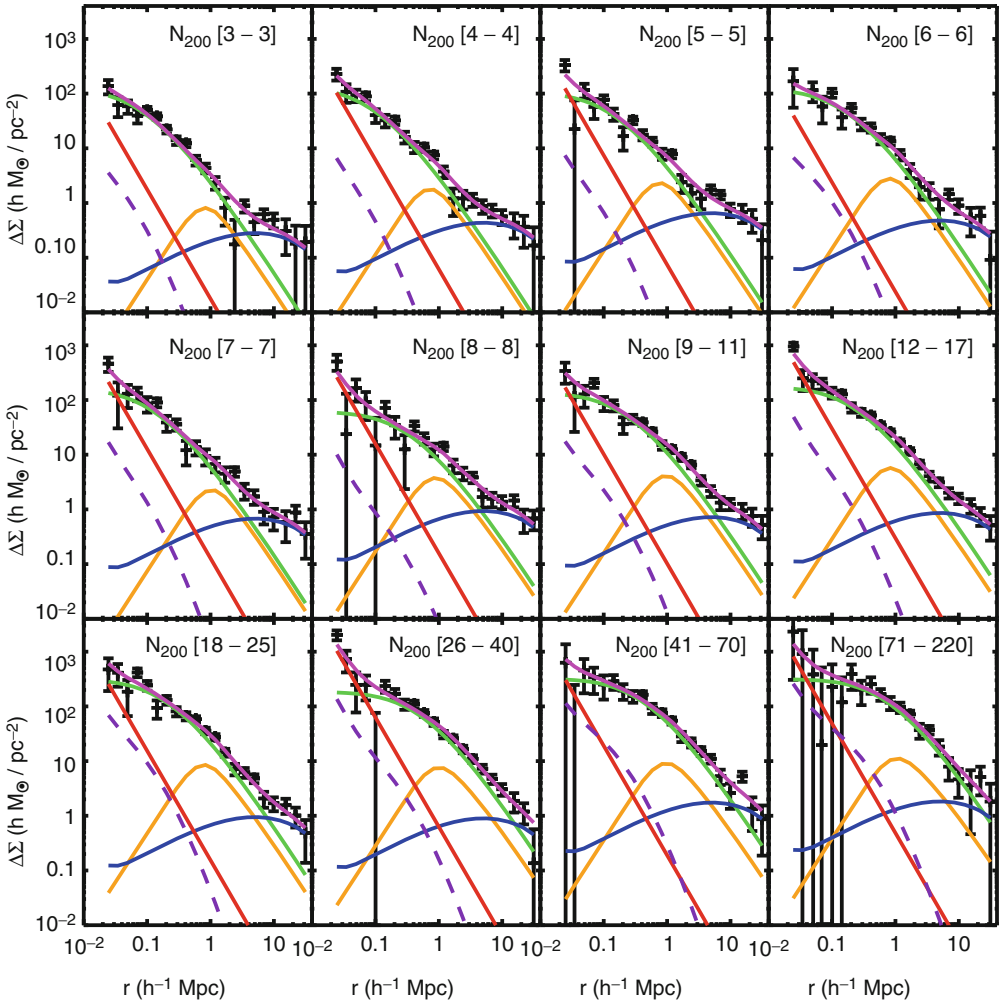


Fig. 7.29 The galaxy-group lensing signal for the maxBCG sample of groups/clusters, where different *panels* correspond to different richness bins. The signal is modeled by a number of different contributions: *Green* shows the NFW signal of the dark matter halo, *orange* is the contribution from miscentering the halo, *red* the signal from the brightest cluster galaxy, and *blue* the contribution from neighboring halos. Source: D.E. Johnston et al. 2007, *Cross-correlation Weak Lensing of SDSS galaxy Clusters II: Cluster Density Profiles and the Mass–Richness Relation*, arXiv:0709.1159, Fig. 8. Reproduced by permission of the author

the upper part of Fig. 7.31, the simulation of a cluster and its substructure is displayed. In fact, this mass distribution looks just like the mass distribution expected in a cluster of galaxies, with the main cluster halo and its distribution of member galaxies. The lower part of Fig. 7.31 shows the simulation of a halo with mass $\sim 2 \times 10^{12} M_\odot$, which corresponds to a massive galaxy. As one can easily see, its mass distribution shows a large number of sub-halos as well. In fact, the two mass distributions are nearly indistinguishable, except for their scaling in the total mass.[12] The presence of substructure over a very wide range in mass is a direct consequence of hierarchical structure formation, in which objects of higher mass each contain smaller structures that have been formed earlier in the cosmic evolution.

Such simulations show that of order $\sim 10\,\%$ of the mass of halos is contained in sub-halos, with a fraction that is slightly smaller for galaxy-mass halos ($\sim 7\,\%$) than for cluster-mass halos. Furthermore, the mass spectrum of the sub-halos follows a simple power law, $n(M) \propto M^{-1.9}$, down to the smallest mass-scale that can be resolved by simulations (which is currently about 10^{-7} times the mass of the parent halo). In fact, the very small velocity dispersion of cold dark matter particles predicts that this mass spectrum should continue down to the smallest halo masses that can be formed by CDM—which is about an Earth mass, or $10^{-6} M_\odot$.

[12] The reason for this is found in the property of the power spectrum of density fluctuations that has been discussed in Sect. 7.5.2, namely that $P(k)$ can be approximated by a power law over a wide range in k. Such a power law features no characteristic scale. For this reason, the properties of halos of high and low mass are scale-invariant, as is clearly visible in Fig. 7.31.

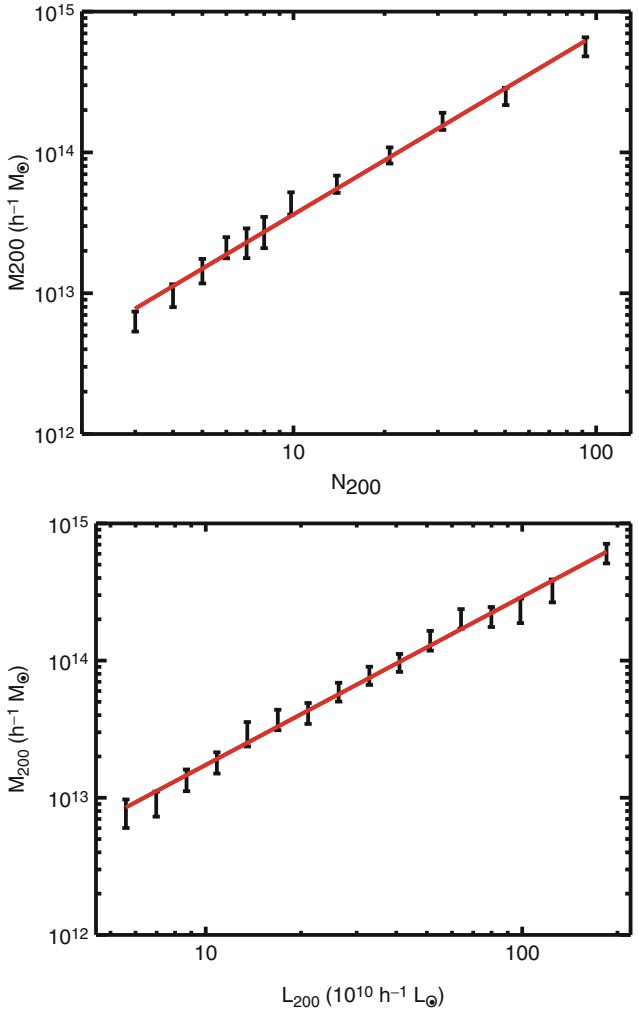


Fig. 7.30 Cluster mass as a function of optical richness (*top*) and cluster luminosity (*bottom*). For the definition of richness and luminosity, the number of red galaxies brighter than $0.4L^*$ and within a projected separation less than $1\,h^{-1}$Mpc from the brightest cluster galaxy (BCG), N_g, is used to define the radius $r_g = 0.156 N_g^{0.6} h^{-1}$ Mpc. This was previously found to be the radius within which the mean luminosity density is 200 times the average cosmic luminosity density of galaxies, which is determined from the galaxy luminosity function. The number of red galaxies within r_g is then defined as N_{200}; likewise, the sum of the luminosities of all red galaxies within r_g is defined as L_{200}. Binning the groups and clusters according to their richness and luminosity, and separately analyzing their weak lensing signal yielded the mass-richness and mass-luminosity relations shown. Note that here also very poor groups are included; some of them consisting of just two or a few galaxies. However, there are still more than 13 000 clusters with $N_{200} \geq 10$. Clusters are restricted to the redshift range $0.1 \leq z \leq 0.3$. Source: D.E. Johnston et al. 2007, *Cross-correlation Weak Lensing of SDSS galaxy Clusters II: Cluster Density Profiles and the Mass–Richness Relation*, arXiv:0709.1159, Fig. 11. Reproduced by permission of the author

Fig. 7.31 Density distribution of two simulated dark matter halos. In the *top image*, the halo has a virial mass of $5 \times 10^{14} M_\odot$, corresponding to a cluster of galaxies. The halo in the *bottom image* has a mass of $2 \times 10^{12} M_\odot$, representing a massive galaxy. In both cases, the presence of substructure in the mass distribution can be seen. It can be identified with individual cluster galaxies in the case of the galaxy cluster. The substructure in a galaxy can not be identified easily with any observable source population; one may expect that these are satellite galaxies, but observations show that these are considerably less abundant than the substructure seen here. Apart from the length-scale (and thus also the mass-scale), both halos appear very similar from a qualitative point of view. Source: B. Moore et al. 1999, *Dark Matter Substructure within Galactic Halos*, ApJ 524, L19, p. L20, Fig. 1. ©AAS. Reproduced with permission

The spatial distribution of the sub-halos is less concentrated towards the halo center than the total mass. The reason for this property lies in the fact that sub-halos whose orbits bring them deep into the potential well of the host halo are subject to strong tidal forces, and they get disrupted in the course of evolution. Simulations which include gas physics (we will describe some of these simulations in Sect. 10.6.1 below) find that the disruption becomes weaker if baryons are included—their dissipational nature leads to more compact, and thus more tightly bound, sub-halos; hence, they can resist the disruptive tidal forces for a longer time.

The 'substructure problem'. As we will discuss in detail in the next chapter, the CDM model of cosmology has proven to be enormously successful in describing and predicting cosmological observations. Because this model has achieved this success and is therefore considered the standard model, results that apparently do not fit into this standard model are of particular interest. The rotation curves of LSB galaxies mentioned above are one such result. Either one finds a good explanation for this apparent discrepancy between observation and the predictions of the CDM model—such as we indicated above—or, otherwise, results of this kind may necessitate to introduce extensions to the CDM model. In the former case, the model would have overcome another hurdle in demonstrating its consistency with observations and would be strengthened even further, whereas in the latter case, new insights would be gained into the physics of cosmology. Besides the rotation curves of dwarf and LSB galaxies, there is another observation that does not seem to fit into the picture of the CDM model at first sight.

Whereas the substructure in clusters is easily identified with the cluster member galaxies, the question arises as to what the sub-halos in galaxy-mass halos can possibly correspond to. The mass spectrum of these halos, as obtained from an early simulation, is displayed in Fig. 7.32. Some of these sub-halos are recognized in our Milky Way, namely the known satellite galaxies like, e.g., the Magellanic Clouds. In a similar way, the satellite galaxies of the Andromeda galaxy may also be identified with sub-halos. However, as we have seen in Sect. 6.1, not more than 40 members of the Local Group were known before 2003—whereas the numerical simulations predict hundreds of satellite galaxies for the Galaxy. This apparent deficit in the number of observed sub-halos, clearly indicated in Fig. 7.32, is considered to be another potential problem of CDM models.

We note that since this discrepancy was first explicitly pointed out, some 25 new dwarf galaxies in the Local Group have been detected from the Sloan Digital Sky Survey. Given that the SDSS observed only a quarter of the sky, one expects to have at least another $\sim$ 70 satellite galaxies in the Local Group which are not contained in the footprint of the SDSS. Therefore, the difference between the number of satellite galaxies and the predictions from numerical simulations has become smaller in recent years. On the other hand, these newly discovered satellites are all of very low mass, and thus do not fully fill in the gap between the curves in Fig. 7.32.

However, one always needs to remember that the dark matter simulations only predict the distribution of mass, and not that of light (which is accessible to observation). One possibility of resolving this apparent discrepancy centers on the interpretation that these sub-halos do in fact exist, but that most of them do not, or only weakly, emit radiation. What appears as a cheap excuse at first sight is indeed already part of the models of the formation and evolution of galaxies. As

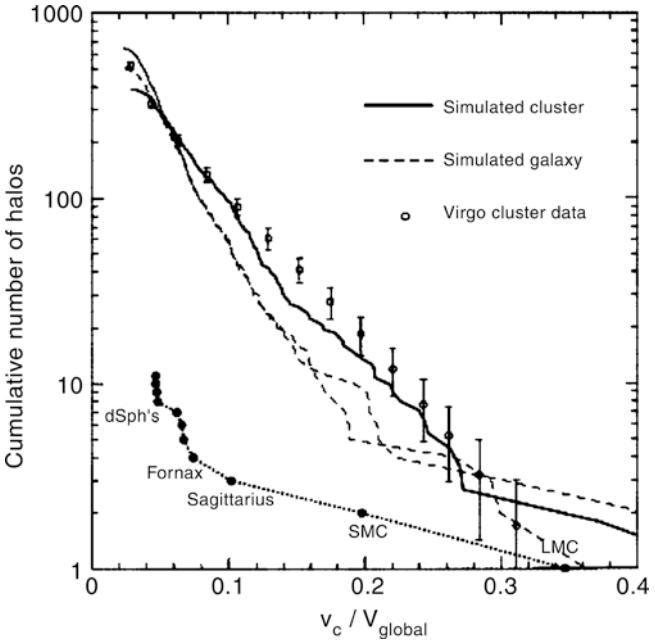


Fig. 7.32 Number density of sub-halos as a function of their mass. The mass is expressed by the corresponding Keplerian rotational velocity v_c, measured in units of the corresponding rotational velocity of the main halo. The *curves* show this number density of sub-halos with rotational velocity $\geq v_c$ for a halo of either cluster mass or galaxy mass. The observed numbers of sub-halos (i.e., of galaxies) in the Virgo cluster are plotted as *open circles* with error bars, and the number of satellite galaxies of the Milky Way as *filled circles*. One can see that the simulations describe the abundance of cluster galaxies quite well, but around the Galaxy significantly fewer satellite galaxies exist than predicted by a CDM model. Source: B. Moore et al. 1999, *Dark Matter Substructure within Galactic Halos*, ApJ 524, L19, p. L20, Fig. 2. ©AAS. Reproduced with permission

will be discussed in Sect. 10.7 in more detail, it is difficult to form a considerable stellar population in halos of masses below $\sim 10^9 M_\odot$. Most halos below this mass threshold will therefore be hardly detectable because of their low luminosity. Thus, in this picture, sub-halos in galaxies are in fact present, as predicted by the CDM models, but most of these would be 'dark'.

The low-mass satellite galaxies in the Local Group that were recently discovered by the Sloan Survey all have a very large mass-to-light ratio, which implies that their stellar mass-to-halo mass ratio is very small. In fact, many of these dwarf galaxies are less luminous than a star cluster, with $L \sim 10^2$–$10^4\, L_\odot$. In contrast to these small luminosities, they all display a rather high stellar velocity dispersion of $\sigma \sim 5\,\mathrm{km/s}$, indicating a fairly high mass. Indeed, these dwarf galaxies are not only the faintest galaxies known, but also those with the largest mass-to-light ratio, $M/L \gtrsim 100$ in Solar units; for some of the newly discovered dwarfs, M/L apparently exceeds 10^4. Their extremely low metallicity argues for a very early epoch of star formation; this is confirmed by the color-magnitude diagrams for some of the

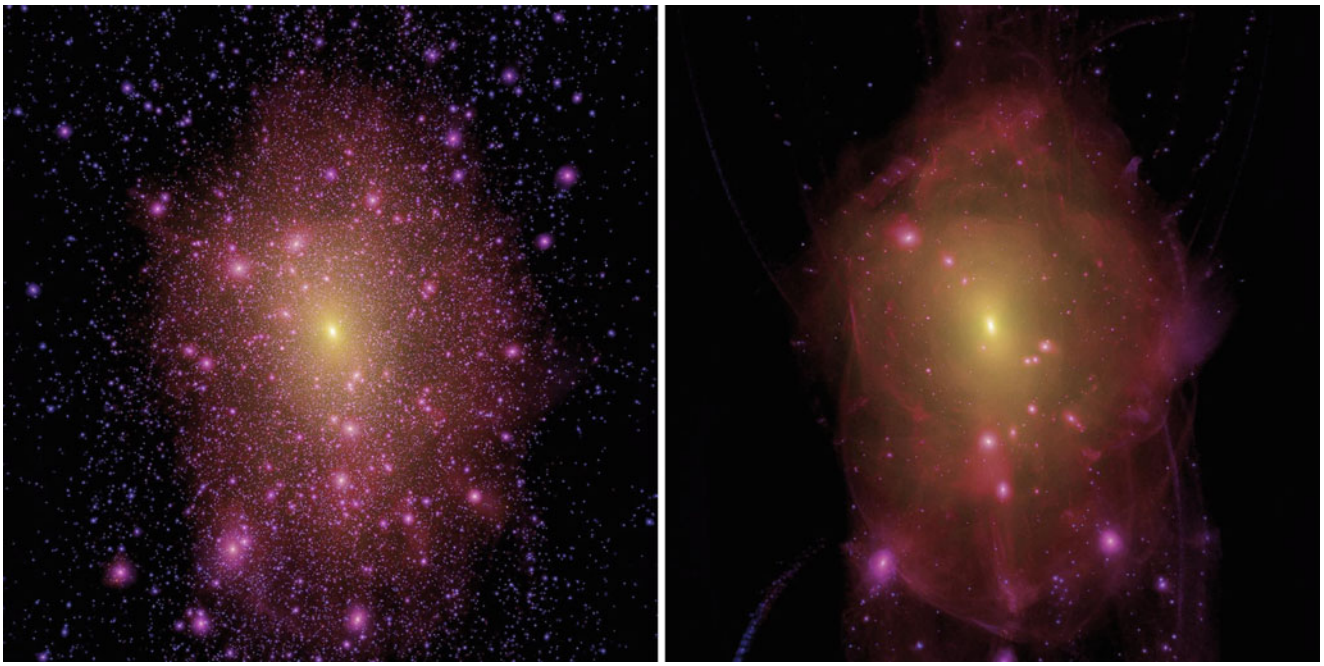


Fig. 7.33 Comparison of a galaxy-mass dark matter halo in the standard CDM model (*left panel*) and in a cosmological model with warm dark matter (*right panel*), within a 1.5 Mpc box. For the simulation, the initial conditions of the density field were chosen to be very similar, except that the power spectrum in the WDM model was truncated, due to the free-streaming of the particles. Obviously, the WDM model halo has far fewer satellite halos; in particular, those of the smallest mass are absent. Source: M. Lovell et al. 2011, *The haloes of bright satellite galaxies in a warm dark matter universe*, arXiv:1104.2929, Fig. 3. Reproduced by permission of the author

dwarfs, which assign them an age of $\sim$ 13 Gys. We shall see in Sect. 10.7 that these observed properties are naturally explained in the framework of galaxy evolution models.

Warm dark matter as alternative. The apparent conflict between the abundance of sub-halos and the observed satellite galaxies in the Milky Way can be potentially avoided if the initial power spectrum of density fluctuations has less power on small spatial scales—corresponding to masses of satellite galaxies. At the same time, the power spectrum at large spatial scales should not be affected, to not endanger the spectacular success of our cosmological model in the description of key cosmological observations (see next chapter). Such a modification of the density fluctuation spectrum would be a consequence of warm dark matter models; as we discussed above, their free streaming would wash out smaller-scale fluctuations. In particular, if the WDM particle has a mass of $\sim$ 2 keV, the cut-off in the power spectrum would correspond to the halos mass of dwarf galaxies.

Figure 7.33 shows the resulting mass distributions of a galaxy-scale halo as predicted by a CDM and a WDM model. In the latter, essentially all small-scale sub-halos are absent, which are found plentiful in the CDM simulation. Hence, in the WDM model, the satellite problem essentially is non-existent.

However, before jumping to conclusions, three points need to be made here. First, a WDM particle appears less natural as seen from the point of view of particle physics, although plausible candidates may exist in some extensions of the Standard Model of particle physics. Second, observations of the Lyman-α forest strongly constrain the allowed mass range of WDM particles (see Sect. 8.5), and lower limits on the mass of the WDM particle obtained from these studies come exceedingly close to the mass needed to substantially reduce the abundance of sub-halos. Third, sub-halos are in fact observed indirectly, as discussed next.

Evidence for the presence of CDM substructure in galaxies. A direct indication of the presence of substructure in the mass distribution of galaxies indeed exists, which originates from gravitational lens systems. As we have seen in Sect. 3.11, the image configuration of multiple quasars can be described by simple mass models for the gravitational lens. Concentrating on those systems with four images of a source, for which the position of the lens is also observed (e.g., with the HST), a simple mass model for the lens has fewer free parameters than the coordinates of the observed quasar images that need to be fitted. Despite of this, it is possible, with very few exceptions, to describe the angular positions of the images with such a model very accurately. This result is not trivial, because for some lens systems which are observed using VLBI techniques, the image positions are known with a precision of better than 10^{-4} arcseconds, with an image separation of the order of $1''$. This result demonstrates that the

mass distribution of lens galaxies is, on scales of the image separation, quite well described by simple mass models.

Besides the image positions, such lens models also predict the magnifications μ of the individual images. Therefore, the ratio of the magnifications of two images should agree with the flux ratio of these images of the background source. The surprising result from the analysis of lens systems is that, although the image positions of (nearly) all quadruply imaged systems are very precisely reproduced by a simple mass model, in not a single one of these systems does the mass model correctly reproduce the flux ratios of the images!

Perhaps the simplest explanation for these results is that the simple mass models used for the lens are not correct and other kinds of lens models should be used. However, this explanation can be excluded for many of the observed systems. Some of these systems contain two or three images of the source that are positioned very closely together, for which one therefore knows that they are located close to a critical curve. In such a case, the magnification ratios can be estimated quite well analytically; in particular, they no longer depend on the exact form of the lens model employed. Hence, the existence of such 'universal properties' of the lens mapping excludes the existence of *simple* (i.e., 'smooth') mass models capable of describing the observed flux ratios. One example of this is presented in Fig. 7.34.

The natural explanation for these flux discrepancies is the fact that a lensing galaxy does not only have a smooth large-scale mass profile, but that there is also small-scale substructure in its density. In the case of spiral galaxies, this may be the spiral arms, which can be seen as a small-scale perturbation in an otherwise smooth mass profile. However, most lens galaxies are ellipticals. The sub-halos that are predicted by the CDM model may then represent the substructure in their mass distribution. For a further discussion of this model, we first should mention that a small-scale perturbation of the mass profile only slightly changes the deflection angle caused by the lens, whereas the magnification μ may be modified much more strongly. As a matter of fact, by means of simulations, it was demonstrated that lens galaxies containing sub-halos of about the same abundance as postulated by the CDM model give rise to a statistical distribution of discrepancies in the flux ratios which is very similar to that found in the observed lens systems. Furthermore, these simulations show that, on average, a particular image of the source is clearly demagnified compared to the predictions by simple, smooth lens models, again in agreement with the observational results. And finally, in the case that a relatively massive sub-halo is located close to one of the images, the image position should also be slightly shifted, compared to the smooth mass model. This effect was in fact directly detected in two lens systems: in these cases, a sub-halo exists in the lens galaxy

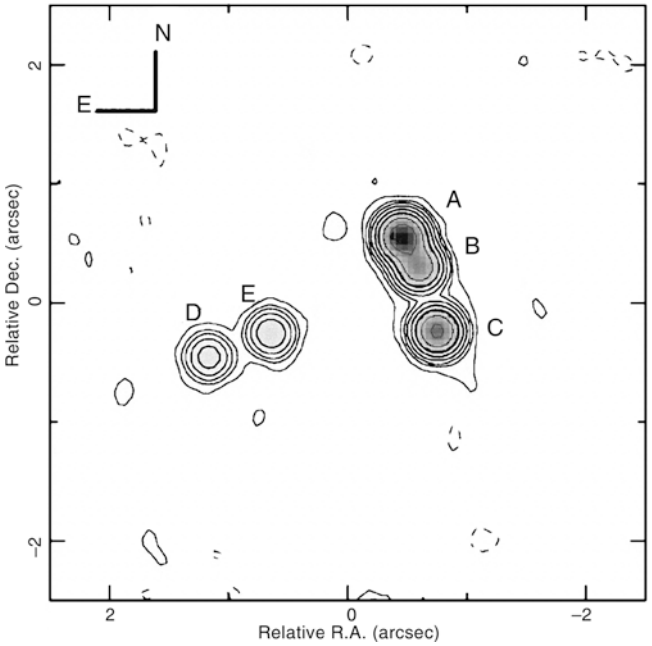


Fig. 7.34 8.5 GHz map of the lens system 2045+265. The source at $z_s = 1.28$ is imaged four-fold (components A–D) by a lens galaxy at $z_d = 0.867$, while component E represents emission from the lens, as is evident from its different radio spectrum. The three images A, B, and C have a separation which is much smaller than the Einstein radius of the lens. From the general properties of the gravitational lens mapping, one can show that any 'smooth' mass model of the lens predicts the flux of the middle one of those (i.e., image B) to be roughly the same as the sum of the fluxes of components A and C. Obviously, this rule is strongly violated in this lens system, because B is weaker than either A or C. This result can only be explained by small-scale structure in the mass distribution of the lens galaxy. Source: C. Fassnacht et al. 1999, *B2045+265: A New Four-Image Gravitational Lens from CLASS*, AJ 117, 658, p. 659, Fig. 1. ©AAS. Reproduced with permission

which is massive enough to form stars, and which therefore can be observed. Its effect on the magnification and the image position can then be inferred from the lens model (see Fig. 7.35).

In strong lensing clusters, one can actually see the impact of mass substructure quite clearly: The three arcs to the left of the center in the cluster Cl 0024+17 (see lower panel of Fig. 6.51) are predicted by any smooth mass model to follow a 'universal' behavior, in that the middle of the arcs should have a length that is the sum of the lengths of the two outer arcs. Clearly, the middle arc is seen to be by far the shortest. This is due to substructure, which is easily identified by the two cluster galaxies located close to the middle arc and thus destroying this universal behavior of the lens mapping.

In addition to these flux-ratio discrepancies, at least two sub-halos have been identified in modeling lens systems with Einstein rings for which very high-quality data were obtained. In one of these cases, where the estimated mass of the sub-halo amounts to $M \sim 3.5 \times 10^9\, M_\odot$, there are

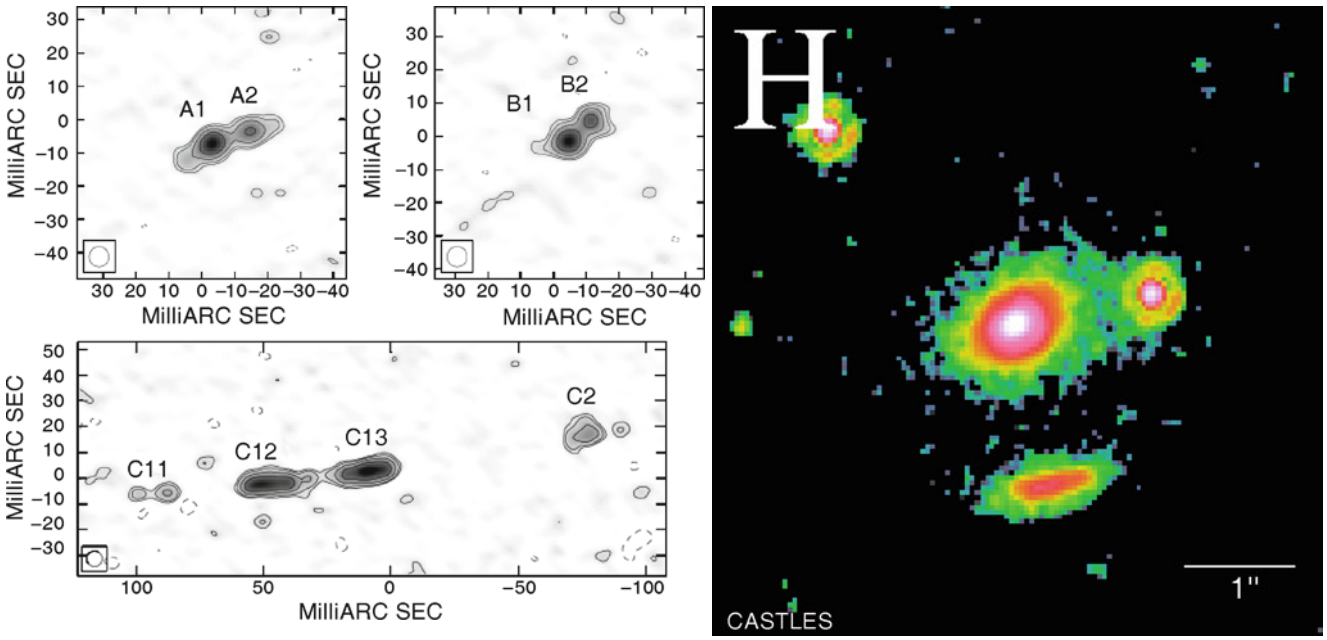


Fig. 7.35 *On the right*, an H-band image of the lens system MG 2016+112 is shown, consisting of a lens galaxy in the center and four images of the background source, the two southernmost of which are nearly merged in this image. *On the left*, VLBI maps of these components are presented; the radio source consists of a compact core and a jet component, clearly visible in images A and B. The VLBI map of component C reveals that is it in fact a double image of the source, in which the core and jet components each are visible twice. Any smooth mass model for the lens galaxy predicts that the separation C12—C11 should roughly be the same as that between C13—C2, which obviously contradicts the observation. In this case, the substructure in the mass distribution is even visible: if one includes the weak emission south of component C, which is visible in the image on the right, into the lens model as a mass component, the separation of the components in image C can be well modeled. Source: *Left*: L.V.E. Koopmans et al. 2002, *2016+112: a gravitationally lensed type II quasar*, MNRAS 334, 39, p. 41, Fig. 1. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society. *Right*: Castles Collaboration/C.S. Kochanek, E.E. Falco, C. Impey, J. Lehar, B. McLeod, H.-W. Rix

stringent limits on its luminosity, which translates into a lower limit of its mass-to-light ratio of 120 in Solar units.

For these reasons, it is probable that galaxies contain sub-halos, as predicted by the CDM model, but most sub-halos, in particular those with low mass, contain only few stars and are therefore not visible. One consequence of this explanation is that the low-mass satellite galaxies that are seen in our Local Group should be dominated by dark matter. Given the faintness and low surface brightness of these galaxies, obtaining kinematical information for them is very difficult and requires large telescopes for spectroscopy of individual stars in these objects. The results of such investigations indicate that the dwarf galaxies in the Local Group are indeed dark matter dominated, with a mass-to-light ratio of ~ 100 in Solar units. Whereas some uncertainty remains, e.g., related to the assumption of dynamical equilibrium, it is clear that these faint satellites represent sub-halos which are unusually poor in stars.

The 'disk of satellite galaxies'. Whereas the abundance of dark matter subhalos in galaxies no longer presents a serious problem for CDM models of structure formation, the spatial distribution of satellite galaxies around the Milky Way requires more explanation. As we mentioned in Sect. 6.1.1, the 11 classical satellites of the Galaxy seem to form a planar distribution. Such a distribution would be extremely unlikely if the satellite population was drawn from a near-isotropic probability distribution. Therefore, the planar satellite distribution has been considered as a further potential problem for CDM-like models. However, using semi-analytic models of galaxy formation, combined with simulations of the large-scale structure, a different picture emerges. Since galaxies preferentially form in filaments of the large-scale structure, the accretion of smaller mass halos onto a high-mass halo occurs predominantly in the direction of the filament. The most massive sub-halos therefore tend to form a planar distribution, not unlike the one seen in the Milky Way's satellite distribution. The anisotropy of the distribution of massive satellites may also serve to explain the Holmberg effect.

7.9 Origin of the density fluctuations

We have seen in Sect. 4.5.3 that the horizon and the flatness problem in the normal Friedmann–Lemaître evolution of the Universe can be solved by postulating an early phase of very rapid—exponential—expansion of the cosmos. In this

inflationary phase of the Universe, any initial curvature of space is smoothed away by the tremendous expansion. Furthermore, the exponential expansion enables the complete currently visible Universe to have been in causal contact prior to the inflationary phase. These two aspects of the inflationary model are so attractive that today most cosmologists consider inflation as part of the standard model, even if the physics of inflation is as yet not understood in detail.

Density fluctuations from inflation. The inflationary model has another property that is considered to be essential. Through the huge expansion of the Universe, microscopic scales are blown up to macroscopic dimensions. The large-scale structure in the current Universe corresponds to microscopic scales prior to and during the inflationary phase. From quantum mechanics, we know that the matter distribution cannot be fully homogeneous, but it is subject to quantum fluctuations, expressed, e.g., by Heisenberg's uncertainty relation. By inflation, these small quantum fluctuations are expanded to large-scale density fluctuations. For this reason, the inflationary model also provides a natural explanation for the origin of initial density fluctuations.

Indeed, it is the only mechanism known in which perturbations can be generated which are larger than the horizon. As we discussed in the framework of the 'horizon problem', two points further apart than $\sim 1^\circ$ have not been in causal contact before recombination when considering standard Friedmann expansion. Despite of this, we observe temperature fluctuations in the CMB on larger scales, which implies that density perturbations larger than the horizon scale were present at $z \sim 1100$. In the same manner as inflation provides an explanation for the horizon problem, it explains the possibility to have superhorizon fluctuations—before the inflationary phase, the whole visible Universe had been in causal contact.

The primordial power spectrum. In fact, one can study the generation of macroscopic density perturbations from quantum fluctuations quantitatively and calculate the initial power spectrum of these fluctuations. The result of such investigations depends slightly on the details of the inflationary model they are based on. However, these models agree in their prediction that the initial power spectrum should have a form very similar to the Harrison–Zeldovich fluctuation spectrum, except that the spectral index n_s of the primordial power spectrum should be slightly smaller than the Harrison–Zeldovich value of $n_\mathrm{s} = 1$. Thus, the model of inflation can be directly tested by measuring the power spectrum and, as we shall see in Chap. 8, the power-law slope n_s indeed seems to be slightly, but significantly flatter that unity, as expected from inflation. The deviation of n_s from unity is called the *tilt* of the initial density fluctuation spectrum.

The various inflationary models also differ in their predictions of the relative strength of the fluctuations of space-time, which should be present after inflation. Such fluctuations are not directly linked to density fluctuations, but they are a consequence of General Relativity, according to which space-time itself is also a dynamical quantity. One consequence of this is the existence of gravitational waves. Although no gravitational waves have been directly detected until now, the analysis of the double pulsar PSR J1915+1606 proves the existence of such waves.[13] Primordial gravitational waves provide an opportunity to empirically distinguish between the various models of inflation. These gravitational waves leave a 'footprint' in the polarization of the cosmic microwave background that is measurable in principle. Several experiments are currently searching for this polarization signature in the CMB.

7.10 Problems

7.1. Homogeneous solution of the Euler–Poisson system. The system of equations (7.2), (7.3), (7.4) admits an exact solution, namely that of a homogeneous universe with an expansion law given by $\boldsymbol{v}(\boldsymbol{r}, t) = H(t)\boldsymbol{r}$, as will be shown here.

1. If we require the density to be homogeneous at all times, show that this implies that $\nabla \cdot \boldsymbol{v}$ is independent of $\boldsymbol{r}$.
2. Show that the continuity equation implies for the Hubble velocity field that the density ρ satisfies (4.11).
3. Determine $\nabla\Phi$ from (7.4), and show that the Euler equation for the Hubble velocity field then reduces to the Friedmann equation (4.19), specialized to the case of vanishing pressure.

7.2. The growth equation. A second-order differential equation such as (7.15) has two linearly independent solutions. In general, they are difficult to find, and in the general case, they must be obtained numerically. However,

[13] The binary pulsar PSR J1915+1606 was discovered in 1974. From the orbital motion of the pulsar and its companion star, gravitational waves are emitted, according to General Relativity. Through this, the system loses kinetic (orbital) energy, so that the size of the orbit decreases over time. Since pulsars represent excellent clocks, and we can measure time with extremely high precision, this change in the orbital motion can be observed with very high accuracy and compared with predictions from General Relativity. The fantastic agreement of theory and observation is considered a definite proof of the existence of gravitational waves. For the discovery of the binary pulsar and the detailed analysis of this system, Russell Hulse and Joseph Taylor were awarded the Nobel Prize in Physics in 1993. In 2003, a double neutron star binary was discovered where pulsed radiation from both components can be observed. This fact, together with the small orbital period of 2.4 h implying a small separation of the two stars, makes this an even better laboratory for studying strong-field gravity.

the growth equation (7.15) can be solved explicitly, as will be shown here.

1. Show that the Hubble function $H(t)$ is a solution of the growth equation. Hint: Make use of the second Friedmann equation (4.19).
2. Unfortunately, $H(t)$ decreases with time, and thus is not the growing solution we are interested in. Show that a second solution of (7.15) is given by

$$D_+(a) = CH(a) \int_0^a \frac{\mathrm{d}a'}{[a'\,H(a')]^3} \,,$$

   where C is a constant, chosen such that $D_+(1) = 1$. Hint: make use of the first part of this problem.
3. Use the expansion law of an Einstein–de Sitter universe, $a(t) = (t/t_0)^{2/3}$, to show that the corresponding Hubble function $H(t)$ and D_+ as calculated from (7.17) solve the growth equation (7.15).

7.3. Bulk properties of dark matter halos. Consider two dark matter halos of mass $10^{12}h^{-1}M_\odot$ and $10^{15}h^{-1}M_\odot$, corresponding to the halos of a massive galaxy and of a massive cluster, respectively. Assume $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0.7$.

1. What is the virial radius r_{200} and the virial velocity V_{200} of these halos at redshift $z = 2$ and today? Hint: make use of the fact that the Schwarzschild radius of the Sun is $\approx 3\,\mathrm{km} = 3 \times 10^5\,\mathrm{cm}$, and $c/H_0 \approx 3h^{-1}\,\mathrm{Gpc} \approx 9 \times 10^{27}h^{-1}\,\mathrm{cm}$
2. In order to assemble this mass into a halo, matter from a large region must have accumulated. Assuming that this region is spherical, determine its comoving radius.
3. This radius can be identified with the typical length-scale of a perturbation out of which such dark matter halos grow. At which redshift do these density fluctuations enter the horizon?

7.4. Behavior of the growth factor. We have seen that, in an Einstein–de Sitter universe, the growth factor equals the scale factor—see (7.19). In universes with curvature and/or a cosmological constant, this is no longer the case, as seen in Fig. 7.3.

1. Show that for sufficiently small values of a during the matter-dominated epoch, the growth factor is proportional to a in all universes.
2. Derive the lowest-order correction to this linear behavior, and estimate from that the epoch when significant deviations from the linear behavior of the growth factor occur. Assume for simplicity that the universe is flat, and compare your estimate with Fig. 7.3.

Cosmology III: The cosmological parameters

8

In Chap. 4 and 7, we described the fundamental aspects of the standard model of cosmology. Together with the knowledge of galaxies, clusters of galaxies, and AGNs that we have gained in the other chapters, we are now ready to discuss the determination of the various cosmological parameters. In the course of this discussion, we will describe a number of methods, each of which is in itself useful for estimating cosmological parameters, and we will present the corresponding results from these methods. The most important aspect of this chapter is that we now have more than one independent estimate for each cosmological parameter, so that the determination of these parameters is highly redundant. This very aspect is considerably more important than the precise values of the parameters themselves, because it provides a test for the consistency of the cosmological model.

We will give an example in order to make this point clear. In Sect. 4.4.5, we discussed how the cosmic baryon density can be determined from primordial nucleosynthesis and the observed ratio of deuterium to hydrogen in the Universe. Thus, this determination is based on the correctness of our picture of the thermal history of the early Universe, and on the validity of the laws of nuclear physics shortly after the Big Bang. As we will see later, the baryon density can also be derived from the angular fluctuations in the cosmic microwave background radiation, for which the structure formation in a CDM model, discussed in the previous chapter, is needed as a foundation. If our standard model of cosmology was inconsistent, there would be no reason for these two values of the baryon density to agree—as they do in a remarkable way. Therefore, in addition to obtaining a more precise value of Ω_b from this combination than from each of the individual methods alone, the agreement is also a strong indication of the validity of the standard model.

We will begin in Sect. 8.1 with discussing the observation of the large-scale distribution of matter, the large-scale structure (LSS). It is impossible to observe the (total) matter distribution itself; rather, only the spatial distribution of visible galaxies can be measured. Assuming that the galaxy distribution follows, at least approximately (which we will specify later), that of the dark matter, the power spectrum (or the correlation function) of the density fluctuations can be estimated from that of the galaxies. As pointed out in the previous chapter, the power spectrum in turn depends on the cosmological parameters. In Sect. 8.2, we will summarize some aspects of clusters of galaxies which are relevant for the determination of the cosmological parameters.

In Sect. 8.3, type Ia supernovae will be considered as cosmological tools, and we will discuss their Hubble diagram. Since SN Ia are approximately standard candles, their Hubble diagram provides information on the density parameters Ω_m and Ω_Λ. These observations provided the first direct evidence, around 1998, that the cosmological constant differs from zero. We will then analyze the lensing effect of the LSS in Sect. 8.4, by means of which information about the statistical properties of the LSS of matter is obtained directly, without the necessity for any assumptions on the relation between matter and galaxies. As a matter of fact, this galaxy-mass relation can be directly inferred from the lens effect. In Sect. 8.5, we will turn to the properties of the intergalactic medium and, in particular, we will introduce the Lyman-α forest in QSO spectra as a cosmological probe.

The anisotropy of the cosmic microwave background is discussed in quite some detail in Sect. 8.6. Through observations and analysis of the temperature (and polarization) fluctuations in the CMB, a vast amount of very accurate information about the cosmological parameters are obtained. In particular, we will report on the recent and exciting results concerning CMB anisotropies, and will combine these findings with the results obtained by other methods in Sect. 8.7. This combination yields a set of parameters for the cosmological model which is able to describe nearly all observations of cosmological relevance in a self-consistent manner, and which today defines the standard model of cosmology.

The final section of this chapter will be dedicated to a discussion of the properties of Dark Energy whose presence

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_8,

has been established with the various cosmological probes. As arguably the greatest mystery in current fundamental physics, it is of utmost interest to find out whether it is really a cosmological constant or has some more interesting features, e.g., being time dependent.

8.1 Redshift surveys of galaxies

8.1.1 Introduction

The inhomogeneous large-scale distribution of matter that was described in Chap. 7 is not observable directly because it consists predominantly of dark matter. If it is assumed that the distribution of galaxies traces the underlying distribution of dark matter fairly, the properties of the LSS of matter could be studied by observing the galaxy distribution in the Universe. Quite a few good reasons exist for this assumption not to be completely implausible. For instance, we observe a high galaxy density in clusters of galaxies, and with the methods discussed in Chap. 6, we are able to verify that clusters indeed represent strong mass concentrations. Qualitatively, this assumption therefore seems to be justified, though in detail, essential modifications are necessary, as we will discuss later.

In any case, the distribution of galaxies on the sphere appears inhomogeneous and features large-scale structure. Since galaxies have evolved from the general cosmic density field, they should contain information about the latter. It is consequently of great interest to examine and quantify the properties of the galaxy distribution.

In principle, two possible ways exist to accomplish this study of the galaxy distribution. With photometric sky surveys, the two-dimensional distribution of galaxies on the sphere can be mapped. To also determine the third spatial coordinate, it is necessary to measure the redshift of the galaxies using spectroscopy, deriving the distance from the Hubble law (1.6) and its generalization for higher redshifts as discussed in Chap. 4. It is obvious that we can learn considerably more about the statistical properties of the galaxy distribution from their three-dimensional distribution; hence, redshift surveys are of particular interest.

The graphical representation of the spatial galaxy positions is accomplished with so-called wedge diagrams. They represent a sector of a circle, with the Milky Way at its center. The radial coordinate is proportional to z (or cz—by this, the distance is measured in km/s), and the polar angle of the diagram represents an angular coordinate in the sky (e.g., right ascension), where an interval in the second angular coordinate is selected in which the galaxies are located. An example for such a wedge diagram is shown in Fig. 8.1.

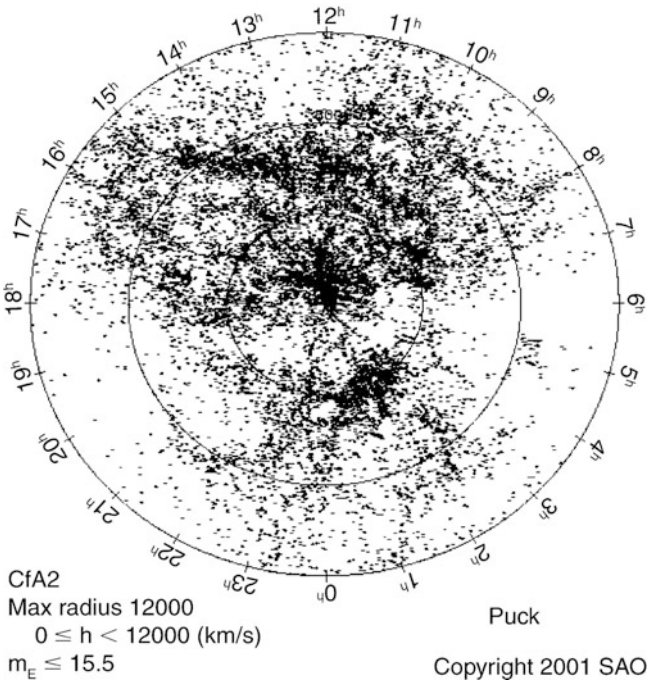


Fig. 8.1 The CfA redshift survey, in equatorial coordinates. Along its radial axis, this wedge diagram shows the escape velocity cz up to 12 000 km/s, and the polar angle specifies the right ascension of a galaxy. The Great Wall extends from 9 to 15^h. The overdensity at 1^h and $cz = 4000$ km/s is the Pisces-Perseus-supercluster. Credit: M. Geller, J. Huchra, Smithsonian Astrophysical Observatory

8.1.2 Redshift surveys

Performing redshift surveys is a very time-consuming task compared to making photometric sky maps, because recording a spectrum requires much more observing time than the mere determination of the apparent magnitude of a source. Hence, the history of redshift surveys, like that of many other fields in astronomy, is driven by the development of telescopes and instruments. The introduction of CCDs in astronomy in the early 1980s provided a substantial increase in sensitivity and accuracy of optical detectors, and enabled redshift surveys of galaxies in the nearby Universe containing several thousand galaxies (see Fig. 8.1). Using a single slit in the spectrograph implied that in each observation the spectra of only one or very few galaxies could be recorded simultaneously. The situation changed with the introduction of spectrographs with high multiplexity which were designed specifically to perform redshift surveys. With them, the spectra of many objects (up to $\sim 10^3$) in the field-of-view of the instrument can be observed simultaneously.

The strategy of redshift surveys. Such a survey is basically defined by two criteria. The first is its geometry: a region of the sky is chosen in which the survey is performed.

Second, those objects in this region need to be selected for which spectra should be obtained. In most cases, for practical reasons the objects are selected according to their brightness, i.e., spectra are taken of all galaxies above a certain brightness threshold. The latter defines the number density of galaxies in the survey, as well as the required exposure time. For the selection of spectroscopic targets, a photometric catalog of sources is required as a starting point. The selection criteria may be refined further in some cases. For instance, a minimum angular extent of objects may be chosen to avoid the inclusion of stars. The spectrograph may set constraints on the selection of objects; e.g., a multi-object spectrograph is often unable to observe two sources that are too close together on the sky.

Examples of redshift surveys. In the 1980s, the *Center for Astrophysics (CfA)-Survey* was carried out which measured the redshifts of more than 14 000 galaxies in the local Universe (Fig. 8.1). The largest distances of these galaxies correspond to about $cz \sim 15\,000\,\mathrm{km/s}$, or $D = cz/H_0 = 150h^{-1}\,\mathrm{Mpc}$. One of the most spectacular results from this survey was the discovery of the 'Great Wall', a huge structure in the galaxy distribution (see also Fig. 7.2).

In the *Las Campanas Redshift Survey (LCRS)*, carried out in the first half of the 1990s, the redshifts of more than 26 000 galaxies were measured. They are located in six narrow strips of 80° length and 1.5° width each. With distances of up to ∼ 60 000 km/s, this survey is considerably deeper than the CfA Redshift Survey. The distribution of galaxies is displayed in Fig. 8.2, from which we can recognize the typical bubble or honey-comb structure. Galaxies are distributed along filaments, which are surrounding large regions in which virtually no galaxies exist—the aforementioned voids.[1] The galaxy distribution shows a structure which is qualitatively very similar to the dark matter distribution generated in numerical simulations (see, e.g., Fig. 7.13). In addition, we see from the galaxy distribution that no structures exist with scales comparable to the extent of the survey. Thus, the LCRS has probed a scale larger than that where significant structures of the mass distribution are found. The survey volume of the LCRS therefore covers a representative section of the Universe.

A different kind of redshift survey became possible through the sky survey carried out with the IRAS satellite (see Sect. 1.3.2). In these redshift surveys of IRAS galaxies, the selection of spectroscopic targets was based on the 60 μm flux measured by IRAS in its (near) all-sky survey. Various

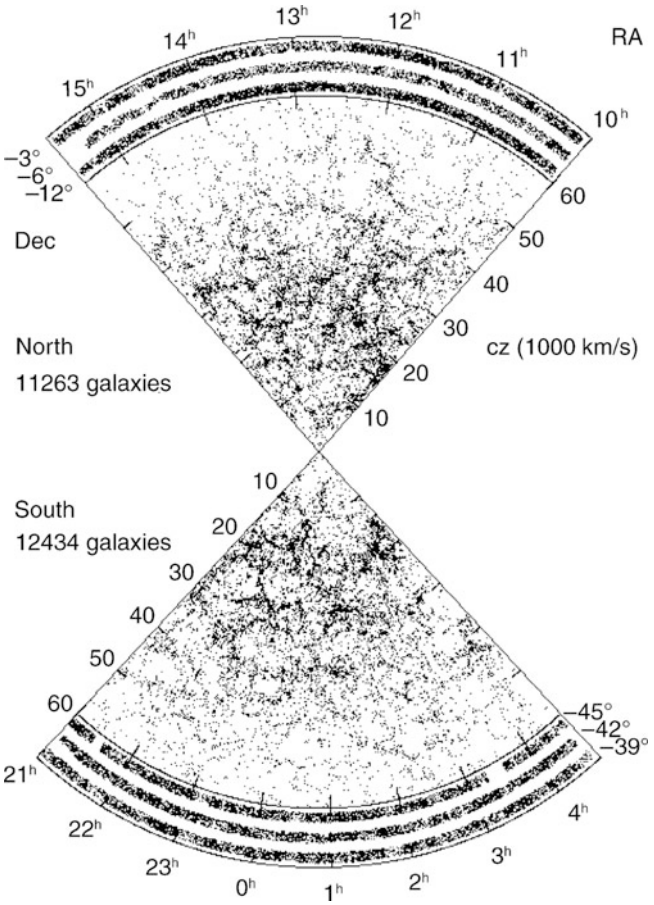


Fig. 8.2 The Las Campanas Redshift Survey consists of three fields each at the North and South Galactic Pole. Each of these fields is a strip 1.5° wide and 80° long. Overall, the survey contains about 26 000 galaxies, and the median of their redshift is about 0.1. The six strips show the distribution of galaxies on the sphere, and the wedge diagram indicates, for galaxies with measured redshift, the right ascension versus distance from the Milky Way, measured in units of $1000\,\mathrm{km\,s^{-1}}$. Source: Lin et al. 1996, *The Power Spectrum of Galaxy Clustering in the Las Campanas Redshift Survey*, ApJ 471, 617, p. 618, Fig. 1. ©AAS. Reproduced with permission

redshift surveys are based on this selection, differing in the flux limit applied; for example the 2 Jy survey (hence, $S_{60\mu m} \geq 2\,\mathrm{Jy}$), or the 1.2 Jy survey. The QDOT and PSCz surveys both have a limiting flux of $S_{60\mu m} \geq 0.6\,\mathrm{Jy}$, where QDOT observed spectra for one out of six randomly chosen galaxies from the IRAS sample, while PSCz is virtually complete and contains ∼ 15 500 redshifts. One of the advantages of the IRAS surveys is that the FIR flux is nearly unaffected by Galactic absorption, an effect that needs to be corrected for when galaxies are selected from optical photometry. Furthermore, the PSCz is an 'all-sky' survey, containing the galaxy distribution in a sphere around us, so that we obtain a complete picture of the local galaxy distribution. However, one needs to be aware of the fact that in selecting galaxies via their FIR emission one is thus

[1] Note that in Fig. 8.2, as in all similar redshift surveys, the number density of galaxies in the diagrams decreases outwards. This is an immediate consequence of the limiting flux of the spectroscopic survey. As one goes to larger distances, only the most luminous galaxies exceed the flux threshold, leading to a decreasing number density.

selecting a particular type of galaxy, predominantly those which have a high dust content and active star formation which heats the dust.

The *Canada-France Redshift Survey (CFRS)* obtained spectroscopy of faint galaxies with $17.5 \leq I \leq 22.5$, with a median redshift of about 0.5. The resulting catalog contains 948 objects, 591 of which are galaxies. This survey was performed by a multi-object spectrograph at the CFHT (see Sect. 1.3.3) which was able to take the spectra of up to 100 objects simultaneously. For the first time, due to its faint limiting magnitude it enabled us to study the evolution of (optically-selected) galaxies, for example by means of their luminosity function and their star formation rate, and to investigate the redshift dependence of the galaxy correlation function—and thus to see the evolution of the large-scale structure.

Recently, two large spectroscopic surveys with faint limiting magnitudes were carried out. Both of them use high multiplex spectrographs mounted on 10-m class telescopes: the VIMOS instrument on the VLT and the DEIMOS instrument on Keck. Both surveys, the VIMOS VLT Deep Survey (VVDS) and the DEEP2 survey, obtained spectra of several tens of thousands of galaxies with $z \sim 1$, thus extending the CFRS by more than an order of magnitude in sample size and by ~ 1.5 magnitudes in depth.

The 2dF-Survey and the Sloan Digital Sky Survey. The scientific results from the first redshift surveys motivated the conduction of considerably more extended surveys. By averaging over substantially larger volumes in the Universe, it was expected that the statistics on the galaxy distribution could be significantly improved. In addition, the analysis of the galaxy distribution at higher redshift would also enable a measurement of the evolution in the galaxy distribution. With these main objectives in mind, two very extensive redshift surveys were performed at the beginning of this millennium: the *2 degree Field Galaxy Redshift Survey (2dFGRS)* and the *Sloan Digital Sky Survey (SDSS)*.

The 2dFGRS was carried out using a spectrograph specially designed for this project, which was mounted at the 4-m Anglo Australian Telescope. Using optical fibers to transmit the light of the observed objects from the focal plane to the spectrograph, up to 400 spectra could be observed simultaneously over a usable field with a diameter of 2°. The positioning of the individual fibers on the location of the pre-selected objects was performed by a robot. The redshift survey covered two large connected regions in the sky, of 75° × 15° and 75° × 7.5°, plus 100 additional, randomly distributed fields. This survey geometry was chosen so as to optimize the cosmological information from the galaxy distribution, that is, the most precise measurement of the correlation function at relevant scales. The photometric input catalog was the APM galaxy catalog which had been compiled from digitized photographic plates. The limiting magnitude of the galaxies for which spectra were obtained is approximately $B \lesssim 19.5$, where this value is corrected for Galactic extinction. The 2dFGRS contains redshifts for more than 230 000 galaxies (see Fig. 7.1). The spectra and redshifts are publicly available. The scientific yield from this large data set is very impressive; some of these results will be shown further below.

The SDSS survey was described already in Sect. 1.4; its strategy was similar to the 2dFGRS, except that besides spectroscopy of more than a million objects—mostly galaxies—also a fifth of the sky was mapped in five optical bands. The selection of targets for spectroscopy was carried out using this photometric information. The photometric data has been used in a large variety of other Galactic and extragalactic projects.

Both the 2dF survey and the SDSS also recorded, besides the spectra of galaxies, those of QSOs which were selected based on their optical colors; this yielded by far the most extensive QSO surveys. It must be pointed out that redshift surveys deliver not only the redshifts of galaxies and AGNs, but also rich additional spectral information which has a wide range of applications for studying the physical properties of these sources. In fact, we made use of many of these results in discussing the properties of galaxies in Chap. 3.

8.1.3 Determination of the power spectrum

We will now return to the question of whether the distribution of (dark) matter in the Universe can be derived from the observed distribution of galaxies. If galaxies trace the distribution of dark matter fairly, the power spectrum or correlation function of dark matter could be determined from the galaxy distribution. However, this is not necessarily the case; for instance, it may be that there is a threshold in the local density of dark matter, below which the formation of galaxies does not occur or is at least strongly suppressed. Thus, the relative spatial distribution of galaxies and dark matter will depend on how galaxies form and evolve. Whereas there has been great progress on that topic in the past years, we cannot accurately predict the relation between galaxies and the underlying matter distribution, at least not without introducing a number of model assumptions. Alternatively we can parametrize this uncertainty, which is the common approach.

Biasing of galaxies. In analogy with the biasing of dark matter halos discussed in Sect. 7.6.3, we parametrize the connection between dark matter and galaxies by the so-called *linear bias factor* b_g of galaxies. It is defined in a similar manner: We consider the discrete galaxy field to be smoothed

over spheres with radius R, to get the continuous galaxy number density field n; then we make the ansatz

$$\delta_{\rm g} := \frac{\Delta n}{\bar{n}} = b_{\rm g} \frac{\Delta \rho}{\bar{\rho}} = b_{\rm g}\, \delta_R \ , \tag{8.1}$$

where $\Delta n = n - \bar{n}$ is the deviation of the local number density of galaxies from their average density. Hence, the bias factor is the ratio of the relative overdensities of galaxies to dark matter. Such a linear relation is not strictly justified from theory. The galaxy bias factor b may depend on the galaxy type and on redshift; in addition, it may depend on the smoothing scale R. However, one might expect that the scale dependence disappears for sufficiently large R where the density field evolves linearly.

The definition given in (8.1) must be understood in a statistical sense. In a volume $V = (4\pi/3)R^3$, we expect on average $\bar{N} = \bar{n}\, V$ galaxies, whereas the observed number of galaxies is $N = n\, V$. Hence,

$$\left(\frac{\Delta n}{\bar{n}}\right)_R = \frac{n - \bar{n}}{\bar{n}} = \frac{N - \bar{N}}{\bar{N}} = b_{\rm g}\, \delta_R \ .$$

Under the assumption of linear biasing, we can then infer the statistical properties of matter from those of the galaxy distribution.

Normalization of the power spectrum. In Sect. 7.4.2, we demonstrated that the power spectrum of the density fluctuations can be predicted in the framework of a CDM model, except for its normalization which has to be measured empirically. A convenient way for its parametrization is through the parameter σ_8. This parameter is motivated by the following observation.

Analyzing spheres of radius $R = 8h^{-1}\,$Mpc in the local Universe, it is found that optically-selected galaxies have, on this scale, a fluctuation amplitude of about unity,

$$\sigma^2_{8,{\rm g}} := \left\langle \left(\frac{\Delta n}{\bar{n}}\right)^2 \right\rangle_8 \approx 1 \ , \tag{8.2}$$

where the averaging is performed over different spheres of identical radius $R = 8h^{-1}\,$Mpc. Accordingly, we define the dispersion of the matter density contrast, averaged over spheres of radius $R = 8h^{-1}\,$Mpc as

$$\sigma_8^2 = \left\langle \left| \delta_{8h^{-1}\,{\rm Mpc}} \right|^2 \right\rangle \ . \tag{8.3}$$

Note that this definition is a special case of (7.50). Using the definition of the bias factor (8.1), we then obtain

$$\sigma_8 = \frac{\sigma_{8,{\rm g}}}{b_{\rm g}} \approx \frac{1}{b_{\rm g}} \ . \tag{8.4}$$

Because of this simple relation, it has become common practice to use σ_8 as a parameter for the normalization of the power spectrum.[2] If $b_{\rm g} = 1$, thus if galaxies trace the matter distribution fairly, then one has $\sigma_8 \approx 1$. If $b_{\rm g}$ is not too different from unity, we see that the density fluctuations on a scale of $\sim 8h^{-1}\,$Mpc are becoming non-linear at the present epoch, in the sense that $\delta_{8h^{-1}\,{\rm Mpc}} \sim 1$. On larger scales, however, the evolution of the density contrast can approximately be described by linear perturbation theory.

Shape of the power spectrum. If one assumes that $b_{\rm g}$ does not depend on the length-scale considered, the *shape* of the dark matter power spectrum can be determined from the power spectrum $P_{\rm g}(k)$ of the galaxies, since $P_{\rm g}(k) = b_{\rm g}^2\, P(k)$, whereas its amplitude depends on $b_{\rm g}$. As we have seen in Sect. 7.4.2, the shape of $P(k)$ is described by the shape parameter $\Gamma = h\,\Omega_{\rm m}$ in the framework of CDM models.

It should be briefly noted here that the determination of the power spectrum from a three-dimensional distribution of galaxies is a non-trivial task. It is not calculated from the correlation function according to (7.29), since for calculating $P(k,z)$ according to this transformation, one would need to know the correlation function at arbitrarily large scales. Instead, the density field of galaxies is used directly for the estimate of the power spectrum. Several issues need to be considered carefully: The survey geometry is non-trivial, due to the finite angular size of the survey and instrumental restrictions about the smallest separation of objects for which spectra can be recorded. The galaxies are flux limited, and so the luminosity threshold for the galaxy population is redshift-dependent. As we shall see below, this effect requires to take into account the luminosity-dependent galaxy bias. Finally, the determination of the power spectrum is affected by peculiar velocities of galaxies, as will be described below in Sect. 8.1.5.

Early results from the comparison of the shape of the power spectrum of galaxies with that of CDM models yielded $\Gamma \sim 0.25$ (see Fig. 8.3). Since $\Gamma = h\Omega_{\rm m}$, this result indicates a universe of low density (unless h is unreasonably low).

In the 2dFGRS, the power spectrum of galaxies was measured with a much higher accuracy than had previously been possible. Since a constant $b_{\rm g}$ can be expected, at best, in the linear domain, i.e., on scales above $\sim 10h^{-1}\,$Mpc, only such linear scales are used in the comparison with the power spectra predicted by CDM models. As the density parameter $\Omega_{\rm m}$ seems to be relatively small, the baryonic density plays a noticeable role in the transfer function [see (7.35)] which depends on $\Omega_{\rm b}$ as well as on Γ. In particular, we saw in Sect. 7.4.3 that the presence of baryons leads to oscillations in the matter distribution, the baryonic acoustic oscillations. In Fig. 8.4, the different curves show the predicted power spectrum for models with different shape parameter Γ, and

[2] More precisely, one considers σ_8 the normalization of the power spectrum linearly extrapolated to the present day, $P_0(k)$, so that the relation (8.4) needs to be modified slightly.

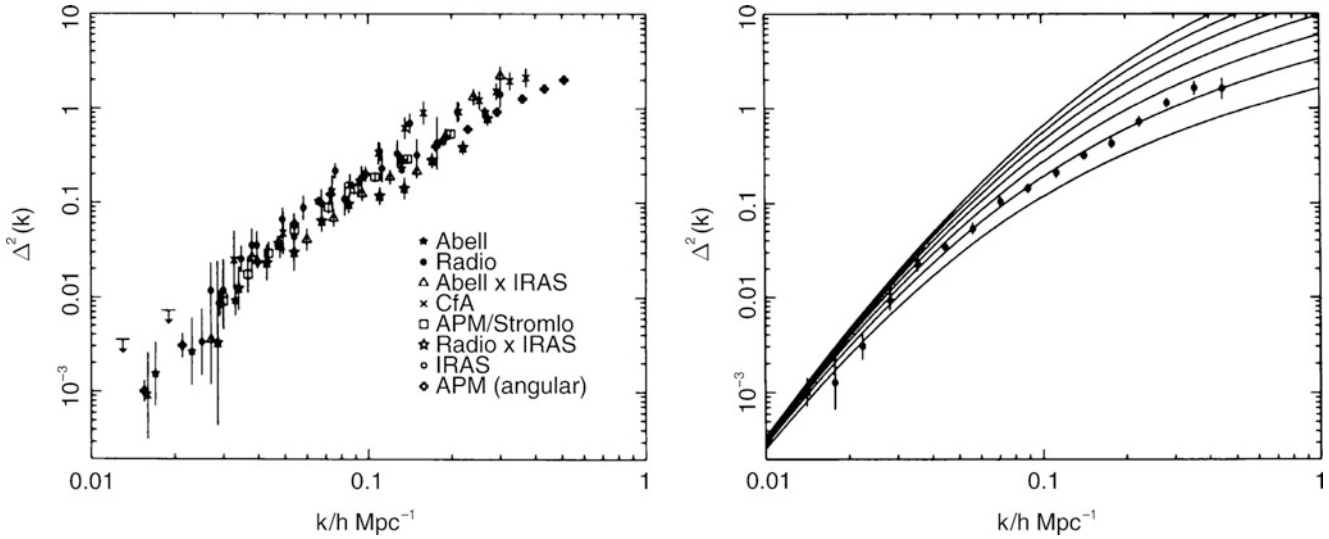


Fig. 8.3 *On the left*, the power spectrum of galaxies is displayed, as determined from different galaxy surveys, where $\Delta^2(k) \propto k^3\, P(k)$ is a dimensionless description of the power spectrum. *On the right*, model spectra for $\Delta^2(k)$ are plotted, where Γ varies from 0.5 (*uppermost curve*) to 0.2; the data from the various surveys have been suitably averaged. We see that a value of $\Gamma \sim 0.25$ for the shape parameter fits these early observations quite well. Source: J.A. Peacock & S.J. Dodds 1994, *Reconstructing the Linear Power Spectrum of Cosmological Mass Fluctuations*, MNRAS 267, 1020, p. 1029, 1031, Figs. 6, 9. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

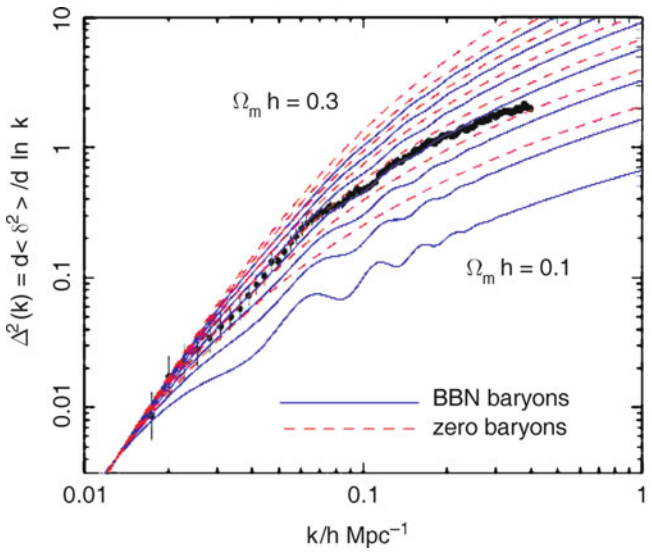


Fig. 8.4 Power spectrum of the galaxy distribution as measured in the 2dFGRS (points with error bars), here represented as $\Delta^2(k) = k^3\, P(k)/(2\pi^2)$. The curves show power spectra from CDM models with different shape parameter $\Gamma = \Omega_{\rm m} h$, and two values of $\Omega_{\rm b}$: one as obtained from primordial nucleosynthesis (BBN, *solid curves*), and the other for models without baryons (*dashed curves*). The Hubble constant $h = 0.7$ and the slope $n_{\rm s} = 1$ of the primordial power spectrum were assumed. A very good fit to the observational data is obtained for $\Gamma \approx 0.2$. Source: J.A. Peacock 2003, *Large-scale surveys and cosmic structure*, astro-ph/0309240, Fig. 5

two different values for the baryon density $\Omega_{\rm b}$, in one case set to zero (dashed curves), in the other case fixed to the value obtained from Big Bang nucleosynthesis (see Sect. 4.4.5). The relative amplitude of the oscillations is the larger, the smaller $\Omega_{\rm m}$ is, because for smaller $\Omega_{\rm m}$, the ratio of baryons to dark matter is larger, and thus their impact of larger relative importance.

The measurement accuracy of the galaxy distribution in the 2dFGRS is high enough to be sensitive to this dependence. Figure 8.4 includes the measured power spectrum of galaxies from the 2dFGRS which is thus compared to the predictions of the models. Apparently, the models which include baryons provide a better fit to the data. The best fit with a CDM model is characterized by

$$\Gamma = \Omega_{\rm m} h = 0.18 \pm 0.02\,, \quad \Omega_{\rm b}/\Omega_{\rm m} = 0.17 \pm 0.06\,. \tag{8.5}$$

In order to combine the correlation functions (or power spectra) of different galaxy populations to obtain the power spectrum of the matter distribution, the bias factor of the galaxies needs to be known. Or, slightly less ambitious, in order to obtain the *shape* of the matter power spectrum from the clustering of different types of galaxies, the *relative* bias of these galaxies needs to be determined. If the assumption of a linear bias is satisfied, than the correlation functions of different populations of galaxies should satisfy $\xi_i(r) = b_i^2 \xi(r)$, where the index labels the galaxy population which has a bias factor b_i, and $\xi(r)$ is the matter correlation function. This relation then implies that $\xi_i(r)/b_i^2 = \xi_j(r)/b_j^2$, and so the relative bias factors can be determined.

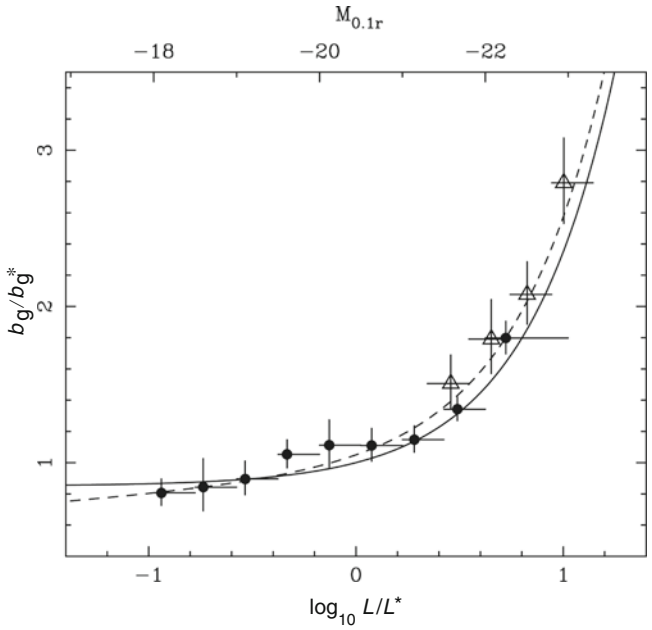


Fig. 8.5 The bias of galaxies in the SDSS as a function of galaxy luminosity, normalized to that of L^*-galaxies. The *filled circles* correspond to the galaxies from the main sample of SDSS galaxies, *open triangles* to the luminous red galaxies. The *solid curve* shows the relation $b_g/b_g^* = 0.85 + 0.15\,L/L^*$, previously obtained from the 2dFGRS, whereas the *dashed curve* shows a slightly different functional form. Source: W.J. Percival et al. 2007, *The Shape of the Sloan Digital Sky Survey Data Release 5 Galaxy Power Spectrum*, ApJ 657, 645, p. 650, Fig. 6. ©AAS. Reproduced with permission

This procedure was applied to the SDSS galaxy survey, which consists of two separate galaxy samples. The first one is the 'main sample', which consists of all galaxies with a (Galactic extinction corrected) limiting magnitude of $r = 17.77$; their median redshift is $z_{\rm med} \sim 0.11$. In addition, for a deeper spectroscopic survey, red galaxies were selected. Those with a high intrinsic luminosity were then combined into the sample of luminous red galaxies (LRGs).

Figure 8.5 shows the bias factor of SDSS galaxies as a function of their luminosity, scaled to the bias of an L^*-galaxy. The bias factor decreases toward fainter luminosities, but apparently reaches a limiting value for $L \to 0$. Conversely, for $L \gtrsim L^*$ the bias increases strongly with luminosity: more luminous galaxies are more strongly correlated than less luminous ones.

As an aside, it is tempting to compare the result in Fig. 8.5 with the behavior of the halo bias with mass (Fig. 7.23). The two figures show a remarkable qualitative similarity. This similarity could be understood if every halo of mass M contains one galaxy of luminosity $L(M)$, in which case the galaxy bias and halo bias are related to each other by $b_h(M) = b_g(L(M))$. We will see later that this very simple model actually works quite well.

Accounting for the luminosity dependence of the galaxy bias, the power spectrum of the SDSS main galaxy sample was constructed and is shown in Fig. 8.6, together with power spectra determined before. If our bias model is correct, then this would resemble the matter power spectrum, up to an overall normalization (which is uncertain due to the unknown value of b_g^*). This power spectrum can now be compared with that of CDM models; if restricted to sufficiently large scales (small k), the perturbations can still be considered to be in the linear regime, so the linear CDM power spectra should apply. Two different model fits are shown that differ by the range of k-values in which the fit was applied. We see that these two model fits differ somewhat, which indicates that the reconstructed power spectrum does not fully resemble that of a CDM model. As a consequence, the best-fitting model parameters are different, as shown in the right-hand panel of Fig. 8.6: whereas the ratio of the baryon-to-total matter density parameter is almost the same for these two models, the best-fitting value of Ω_m is significantly different.[3] Furthermore, the three different power spectra shown are slightly different, in particular on smaller scales. Apart from the potential systematics or uncertainties remaining in the analysis, these differences could be traced back to a scale-dependent galaxy bias, in combination with the different mix of red and blue galaxies in both redshift surveys. In fact, if the galaxy samples in both surveys are restricted to red galaxies, then the resulting power spectra do agree well.

8.1.4 Baryonic acoustic oscillations

We have shown in Sect. 7.4.3 that the photon-baryon fluid contained sound waves prior to recombination, which then were frozen when the Universe turned neutral due to the drop of the baryonic sound speed to effectively zero. The length-scale of the resulting density perturbation (7.46) is given by the sound horizon at recombination; it depends only on the baryon-to-photon and the matter-to-radiation density ratios. The former is proportional to $\Omega_b h^2$ and is determined, e.g., from Big Bang Nucleosynthesis; the latter is proportional to $\Omega_m h^2$ which, as we will see in Sect. 8.7, is well determined from measurements of the CMB. Together, one finds that the acoustic scale has a comoving value of $r_s \approx 150\,$Mpc.

As mentioned in Sect. 7.4.3, we expect that these frozen sound waves of the baryons leave an imprint on the overall matter correlation function, best seen in Fig. 7.7. This unique feature in the correlation function, if it can be detected in the galaxy correlation function, would provide a well-defined 'standard rod' in the observable Universe. Since we can observe only angular scales on the sky, the relation between

[3] This can be understood from the fact that the ratio Ω_b/Ω_m affects the power spectrum mainly on large scales—where the two fits agree quite well—whereas the value of Ω_m determines the shape parameter Γ. Changing Γ shifts the peak of the power spectrum, and due to its curvature affects the local slope in the k-range considered here.

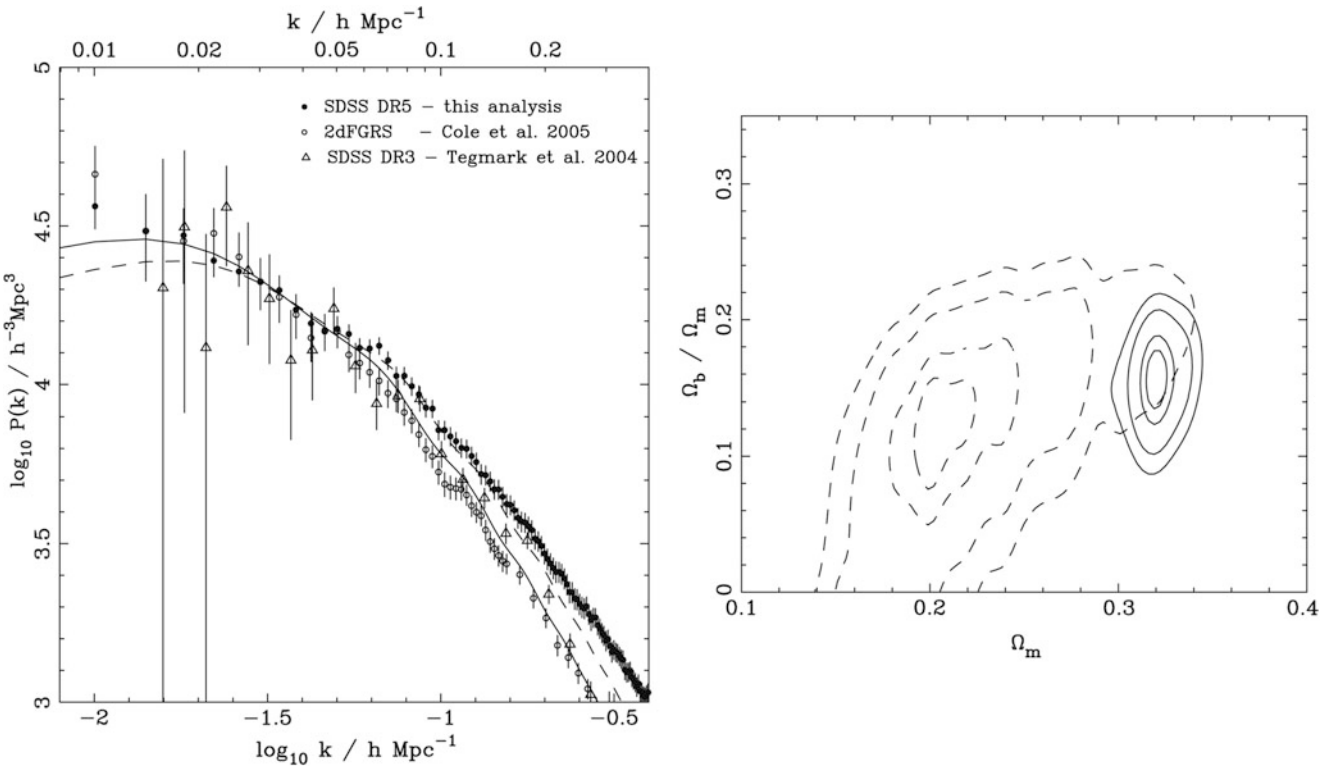


Fig. 8.6 The *left panel* shows the power spectrum of the SDSS main galaxy sample (*filled circles*), compared to that of a previous analysis with a smaller subset of SDSS galaxies (from a previous data release; *open triangles*) and that of the 2dFGRS (*open circles*). The normalization of the different power spectra were shifted so that they match on large scales (i.e., small k). The shape of the power spectra agree well on large scales, whereas there are some differences on smaller scales. The *dashed curve* shows a CDM model fit to the SDSS power spectrum over the range $0.01h\,\mathrm{Mpc}^{-1} \leq k \leq 0.06h\,\mathrm{Mpc}^{-1}$, whereas the solid curve is the best-fit CDM model over the range $0.01h\,\mathrm{Mpc}^{-1} \leq k \leq 0.15h\,\mathrm{Mpc}^{-1}$. The *right panel* shows confidence regions from such fits in the parameter plane of Ω_m and $\Omega_\mathrm{b}/\Omega_\mathrm{m}$. The *dashed* (*solid*) contours correspond to the narrow (broader) range of k-values indicated above. Source: W.J. Percival et al. 2007, *The Shape of the Sloan Digital Sky Survey Data Release 5 Galaxy Power Spectrum*, ApJ 657, 645, p. 655, 656, Figs. 13, 14. ©AAS. Reproduced with permission

the (comoving) length of the standard rod and the associated angular scale provides a measure of distance. Therefore, a measurement of the baryonic acoustic oscillations (BAOs) in the correlation of galaxies at a given redshift z can be used to determine the (comoving) angular diameter distance $f_k(z)$—which depends on the density parameters Ω_m and Ω_Λ.

In fact, the three-dimensional correlation function does not only depend on the transverse length scale which is related to the angular scale via the angular-diameter distance, but also on the separation of galaxies along the line-of-sight. The comoving distance interval corresponding to a redshift interval Δz is given by $\Delta x = c\,\Delta z/H(z)$. Since there are two transversal dimensions, and one along the line-of-sight, the distance measure that is determined best from BAOs is the geometric mean

$$D_V(z) = \left(f_k^2(z)\,\frac{c\,z}{H(z)} \right)^{1/3} . \tag{8.6}$$

The large redshift surveys 2dFGRS and SDSS allowed the first detection of these BAOs in the galaxy distribution in 2005. Figure 8.7 shows the discovery of BAOs from the SDSS, where a clear feature in the galaxy correlation function is seen at the expected length scale. The mean redshift of the galaxies from which the correlation function was determined is $z \sim 0.35$; thus, this measurement yields an estimate of the angular diameter distance to that redshift, with about a 5 % accuracy. In particular, we point out that the sound horizon at recombination is visible in the current Universe!

The power of the method depends on whether the galaxies trace the underlying matter distribution sufficiently well, so that the measured galaxy correlation function reflects the correlation function of matter. Given the large spatial scale on which BAOs are observed, the proportionality between the galaxy and matter fluctuation fields, assumed by the simple bias model (8.1), is expected to hold very well. This then turns BAOs into a straightforward, almost purely geometrical tool for measuring the geometry of our Universe.

For this reason, several surveys are underway to measure the acoustic scale as a function of redshift. Figure 8.8 shows recent measurements of BAOs over a range of red-

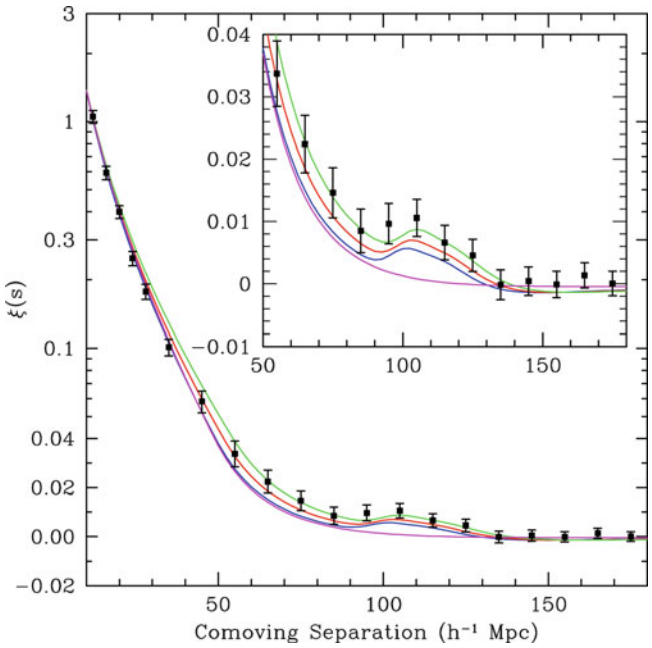


Fig. 8.7 The correlation function of galaxies, as observed in the SDSS, shows a clear indication of a secondary peak on a comoving scale of about $100h^{-1}$Mpc $\sim$ 150 Mpc. Curves show models with slightly different density parameter $\Omega_{\rm m}h^2 = 0.12, 0.13, 0.14$, with fixed baryon density of $\Omega_{\rm b}h^2 = 0.024$. The lowest, *smooth curve* is a ΛCDM model without baryons which thus shows no features due to baryonic oscillations. Source: D. Eisenstein et al. 2005, *Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies*, ApJ 633, 560, p. 563, Fig. 2, ©AAS. Reproduced with permission

shifts. Amazingly, the measurements are in perfect agreement with the cosmological parameters as determined by CMB anisotropy measurements (see Sect. 8.6). In particular, the spatial flatness of our Universe is confirmed, and any curvature is constrained to be very small, $|\Omega_{\rm m}+\Omega_\Lambda-1| \lesssim 0.01$.

8.1.5 Effect of peculiar velocities

Redshift space. The relative velocities of galaxies in the Universe are not only due to the Hubble expansion but, in addition, galaxies have peculiar velocities. The peculiar velocity of the Milky Way is measurable from the CMB dipole (see top panel of Fig. 1.21). Owing to these peculiar motions, the observed redshift of a source is the superposition of the cosmic expansion velocity and its peculiar velocity v along the line-of-sight,

$$c\,z = H_0\,D + v\ . \tag{8.7}$$

The measurement of the other two spatial coordinates (the angular position on the sky) is not affected by the peculiar velocity. The peculiar velocity therefore causes a distortion of galaxy positions in wedge diagrams, yielding a shift in the radial direction relative to their true positions. Since for most galaxies only the redshift is measurable and not the true distance D, the observed three-dimensional position of a source is specified by the angular coordinates and the redshift distance

$$s_3 = \frac{c\,z}{H_0} = D + \frac{v}{H_0}\ . \tag{8.8}$$

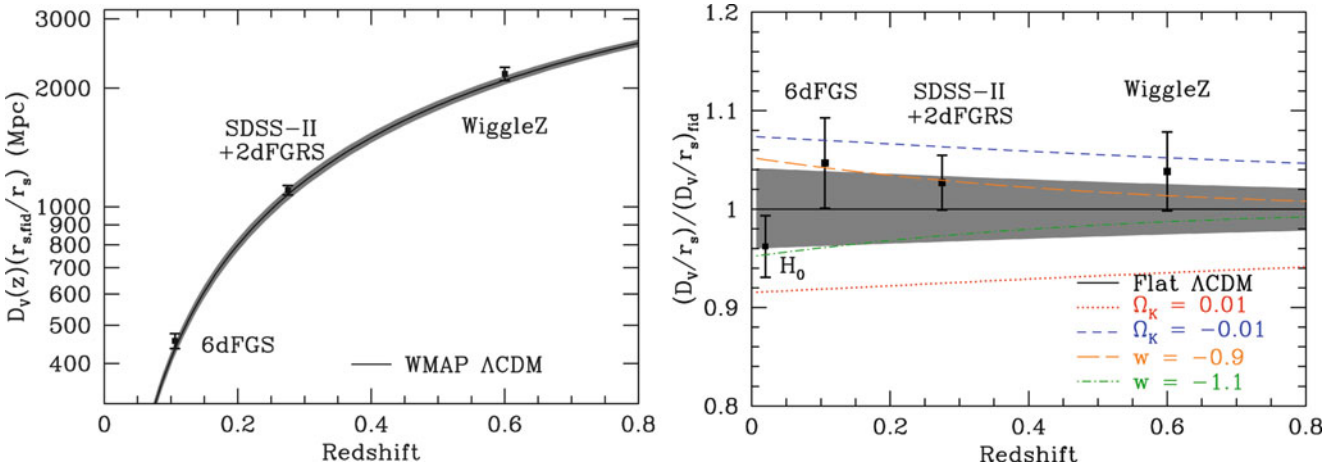


Fig. 8.8 Recent results on the distance-redshift relation from measurements of BAOs. *Left panel*: The *black line* and the *grey band* shows the best-fit cosmological model obtained from the WMAP CMB anisotropy measurements (see Sect. 8.6.5) and its 1-σ uncertainty range. The BAO distance determination from three surveys are indicated, which are located right on this best-fit model. *Right panel*: The same data are shown, now divided by the prediction of a flat ΛCDM model with parameters as determined from WMAP. Also shown are curves of this distance ratio for models with a small positive or negative curvature, or for different equations-of-state of dark energy (see Sect. 8.8 below). In particular, the BAO measurements exclude any appreciable curvature parameter of our Universe. Source: D.H. Weinberg et al. 2012, *Observational Probes of Cosmic Acceleration*, arXiv:1201.2434, Fig. 8. Reproduced by permission of the authors

The space that is spanned by these three coordinates is called *redshift space*. In particular, we expect that the correlation function of galaxies is not isotropic in redshift space, whereas the statistical isotropy of the Universe, according to the cosmological principle, yields an isotropic correlation function in real space.

Galaxy distribution in redshift space. The best known example of this effect is the "Fingers of God". To understand their origin, we consider galaxies in a cluster. They are located in a small region in space, thus all at roughly the same distance D and within a small solid angle on the sphere. In real space, they would therefore appear as a three-dimensional galaxy concentration. However, due to the high velocity dispersion, the galaxies span a broad range in s_3, which is easily identified in a wedge diagram as a highly stretched structure pointing towards us, as can be seen in, e.g., Fig. 7.2.

On larger scales, mass concentrations cause the opposite effect: galaxies that are closer to us than the center of this overdensity move towards the concentration, due to the gravitational attraction, hence away from us. Therefore their redshift distance s_3 is larger than their true distance D. Conversely, the peculiar velocity of galaxies behind the mass concentration is pointing towards us, so their s_3 is smaller than their true distance. If we now consider galaxies that are located on a spherical shell around this mass concentration, this sphere in physical space becomes an oblate ellipsoid with symmetry axis along the line-of-sight in redshift space. This effect is illustrated in Fig. 8.9.

Hence, the distortion between physical space and redshift space is caused by peculiar velocities which manifest themselves in the transformation (8.8) of the radial coordinate in space (thus, the one along the line-of-sight). Due to this effect, the correlation function of galaxies is not isotropic in redshift space. The reason for this is the relation between the density field and the corresponding peculiar velocity field. Specializing (7.25) to the current epoch and using the relation (8.1) between the density fields of matter and of galaxies, we obtain

$$\boldsymbol{u}(\boldsymbol{x}) = \beta \frac{H_0}{4\pi} \int \mathrm{d}^3 y \; \delta_\mathrm{g}(\boldsymbol{y}) \frac{\boldsymbol{y}-\boldsymbol{x}}{|\boldsymbol{y}-\boldsymbol{x}|^3} \,, \tag{8.9}$$

where we defined the parameter

$$\beta := \frac{\Omega_\mathrm{m}^{0.6}}{b_\mathrm{g}} \,. \tag{8.10}$$

This relation between the density field of galaxies and the peculiar velocity is valid in the framework of linear perturbation theory under the assumption of linear biasing. The anisotropy of the correlation function is now caused by this correlation between $\boldsymbol{u}(\boldsymbol{x})$ and $\delta_\mathrm{g}(\boldsymbol{x})$, and the degree of anisotropy depends on the parameter β. Since the correlation function is anisotropic, this likewise applies to the power spectrum.

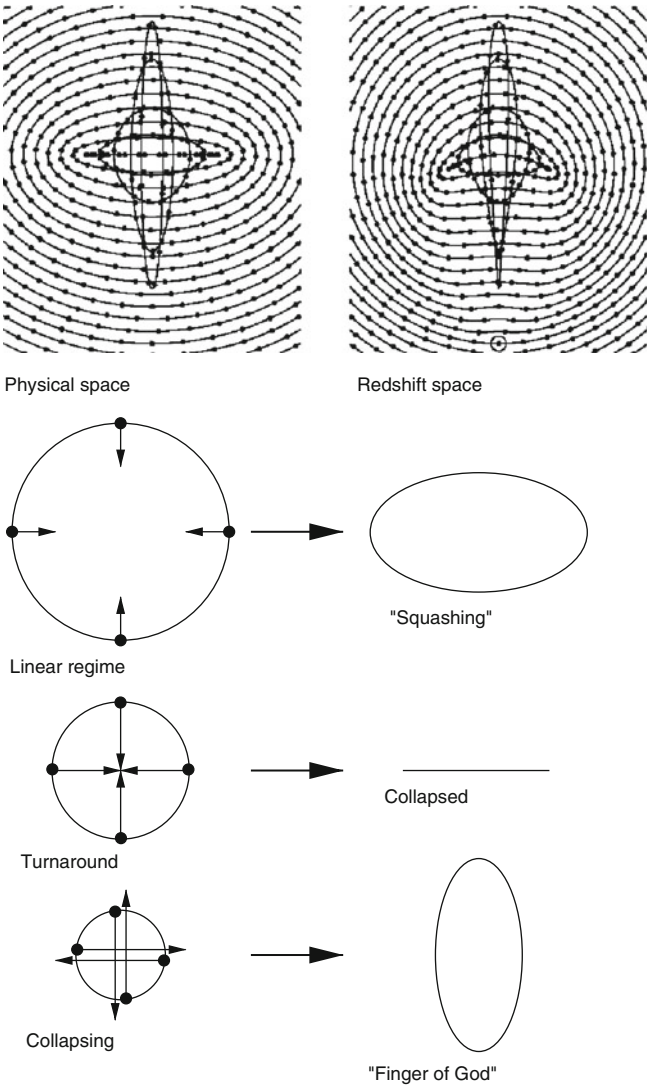


Fig. 8.9 The influence of peculiar velocities on the location of galaxies in redshift space. The *upper left panel* shows the positions of galaxies (points) in redshift space, which are in reality located on spherical shells. Galaxies connected by curves have the same separation from the center of a spherically-symmetric overdensity (such as a galaxy cluster) in real space. The explanation for the distortion in redshift space is given in the *lower panel*. On large scales, galaxies are falling into the cluster, so that galaxies closer to us have a peculiar velocity directed away from us. Thus, in redshift space they appear to be more distant than they in fact are. The inner virialized region of the cluster generates a 'Finger of God', shown by the highly elongated ellipses in redshift space directed toward the observer. Here, galaxies from a small spatial region are spread out in redshift space due to the large velocity dispersion yielding large radial patterns in corresponding wedge diagrams. In the *upper right panel*, the same effect is shown for the case where the cluster is situated close to us (*small circle* in lower center). Source: A.J.S. Hamilton 1997, *Linear Redshift Distortions: A Review*, astro-ph/9708102, Figs. 1,2

Cosmological constraints. Indeed, the anisotropy of the correlation function can be measured, as is shown in Fig. 8.10 for the 2dFGRS (where in this figure, the usual convention of denoting the transverse separation as σ and that along the line-of-sight in redshift space as π is followed). Clearly visible is the oblateness of the curves of equal correlation strength along the line-of-sight for separations $\gtrsim 10h^{-1}$ Mpc, for which the density field is still approximately linear, whereas for smaller separations the finger-of-god effect emerges. This oblateness at large separations depends directly on β, due to (8.9), so that β can be determined from this anisotropy.[4] However, one needs to take into account the fact that galaxies are not strictly following the (linear) cosmic velocity field. Due to small-scale gravitational interactions, they have a velocity dispersion σ_p around the velocity field predicted by linear theory. A quantitative interpretation of the anisotropy of the correlation function needs to account for this effect, which causes an additional smearing of galaxy positions in redshift space along the line-of-sight, as best seen in the Fingers of God which happen on non-linear scales and are therefore not covered by (8.9). Therefore, the derived value of β is related to σ_p. It is possible to determine both quantities simultaneously, by comparing the observed correlation function with models for different values of β and σ_p. From this analysis, confidence regions in the β-σ_p-plane are obtained, which feature a distinct minimum in the corresponding χ^2 function and by which both parameters can be estimated simultaneously. For the best estimate of these values, the 2dFGRS yielded

$$\beta = 0.51 \pm 0.05 \; ; \quad \sigma_p \approx 520\,\mathrm{km/s} \; . \tag{8.11}$$

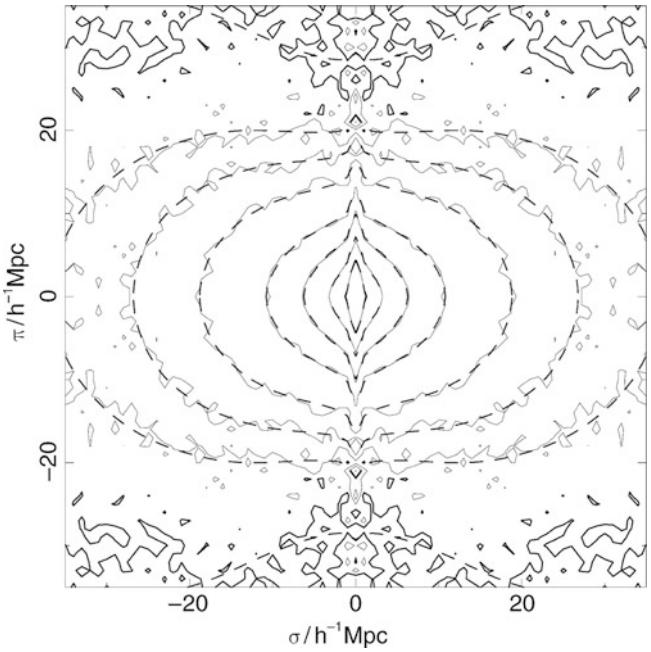


Fig. 8.10 The 2-point correlation function ξ_g, as measured from the 2dFGRS, plotted as a function of the transverse separation $\sigma = D\,\Delta\theta = c\,z\,\Delta\theta/H_0$ and the radial separation $\pi = c\,\Delta z/H_0$ in redshift space. *Solid* contours connect values of constant ξ_g. The *dashed curves* show the same correlation function, determined from a cosmological simulation that accounts for small-scale velocities. The oblateness of the distribution for large separations and the Fingers of God are clearly visible. Source: E. Hawkins et al. 2003, *The 2dF Galaxy Redshift Survey: correlation functions, peculiar velocities and the matter density of the Universe*, MNRAS 346, 78, p. 92, Fig. 22. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

8.1.6 Projected correlation function

Since the correlation function of galaxies is affected by peculiar velocities, these have to be accounted for if the measured ξ_g (or its equivalent, the power spectrum) is used to determine the power spectrum of the cosmic matter distribution. In fact, such corrections were applied in the results shown in the previous subsection.

Projected correlation function. One widely-used method to account for redshift-space distortions in correlation functions is to define the projected correlation function $w_p(r_p)$. It is based on the fact that peculiar motions affect only the line-of-sight dependence of the correlation function, whereas the transverse part of it is unchanged. In other words, if $\xi^s(r_p, \pi)$ denotes the correlation function in redshift space, where $r_p = D\,\Delta\theta = c\,z\,\Delta\theta/H_0$ is the transverse separation and $\pi = c\,\Delta z/H_0$ the line-of-sight separation, then $\xi^s(r_p, 0) = \xi_g(r_p)$, where $\xi_g(r)$ denotes the isotropic correlation function in physical space.[5] This fact allows us to 'integrate out' the redshift-space distortions, by defining

$$w_p(r_p) = \int_{-\infty}^{\infty} \mathrm{d}\pi \; \xi^s(r_p, \pi) \; . \tag{8.12}$$

The same result is obtained when we integrate the correlation function in physical space along the line-of-sight, i.e.,

$$w_p(r_p) = \int_{-\infty}^{\infty} \mathrm{d}r_3 \; \xi_g\left(\sqrt{r_p^2 + r_3^2}\right)$$

[4] In fact, one can decompose the correlation function into multipole components, such as the monopole (which is the isotropic part of the correlation function, i.e., its angle-averaged value), quadrupole (describing the oblateness) etc. The ratio of the quadrupole and monopole components in the linear regime is independent of the underlying power spectrum, and depends only on β.

[5] Note that the explicit expressions for w_p and π are valid only in the local Universe; for higher redshifts, these expressions need to be modified according to the relations given in Chap. 4.

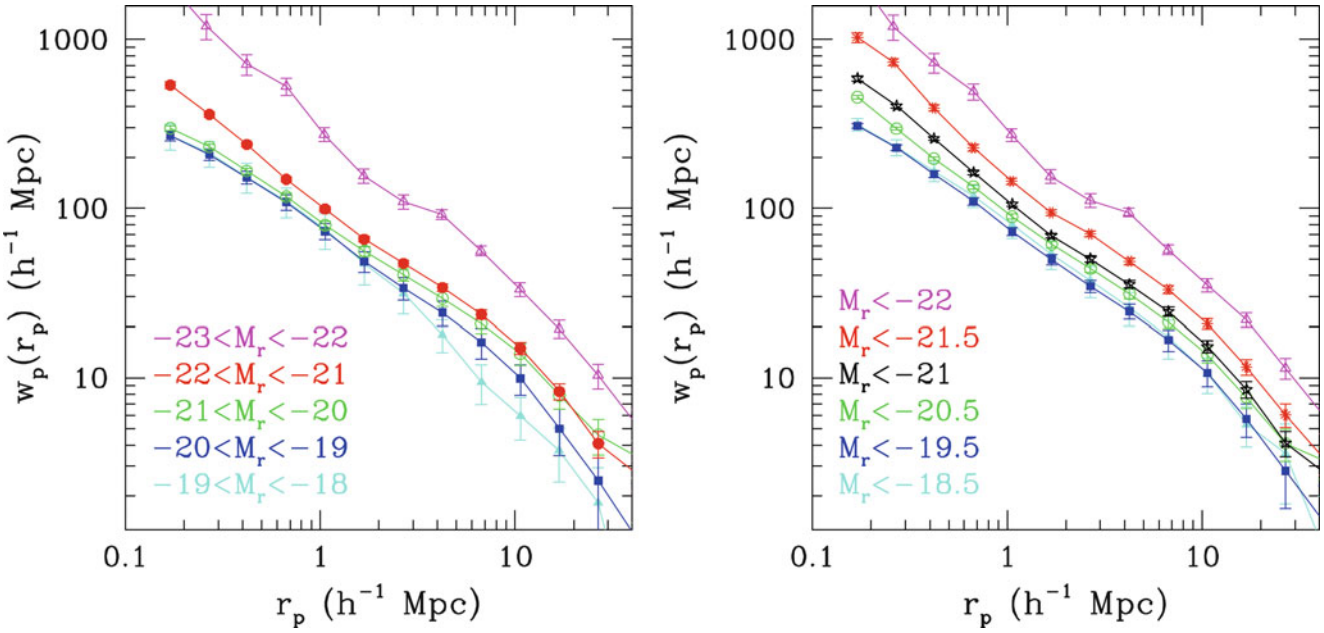


Fig. 8.11 The projected correlation function as determined from the SDSS, shown for bins in absolute r-band magnitude (*left*) and for samples with limiting absolute magnitude (*right*). Source: I. Zehavi et al. 2011, *Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity*, ApJ 736, 59, p. 10, Fig. 6. ©AAS. Reproduced with permission

$$= 2 \int_{r_p}^{\infty} \frac{\mathrm{d}r\, r}{\sqrt{r^2 - r_p^2}} \xi_g(r) \,, \tag{8.13}$$

where in the last step we changed the integration variable from r_3 to r and used that $r^2 = r_p^2 + r_3^2$, so that $\mathrm{d}r_3 = r \left(r^2 - r_p^2\right)^{-1/2} \mathrm{d}r$. Hence, $w_p(r_p)$ can be determined directly from observations using (8.12), and is intimately related to the correlation function in real space, as seen in (8.13).[6] In particular, if $\xi_g(r)$ is a power law,

$$\xi_g(r) = \left(\frac{r}{r_0}\right)^{-\gamma} \,, \tag{8.14}$$

then $w_p(r_p) = A(\gamma)\, r_0\, (r_p/r_0)^{1-\gamma}$, where the coefficient A depends on the slope. Thus from measuring w_p and fitting a power law to it, one can directly determine the slope γ and the correlation length r_0. Note that w_p has the dimension of a length.

Dependence on luminosity. In the following, we present selected results from the completed SDSS survey concerning the measurement of w_p. Figure 8.11 shows the projected correlation function, both for different bins in absolute magnitude as well as for luminosity-limited samples; for the latter, the number of galaxies in each sample is larger, so the statistical accuracy is improved, but the different curves are then no longer statistically independent. Two aspects are immediately visible: first, the correlation function increases with increasing galaxy luminosity; this increase is fairly mild at lower luminosities, but rather strong for the brighter samples. Second, over a large range, the correlation function can roughly be approximated by a power in r_p. However, for the most luminous samples, there are significant deviations from a power on smaller scales.

Galaxy bias. The implication of an increasing correlation with luminosity is a luminosity-dependent galaxy bias b_g, as we saw already in Fig. 8.5. The results shown in Fig. 8.11 can thus be used immediately to obtain the *relative* bias of galaxies as a function of luminosity. If in addition one assumes the cosmological parameters of the concordance model, one can calculate from it the corresponding $\xi(r, z)$ (where $z \sim 0.1$ is the median redshift of the SDSS galaxies), and from (8.12) the resulting projected correlation function

[6] In principle, (8.13) can be inverted to yield

$$\xi_g(r) = -\frac{1}{\pi} \int_r^{\infty} \frac{\mathrm{d}x}{\sqrt{x^2 - r^2}} \frac{\mathrm{d}w_p(x)}{\mathrm{d}x} \,;$$

however, this would require measurements of w_p out to infinite separation. But this inversion is not necessary in most applications, since for the comparison of data to model predictions, the corresponding w_p can be calculated from the models. As another technical remark, in real applications the integral in (8.12) is cut off at a sufficiently large scale, i.e., the integration is taken between $-\Delta\pi$ and $+\Delta\pi$, where $\Delta\pi$ is chosen such that the redshift-space correlation essentially has decreased to zero for $\pi > \Delta\pi$. This suppresses noise in the measurement, coming from uncorrelated foreground and background galaxies.

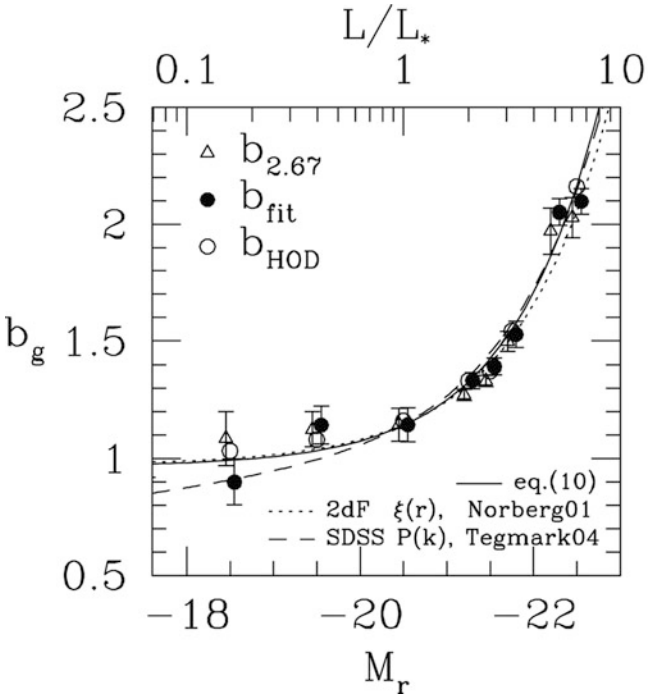


Fig. 8.12 For an assumed flat ΛCDM model with $\Omega_m = 0.3$, the galaxy bias can be obtained from the projected correlation functions of galaxies and that of dark matter. Shown is b_g as a function of absolute magnitude; the upper axis indicates the corresponding L/L^*. Three different methods for matching the observed and theoretical correlation functions were used, corresponding to the three different symbol types. The *solid curve* shows the fit (8.15), whereas the *dashed* and *dotted curves* are the same as in Fig. 8.5. Source: I. Zehavi et al. 2011, *Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity*, ApJ 736, 59, p. 10, Fig. 7. ©AAS. Reproduced with permission

of dark matter can be obtained. The ratio between the w_p of the galaxy sample and that of the dark matter then yields the square of the bias factor, now measured in absolute terms.

The resulting bias as a function of luminosity is shown in Fig. 8.12. Qualitatively, the result shown is very similar to that of Fig. 8.5, except that the current sample is slightly larger. The galaxy bias is a rather weak function of luminosity for $L \lesssim L^*$, but increases steeply for larger L. A fit to the measurements is provided by

$$b_g(L) = \left(\frac{\sigma_8}{0.8}\right)^{-1} \left[0.97 + 0.17\left(\frac{L}{L_*}\right)^{1.04}\right] , \qquad (8.15)$$

where the remaining dependence on the normalization of the matter power spectrum is indicated. This result is in a very good agreement with those found before, either from a subset of SDSS or from the 2dFGRS.

Dependence on galaxy color. The dependence on luminosity is not the only one found for the galaxy bias factor. At given luminosity, the correlation function of red galaxies has a larger amplitude than that of blue galaxies, so that the bias of red galaxies is larger than that of blue galaxies at fixed luminosity. This is seen in the two top panels of Fig. 8.13, where the projected correlation function for several luminosity bins is shown separately for red and blue galaxies. The most obvious conclusion is that the correlation is considerably stronger for red galaxies than for blue ones. This may be interpreted such that, at fixed luminosity, red galaxies live in more massive halos than blue galaxies.

Furthermore, we see that the correlation function decreases faster with radius for the red galaxies; this is also seen in Fig. 8.14 where the slope γ and the correlation length r_0 is shown, as obtained by fitting a power law to $w_p(r_p)$ in the range $4h^{-1}\,\mathrm{Mpc} \le r_p \le 30h^{-1}\,\mathrm{Mpc}$. We see that the slope γ is smaller for the blue sample, at all luminosities. In addition, for blue galaxies, γ seems to be rather independent of L. This is different for the red galaxies: whereas the slope is roughly constant for $L \gtrsim 0.5L^*$ objects, the correlation function becomes considerably steeper for the faintest red samples. This can be seen also in the top-left panel of Fig. 8.13 where at small separation, the correlation function of the faintest red galaxies is actually larger than that of the most luminous sample. The size of the error bars indicates that this measurement has a large statistical error—very low luminosity red galaxies are quite rare.

The correlation length r_0 for red galaxies is roughly constant for $L \lesssim L^*$, $r_0 \sim 7h^{-1}\,\mathrm{Mpc}$, but increases steeply for more luminous ones. In contrast, r_0 steadily increases for the blue galaxies over the luminosity range shown, though at considerably smaller values. Finally, the lower panel in Fig. 8.13 shows a monotonic behavior of the correlation properties with galaxy color.

Interpretation. Besides yielding information about cosmological parameters, the galaxy correlations shown in this section also provide us with deeper insight about biasing. We expect galaxies to be located in dark matter halos; hence, the spatial distribution of galaxies should resemble that of dark matter halos much closer than that of the mass distribution as a whole. We recall that the correlation properties of dark matter halos depend on their mass, with a halo bias b_h that increases with M. The simplest interpretation of the galaxy correlation function would be to associate galaxies with a given property (such as luminosity) with halos of mass M, such that the correlation of both agree, i.e., $b_g(L) = b_h(M(L))$, as mentioned before. In this way, one can obtain the characteristic halo mass corresponding to galaxies of a given type. This method is indeed frequently used: For example, when studying the properties of galaxies at high redshift (as will be discussed in the next chapter), one can determine their correlation function, compare that with the halo correlation at the same redshift as obtained from cosmological simulations, and thus determine that halo mass for which the correlation function is most similar. This method yields quite good estimates for the host halo mass of these high-redshift galaxies.

Applying this interpretation to the galaxy correlations discussed here, we see from Fig. 8.11 that the correlation strength increases with luminosity, so that the characteristic halo mass is a monotonic function of L—this function can be obtained by matching the galaxy bias b_g of Fig. 8.12 with the halo bias in Fig. 7.23. The fact that red galaxies are more strongly correlated than blue galaxies with the same luminosity could be interpreted such that galaxy bias is mainly a function of stellar mass—instead of optical luminosity—and that for fixed L, the stellar mass is larger in red galaxies. This effect may explain part of the clustering differences shown in Fig. 8.14, but cannot be the whole truth. In particular, this effect cannot account for the increase of the clustering strength for low-luminosity red galaxies.

A better quantitative interpretation of the results from galaxy clustering has to take into account that high-mass halos can host more than one galaxy—the best examples are galaxy clusters. An appropriate framework with which these effects can be described is offered by the halo model, which we discussed in Sect. 7.7.3. According to the results from galaxy-galaxy lensing, most of the blue galaxies are central galaxies of halos, and hence their clustering properties should very closely resemble that of halos; the monotonic behavior of the clustering strength with luminosity for blue galaxies, shown in Fig. 8.14, is consistent with this picture. Red galaxies, however, often are so-called satellite galaxies,

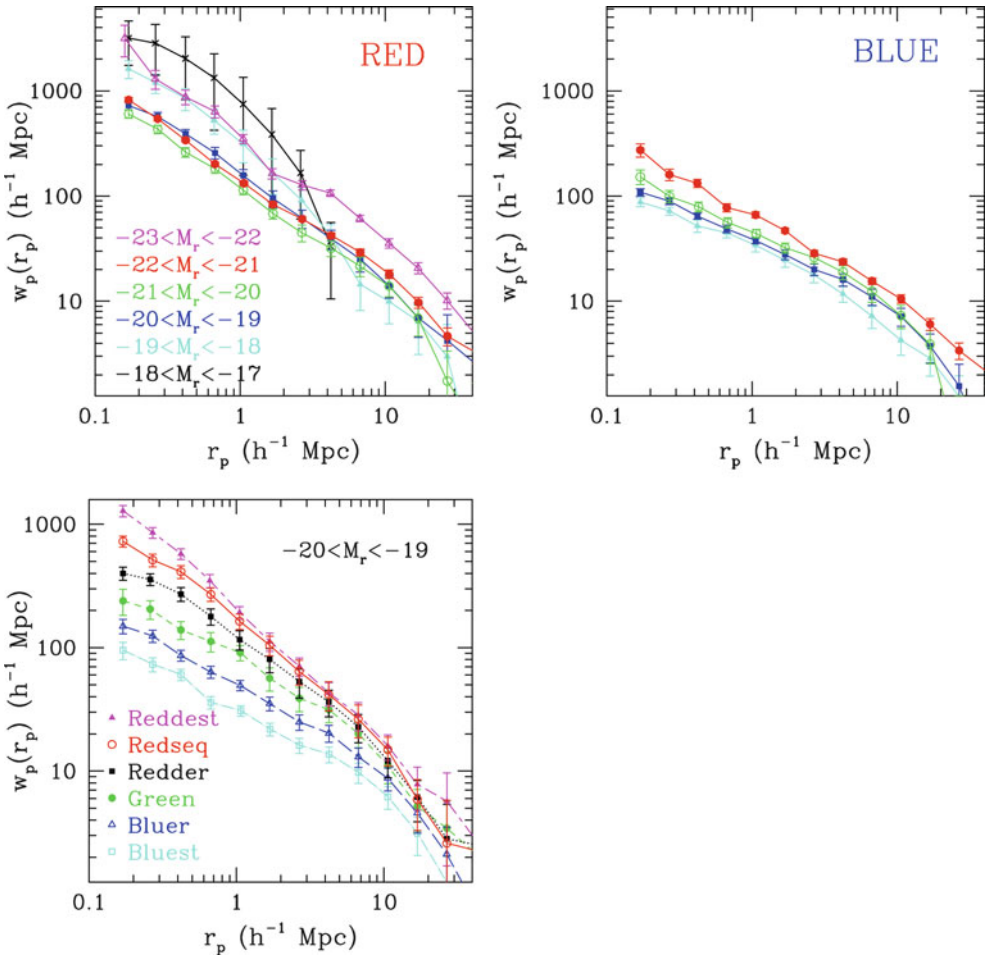


Fig. 8.13 Projected correlation function from the SDSS for red and blue galaxies of different luminosity (*top panels*) and, for a fixed luminosity range, for a finer color selection (*lower left panel*). Hence, the correlation properties of galaxies do not only depend on their luminosity, as shown in Fig. 8.11, but also strongly on their colors. Accordingly, also the galaxy bias depends on stellar population of the galaxy. Source: I. Zehavi et al. 2011, *Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity*, ApJ 736, 59, p. 18 & 19, Figs. 15, 16. ©AAS. Reproduced with permission

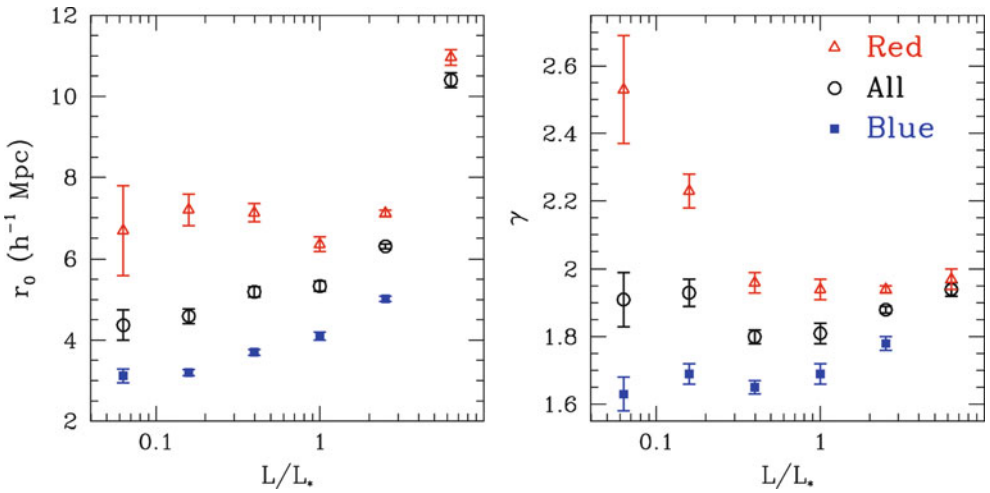


Fig. 8.14 The best fitting power-law model (8.14) for the spatial correlation function $\xi_g(r)$, applied to the data shown in Fig. 8.13, yields the correlation length (*left*) and slope (*right*) as a function of galaxy luminosity and color. *Red triangles*, *blue squares* and *black circles* show the results for red, blue, and all galaxies of the sample, respectively. Source: I. Zehavi et al. 2011, *Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity*, ApJ 736, 59, p. 18, Fig. 17. ©AAS. Reproduced with permission

and the satellite fraction increases towards lower luminosity. Thus, the clustering of red galaxies can be quite different from the clustering of halos, as it partially is due to pairs of galaxies within the same halo, providing a plausible explanation for the steepening of the correlation function towards low-luminosity red galaxies.

8.1.7 Angular correlations of galaxies

Measuring the correlation function or the power spectrum is not only possible with extensive redshift surveys of galaxies, which have become available only relatively recently. In fact, the correlation properties of galaxies can also be determined from their angular positions on the sphere. The three-dimensional correlation of galaxies in space implies that their angular positions are likewise correlated. These angular correlations are easily visible in the projection of bright galaxies onto the sphere (see Fig. 6.2).

The angular correlation function $w(\theta)$ is defined in analogy with the three-dimensional correlation function $\xi(r)$ (see Sect. 7.3.1). Considering two solid angle elements $\mathrm{d}\omega$ at $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the probability of finding a galaxy at $\boldsymbol{\theta}_1$ is $P_1 = \bar{n}\,\mathrm{d}\omega$, where $\bar{n}$ denotes the average density of galaxies on the sphere (with well-defined properties like, for instance, a minimum flux). The probability of finding a galaxy near $\boldsymbol{\theta}_1$ and another one near $\boldsymbol{\theta}_2$ is then

$$P_2 = (\bar{n}\,\mathrm{d}\omega)^2\,[1 + w(|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|)]\ , \qquad (8.16)$$

where we utilize the statistical homogeneity and isotropy of the galaxy distribution, by which the correlation function w depends only on the absolute angular separation. The angular correlation function $w(\theta)$ is of course very closely related to the three-dimensional correlation function ξ_g of galaxies. Furthermore, $w(\theta)$ depends on the redshift distribution of the galaxies considered; the broader this distribution is, the smaller is the fraction of galaxy pairs at a given angular separation which are also located close to each other in three-dimensional space, and hence are correlated. This means that the broader the redshift distribution of galaxies, the smaller the expected angular correlation.

The relation between $w(\theta)$ and $\xi_\mathrm{g}(r)$ is given by the *Limber equation*, which can, in its simplest form, be written as (see Problem 8.1)

$$w(\theta) = \int \mathrm{d}z\ p^2(z) \int \mathrm{d}(\Delta z) \times \xi_\mathrm{g}\left(\sqrt{[D_\mathrm{A}(z)\theta]^2 + \left(\frac{\mathrm{d}D}{\mathrm{d}z}\right)^2 (\Delta z)^2}\right)\ , \qquad (8.17)$$

where $D_\mathrm{A}(z)$ is the angular diameter distance (4.49), $p(z)$ describes the redshift distribution of galaxies, and $\mathrm{d}D$ specifies the physical distance interval corresponding to a redshift interval $\mathrm{d}z$,

$$\mathrm{d}D = -c\,\mathrm{d}t = -\frac{c\,\mathrm{d}a}{a\,H} \quad \Rightarrow \quad \frac{\mathrm{d}D}{\mathrm{d}z} = \frac{c}{(1+z)\,H(z)}\ .$$

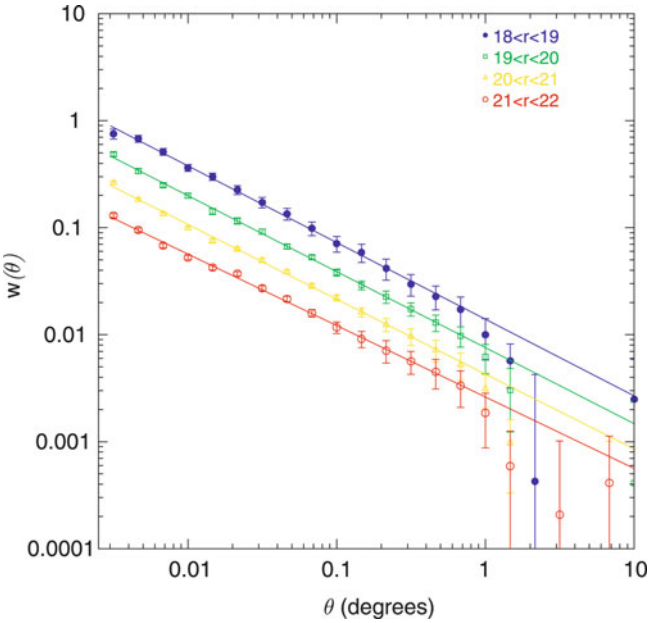


Fig. 8.15 The angular correlation function $w(\theta)$ in the four magnitude intervals $18 < r^* < 19$, $19 < r^* < 20$, $20 < r^* < 21$, and $21 < r^* < 22$, as measured from the first photometric data of the SDSS, together with a power law fit to the data in the angular range $1' \le \theta \le 30'$; the slope in all cases is very close to $\theta^{-0.7}$. Source: A.J. Connolly et al. 2002, *The Angular Correlation Function of Galaxies from Early Sloan Digital Sky Survey Data*, ApJ 579, 42, p. 45, Fig. 2. ©AAS. Reproduced with permission

Note that the galaxy correlation function in (8.17) is written in terms of proper coordinates, instead of comoving separation.

Long before extensive redshift surveys were performed, the correlation $w(\theta)$ had been measured. Since it is linearly related to ξ_g, and since ξ_g in turn is related to the power spectrum of the matter fluctuations and to the bias factor, the measured angular correlation function could be compared to cosmological models. For some time, such analyses have hinted at a small value for the shape parameter $\Gamma = \Omega_\mathrm{m} h$ of about 0.25 (see Fig. 8.3), which is incompatible with an Einstein–de Sitter model. Fig. 8.15 shows $w(\theta)$ for four magnitude intervals measured from early data of the SDSS. We see that $w(\theta)$ follows a power law over a wide angular range, which we would also expect from (8.17) and from the fact that $\xi(r)$ follows a power law.[7] In addition, the figure shows that $w(\theta)$ becomes smaller the fainter the galaxies are, because fainter galaxies have a higher redshift on average and they define a broader redshift distribution.

[7] It is an easy exercise to show that a power law $\xi(r) \propto r^{-\gamma}$ implies an angular correlation function $w(\theta) \propto \theta^{-(\gamma-1)}$ (see Problem 8.4).

8.1.8 Cosmic peculiar velocities

The relation (8.9) between the density field of galaxies and the peculiar velocity can also be used in a different context. To see this, we assume that the distance of galaxies can be determined independently of their redshift. In the relatively local Universe this is possible by using secondary distance measures (such as, e.g., the scaling relations for galaxies that were discussed in Sect. 3.4). With the distance known, we are then able to determine the radial component of the peculiar velocity by means of the redshift,

$$v = cz - H_0\, D \ .$$

We can relate the accuracy with which v can be determined to the uncertainty in distance measures, $|\delta v| = H_0\, |\delta D|$, since the spectroscopic redshifts can be assumed to have negligible uncertainty. Hence,

$$|\delta v| = H_0\, \frac{|\delta D|}{D}\, D \ .$$

Thus, to measure values of v with an accuracy of $|\delta v|$, then for a given relative accuracy of distance measurements, there is a maximum distance where this accuracy can be achieved. Hence, the peculiar velocity field can be determined only relatively locally. For example, assuming $|\delta D|/D \sim 0.1$, which is typical for distance determinations from the Tully–Fisher relation of individual galaxies, and requesting $|\delta v| \leq 500\,\mathrm{km/s}$, we find that $D \leq 5000\,\mathrm{km/s}/H_0 \sim 50h^{-1}\,\mathrm{Mpc}$. In most cases, these measurements are carried out for groups of galaxies which then all have roughly the same distance; in this way, the measurement accuracy of their common (or average) distance is improved and thus the range of this method is increased.

Equation (8.9) now allows us to predict the peculiar velocity field from the measured density field of galaxies, which can then be compared with the measured peculiar velocities—where this relation depends on β [see (8.10)]. Therefore, we can estimate β from this comparison. The inverse of this method is also possible: to derive the density distribution from the peculiar velocity field, and then to compare this with the observed galaxy distribution.[8] Such a comparison is displayed in Fig. 8.16.

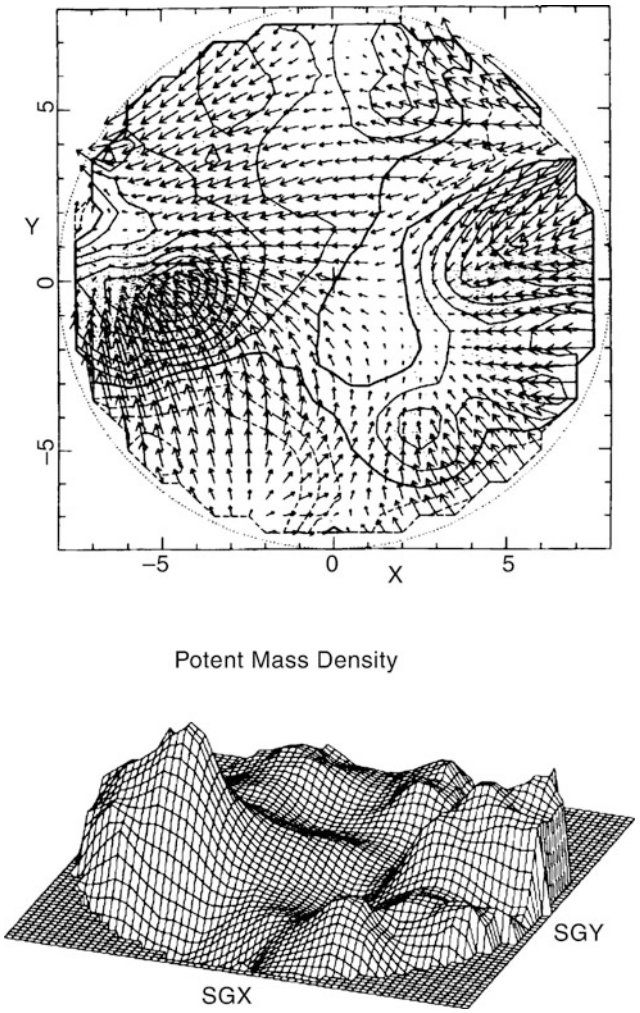


Fig. 8.16 The peculiar velocity field (*top panel*) and the derived density field (*bottom panel*) in our neighborhood. The distances here are specified as expansion velocities in units of 1000 km/s. The mass concentration on the left, towards which the velocity vectors are pointing, is the Great Attractor (see text), and on the right is the Pisces-Perseus supercluster. By comparing this reconstructed mass distribution with the distribution of galaxies, β can be determined. Early analyses of this kind resulted in relatively large values of β, whereas more recent results show $\beta \sim 0.5$. Since the bias factor may be different for the various types of galaxies, β may depend on the type of observed galaxies as well. For instance, one finds that IRAS galaxies have a lower β, thus a higher b_g, than galaxies which are selected optically. Source: Dekel, A. 1994, *Dynamics of Cosmic Flows*, ARA&A 32, 371, Fig. 3, p. 389. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 32 ©1994 by Annual Reviews www.annualreviews.org

Measurements of the peculiar velocity field in the mid-1980s led to the conclusion that an unseen mass concentration, i.e., one that could not, at that time, be identified with a large concentration of galaxies, was having a significant effect on the local velocity field. This mass concentration (which was termed the 'Great Attractor') was located roughly in the direction of the Galactic Center, which is the reason why it was not directly observable.

[8] At first sight this seems to be impossible, since only the radial component of the peculiar velocity can be measured—proper motions of galaxies are far too small to be observable. However, in the linear regime we can assume the velocity field to be a gradient field, $\boldsymbol{u} = \nabla\psi$; see Sect. 7.2.3. The velocity potential ψ can be obtained by integrating the peculiar velocity, $\psi(\boldsymbol{x}) - \psi(\boldsymbol{0}) = \int_{\boldsymbol{0}}^{\boldsymbol{x}} \mathrm{d}\boldsymbol{l} \cdot \boldsymbol{u}$, where the integral is taken over a curve connecting the observer at $\boldsymbol{0}$ to a point $\boldsymbol{x}$. Choosing radial curves, only the radial component of the peculiar velocity enters the integral. Therefore, this component is sufficient, in principle, to construct the velocity potential ψ, and therefore the three-dimensional velocity field.

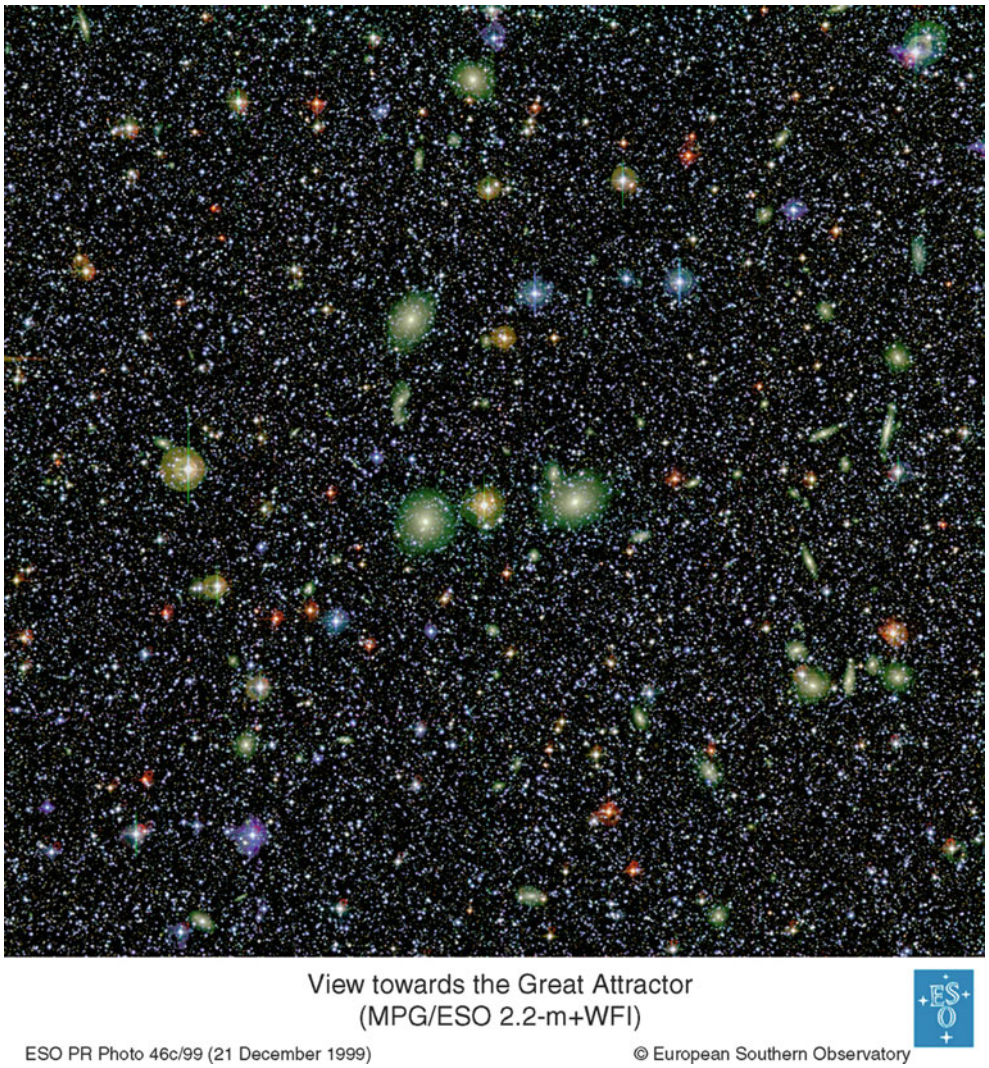


Fig. 8.17 An optical image taken in the direction of the Great Attractor. This image has a side length of half a degree and was observed by the WFI at the ESO/MPG 2.2-m telescope on La Silla. The direction of this pointing is only $\sim 7^\circ$ away from the Galactic disk. For this reason, the stellar density in the image is extremely high (about 200 000 stars can be found in this image) and, due to extinction in the disk of the Milky Way, much fewer faint galaxies at high redshift are found in this image than in comparable images at high Galactic latitude. Nevertheless, a large number of galaxies are visible (greenish), belonging to a huge cluster of galaxies (ACO 3627, at a distance of about 80 Mpc), which is presumably the main contributor to the Great Attractor. Source: European Southern Observatory

X-ray cluster samples are much less affected by Galactic absorption than optically selected clusters, and therefore provide a much clearer view of the mass distribution surrounding the Local Group, including the Zone of Avoidance (see Sect. 2.1). In recent years, based on such X-ray selected clusters, the simple picture of the Great Attractor has been modified. In fact, at the proposed distance to the Great Attractor of $\sim$ 80 Mpc, the matter density seems to be considerably smaller than originally thought. However, behind the Great Attractor there seems to be a significant overdensity of clusters at larger distances (see Fig. 8.17).

The velocity dipole of the galaxy distribution. A related aspect of these studies is the question of whether the observed peculiar velocity of the Local Group, as determined from the dipole anisotropy of the CMB, can be traced back to the matter distribution around us. We would expect to find a related dipole in the matter distribution which caused an acceleration of the Local Group to the observed value of the peculiar velocity of $627 \pm 22\,\mathrm{km/s}$ towards the direction $\ell = 273^\circ \pm 3$, $b = 29^\circ \pm 3^\circ$ in Galactic coordinates (this value is obtained from the direct measurement of the dipole velocity in the rest-frame of the Sun, to which the motion of the Sun relative to the Local Group rest-frame is added).

In principle, this question can be answered from photometric galaxy surveys alone. We found a relation (8.9) between the fractional galaxy overdensity δ_g and the peculiar velocity $\boldsymbol{u}(\boldsymbol{x})$ which we can specialize to the point of origin $\boldsymbol{x} = \boldsymbol{0}$. This relation is based on the assumption of linear biasing. A galaxy at distance D contributes to the peculiar velocity by an amount $\propto m/D^2$, where m is its mass. If we assume that the mass-to-light ratio of galaxies are all the same, then $m \propto L$ and the contribution of this galaxy to $\boldsymbol{u}$ is $\propto L/D^2 \propto S$. Hence, under these simplifying assumptions the contribution of a galaxy to the peculiar velocity depends only on its observed flux.

To apply this simple idea to real data, we need an all-sky map of the galaxy distribution. This is difficult to obtain,

due to the presence of extinction towards the Galactic plane. However, if the galaxy distribution is mapped at infrared wavelengths, these effects are minimized. It is therefore not surprising that most of the studies on the dipole distribution of galaxies concentrate on infrared surveys. The IRAS source catalog still provides one of the major catalogs for such an analysis. More recently, the Two-Micron All Sky Survey (2MASS) catalog provided an all-sky map in the near-IR (see Fig. 1.52) which can be used as well. The NIR also has the advantage that the luminosity at these wavelengths traces the mass of the stellar population of a galaxy quite well, in contrast to shorter wavelength for which the mass-to-light ratio among galaxies varies much more. The results of these studies is that the dipole of the galaxy distribution lies within $\sim 20^\circ$ of the CMB dipole. This is quite a satisfactory result, if we consider the number of assumptions that are made in this method. The amplitude of the expected velocity depends on the factor $\beta = \Omega_{\rm m}^{0.6}/b_{\rm g}$. Thus, by comparing the predicted velocity from the galaxy distribution with the observed dipole of the CMB this factor can be determined, yielding $\beta = 0.49 \pm 0.04$, which is seen to be in excellent agreement with the estimate (8.11) from redshift-space distortions.

Supplementing the photometric surveys with redshifts allows the determination of the distance out to which the galaxy distribution has a marked effect on the Local Group velocity, by adding up the contributions of galaxies within a maximum distance from the Local Group. Although the detailed results from different groups vary slightly, the characteristic distance turns out to be $\sim 150h^{-1}$ Mpc, i.e., larger than the distance to the putative Great Attractor. In fact, earlier results suggested a considerably smaller distance, which was one of the reasons for postulating the presence of the Great Attractor.

8.2 Cosmological parameters from clusters of galaxies

Being the most massive and largest gravitationally bound and relaxed objects in the Universe, clusters of galaxies are of special value for cosmology. In this section, we will explain various methods by which cosmological parameters can be derived from observations of galaxy clusters.

8.2.1 Cluster abundance

In Sect. 7.5.2, we demonstrated that it is possible, for a given cosmological model, to predict the number density of halos as a function of mass and redshift. If we make the assumption that the massive end of this halo population can be identified with galaxy clusters, then we can compare the observed number density of galaxy clusters with these theoretical results and in this way constrain the cosmological parameters on which the expected halo abundance depend.

We saw in Chap. 6 that the selection of clusters by their X-ray emission has long been viewed as the most reliable method of finding clusters, due to the $n_{\rm e}^2$-dependence of the X-ray emissivity. Hence, we will concentrate much of the discussion on the X-ray cluster catalogs described in Sect. 6.4.5 for a comparison of the halo number density with model predictions. However, the selection of clusters from their Sunyaev–Zeldovich effect is a quickly evolving technique, and significant cosmological constraints have been derived from those as well. Finally, modern optical cluster catalogs also play a significant role in cluster cosmology.

Scaling relations and their calibration. In order to perform this comparison, the masses of clusters need to be determined. We discussed various methods of cluster mass determination in Chap. 6. Since a detailed mass determination is possible only for individual clusters, but not for a large sample (which is required for a statistical comparison with model predictions), one usually applies the scaling relations discussed in Sect. 6.5. In other words, one needs to find a relation between an observable X and the cluster mass; the best choice for X is one where, for a given mass M, the scatter of X is smallest, so that X becomes a good mass proxy. In particular, the relation (6.56) between X-ray temperature, X-ray luminosity, and virial mass plays a central role, as does the relation (6.60) between the Y-parameter (either determined from X-ray or SZ-measurements) and the cluster mass. The scaling relations are then calibrated on clusters for which detailed mass estimates were performed.

It should be stressed that the estimated mass of an individual cluster is bound to be uncertain. Weak lensing mass estimates suffer from various effects, such as the sensitivity to the orientation of triaxial mass distributions and the mass structures between us and the cluster, and between the cluster and the source population which is employed for determining the shear. Similarly, the mass obtained from X-ray data using hydrostatic equilibrium is probably problematic, due to additional forces, e.g., from cluster rotation, magnetic field pressure or, in particular, bulk random gas motion which acts like an additional pressure. Either of these methods for the mass determination of individual clusters thus has an intrinsic scatter, but this can be accounted for in the calibration of the scaling relations. In addition, though, these method also may have a non-vanishing bias, which would lead to systematic shifts in the scaling relations.

Statistical mass calibration. There are complementary methods for the calibration of mass-observable relations, which no longer aim at estimating the mass of individual

clusters, but the mean mass of a set of clusters with similar properties.

Thus consider a set of clusters whose observable X is confined to a narrow interval. If X is a good mass proxy, then the masses of these cluster should also be quite similar. One method to get the mean mass of these clusters consists in determining their spatial correlation function $\xi_{\rm cl}(r;X)$ and compare that with the dark matter correlation function $\xi(r)$ which is calculated from the cosmological model. The ratio of these two correlations is the square of the bias $b_{\rm cl}(X)$ for these clusters. If we now identify clusters with dark matter halos, then we can find for a halo mass M such that the halo bias $b_{\rm h}(M)$ equals $b_{\rm cl}(X)$; this method, which is essentially the same as that mentioned before for galaxy halo masses, then yields the mass as a function of the observable X. In the literature, this method is often called mass self-calibration.

You may have noticed that this method has to assume a cosmological model to determine M, whereas the estimated values of M should be used to determine the cosmological parameters. For that reason, the method must be used iteratively: assume a model, get the mass estimates, use them to get the best-fitting cosmological parameters, and use those to start the loop again; it converges quickly. In practice, one also cross-correlates cluster samples with different values of X, say X_1 and X_2; their correlation function is then expected to behave like $\xi_{\rm cl}(r;X_1,X_2) = b_{\rm h}(X_1)b_{\rm h}(X_2)\,\xi(r)$. In this way, one can also estimate the scatter in the mass for clusters at fixed X.

A second possibility for mass calibration is a weak lensing analysis of a set of clusters, often also called cluster stacking. In this method, one statistically superposes the weak lensing signal of a set of clusters, to find the mean lensing signal of this set, and from that to obtained the mean cluster mass. We have discussed this method in Sect. 7.7.4 and shown that it yields scaling relations between observables, such as luminosity or richness, and the cluster mass (see Fig. 7.30). We want to point out that this method is different from obtaining mass estimates from weak lensing studies of individual clusters, and combining their results. Since the weak lensing mass estimate of an individual cluster is obtained by fitting an NFW-profile to its shear measurement, the resulting mass is not linear in the shear signal (which causes the aforementioned bias). In other words: the mean of the mass determined from a weak lensing mass estimate of individual clusters is not the mass obtained from the mean weak lensing signal of the set of clusters.

A third possibility to calibrate the mass-observable relation is through numerical simulations. With increasing sophistication of numerical methods, it becomes possible to include a rich spectrum of physical processes in the simulations, as we will describe in more detail in Sect. 10.6.1; in particular, one can follow the evolution of the intracluster gas. From such simulated clusters, one can generate synthetic data, say the X-ray emission from the gas, which can then be analyzed in the same way as observational data. This process then yields a mass estimate, which can be compared to the true mass known from the simulation. The ratio of these two then estimates the bias of the method for mass determination. However, it must be pointed out that the physical processes which are relevant for the evolution of the cluster gas are not all well understood (such as the impact of heating from a central AGN, which suppresses the cooling flows), and thus this bias estimate carries some uncertainties.

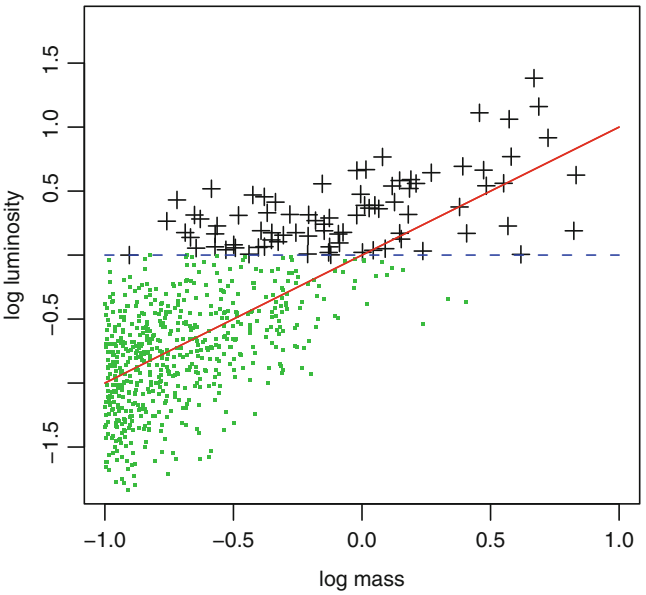


Fig. 8.18 Illustration of selection effects in determining scaling relations. The *green* points show the distribution of clusters which scatter around the mean scaling relation, shown as *red solid line*. A luminosity-limited sample finds only objects above the *horizontal dashed line*, shown here as crosses. A fit through the observed distribution will yield a rather different 'scaling relation', being much flatter than the true one. The zero point on both axes was chosen such that the intersection point between the *red line* and the *dashed black line* lies at the origin. Source: S.W. Allen et al. 2011, *Cosmological Parameters from Observations of Galaxy Clusters*, ARA&A 49, 409, p. 426, Fig. 5. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 49 ©2011 by Annual Reviews www.annualreviews.org

Selection effects. A further important issue in cluster statistics, like in many other fields of astronomy, is that of selection biases. We will illustrate this with a simple example (see also Fig. 8.18): Suppose that there exists a (power-law) scaling relation between X-ray luminosity (or another observable, such as $Y_{\rm SZ}$) and cluster mass, however with a non-negligible scatter. Since surveys characteristically are flux limited (and thus, for a fixed redshift, luminosity limited), it is more likely to detect those objects which lie above the mean luminosity-mass relation. As a consequence, if a power law is fitted to the observed objects in the survey, its slope is biased relative to the underlying scaling

relation. This is a particular example of the Malmquist bias. Clearly, this selection effect needs be accounted for by proper modeling.[9]

Normalization of the power spectrum. The comparison of the number density of observed clusters to the halo density in cosmological models can be performed either in the local Universe or at a range of redshifts. In the former case, one primarily obtains the normalization of the power spectrum from this comparison, hence σ_8, for a given matter density parameter Ω_m. More precisely, the number density of halos depends on the combination $\sigma_8 \Omega_\mathrm{m}^\alpha$, where the exact value of the exponent α depends on the mass range of the halos that are considered. The analysis of cluster catalogs like the ones compiled from the ROSAT All-Sky Survey (RASS) yields a value of about

$$\sigma_8 \left(\frac{\Omega_\mathrm{m}}{0.3} \right)^{0.45} \approx 0.74 \pm 0.03 \,, \tag{8.18}$$

where the uncertainty in this value mainly comes from the uncertainties in the calibration of the scaling relations. It should be pointed out that clusters are particularly suitable for determining σ_8, since the mass within a sphere of comoving radius $8h^{-1}\,\mathrm{Mpc}$ and the mean matter density of the Universe corresponds to a typical cluster mass, as shown in Problem 7.3.

Breaking the Ω_m-σ_8 degeneracy. The degeneracy between Ω_m and σ_8 can be broken either by considering the cluster abundance over a large range in mass, or by studying the redshift evolution of the number density of clusters. As we have seen in Sect. 7.2.2, the growth factor D_+ of the density perturbations depends on the cosmological parameters. For a low-density universe, the growth factor D_+ at high redshift is considerably larger than in an Einstein–de Sitter universe (see Fig. 7.3). Hence, the expected number density of clusters at high redshift is considerably smaller in an EdS model than in one of low density, for a fixed local number density of clusters. Indeed, in an EdS universe virtually no high-mass clusters are expected at $z \gtrsim 0.5$ (see Fig. 7.8), whereas the evolution of the halo number density is significantly weaker for cosmological models with small Ω_m. The fact that very massive clusters have been discovered at redshift $z > 0.5$ is therefore incompatible with a cosmological model of high matter density.

Figure 8.19 illustrates the sensitivity of cluster counts on the cosmological models. In both panels, the comoving number density of clusters is shown for two redshift intervals, together with the prediction of the halo abundance. In the left panel, the ΛCDM model was chosen, while the right panel assumes a low-density open universe. The low-redshift data are equally well fitted by both models—the local cluster abundance is essentially independent of Λ if σ_8 is properly adjusted. However, there are substantial differences for the high-redshift data; the open model simply fails to give an acceptable fit.

Therefore, the evolution of cluster abundance provides more constraints on the model parameters; in particular, the Ω_m-σ_8 degeneracy can be broken. Different samples of X-ray clusters consistently obtained, by assuming a flat cosmology,

$$\sigma_8 = 0.81 \pm 0.04 \,; \quad \Omega_\mathrm{m} = 0.26 \pm 0.08 \,. \tag{8.19}$$

Results from optical clusters. The availability of wide-field multi-band imaging surveys has led to a revival of optical cluster selection. Of particular interest in recent years have been cluster samples selected from large redshift surveys. In Sect. 6.2.4 we have discussed the maxBCG cluster catalog in quite some detail, which is based on the concentration of red galaxies on the sky and in color space, the latter being motivated by the tight color-magnitude relation of red galaxies in cluster—the cluster red sequence. The redshift is estimated from the color of the red sequence, and should be quite accurate in the range $0.1 \le z \le 0.3$. This redshift range is then chosen for the cosmological application described here. The corresponding scaling relation between optical luminosity, or richness, and cluster mass were obtained by weak lensing methods, as described in Sect. 7.7.4.

Studying the abundance of clusters with richness $N_{200} \ge 12$ as a function of mass and comparing this with the abundance of dark matter halos in a flat universe, the constraint

$$\sigma_8 \left(\frac{\Omega_\mathrm{m}}{0.3} \right)^{0.41} \approx 0.77 \pm 0.03 \tag{8.20}$$

was obtained. In addition, the large mass range of the cluster sample allowed the breaking of the Ω_m-σ degeneracy, i.e., to estimate both parameters separately:

$$\sigma_8 = 0.80 \pm 0.07 \,; \quad \Omega_\mathrm{m} = 0.28 \pm 0.07 \,. \tag{8.21}$$

We see that these results are in very good agreement with those obtained from X-ray clusters. This fact is reassuring, since the optical and X-ray analyses are quite different; in particular, they are subject to different potential systematic effects.

Extreme clusters. Due to the exponentially decreasing mass function of dark matter halos, seen in Fig. 7.10,

[9] For example, for an assumed slope of the scaling relation and the scatter of luminosity around it, one can predict the distribution in the observed sample, and compare this prediction with the real sample. By minimizing the differences between the two through varying slope and scatter, the true scaling relation can be recovered.

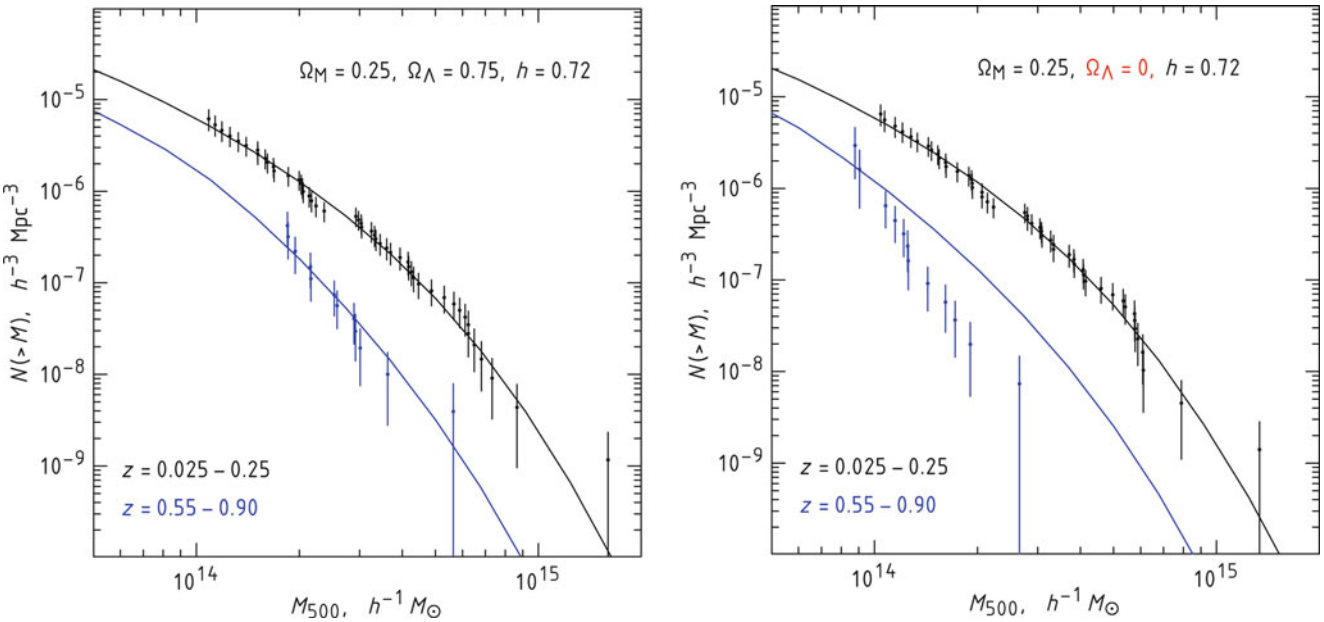


Fig. 8.19 The cluster mass function in two redshift bins, for a flat ΛCDM model (*left*) and an open $\Lambda = 0$ model (*right*). In both cases, the value of σ_8 was chosen as to best fit the low-redshift number density. The low-redshift sample was taken as the 49 brightest clusters from the ROSAT All-Sky Survey, the high-redshift sample for this analysis was taken from the 400 deg^2 ROSAT serendipitous survey; from the latter, only those with $z \geq 0.55$ are included in this figure. Both samples were reobserved with Chandra, yielding much better sensitivity. The mass used in this comparison is M_{500}, i.e., the mass of the sphere around the cluster center inside of which the mean density is 500 times the critical density of the Universe at the corresponding redshift. Note that both panels are based on the same data, but the plotted points are quite different. This is because the conversion of observables to a mass, and the conversion of number counts to a comoving number density, both depend on the adopted cosmological model (for example, the volume element corresponding to a given Δz is cosmology-dependent, as shown in Fig. 4.13). Source: A. Vikhlinin et al. 2009, *Chandra Cluster Cosmology Project III: Cosmological Parameter Constraints*, ApJ 692, 1060, p. 1064, Fig. 2. ©AAS. Reproduced with permission

there is, at any given redshift, an upper bound on cluster masses—essentially given by the condition that the mass-dependent number density times the observable volume at a given redshift is about unity. The exponential decline at the high-mass end of the abundance then implies that the expected number of clusters with twice that mass should be essentially zero. Hence, if one finds such extremely massive clusters, one could in principle falsify the parameters of the cosmological model with which the expected abundance was calculated. Since the cut-off mass decreases with increasing redshift, one might find such extreme clusters in particular at higher redshifts.

Indeed, there have been several claims for the detection of very massive high-redshift clusters whose mass is so large that the probability for finding one of them in our Universe is incredibly small (see Fig. 8.20 for an example)—and hence there may be a serious conflict with our standard cosmological model. Whereas these detections are currently subject to intense research, one must keep in mind a number of issues: First, these clusters are typically at redshifts $z \gtrsim 0.5$, and hence their mass determination—either from the intracluster gas through X-ray or Sunyaev–Zeldovich observations, or from weak lensing—is difficult. In most cases, lowering the estimated mass by $\sim 30\,\%$ reduces the discrepancy significantly. Second, the high-redshift clusters are more likely not to be in dynamical equilibrium, and hence methods for mass determinations based on equilibrium assumptions, such as the hydrostatic mass estimate from X-rays, may yield significantly wrong results. Indeed, these clusters show a rather complex morphology, for example, their critical curves for strong lensing effects are very highly elongated. Again, this is no surprise, since the most massive cluster at a given epoch is so massive that its mass at some earlier epoch must have been significantly smaller; thus, it has just attracted a significant part of its mass, presumably through a merging event, which explains the complex structure. Third, the accuracy of the abundance as determined from simulations (as in Fig. 7.10) is also limited—once the total number of dark matter halos in the simulation volume approaches unity, the mass function becomes rather uncertain.

Taking these effects into account, one concludes that these 'monsters' do not provide a serious challenge to the standard model of cosmology yet.

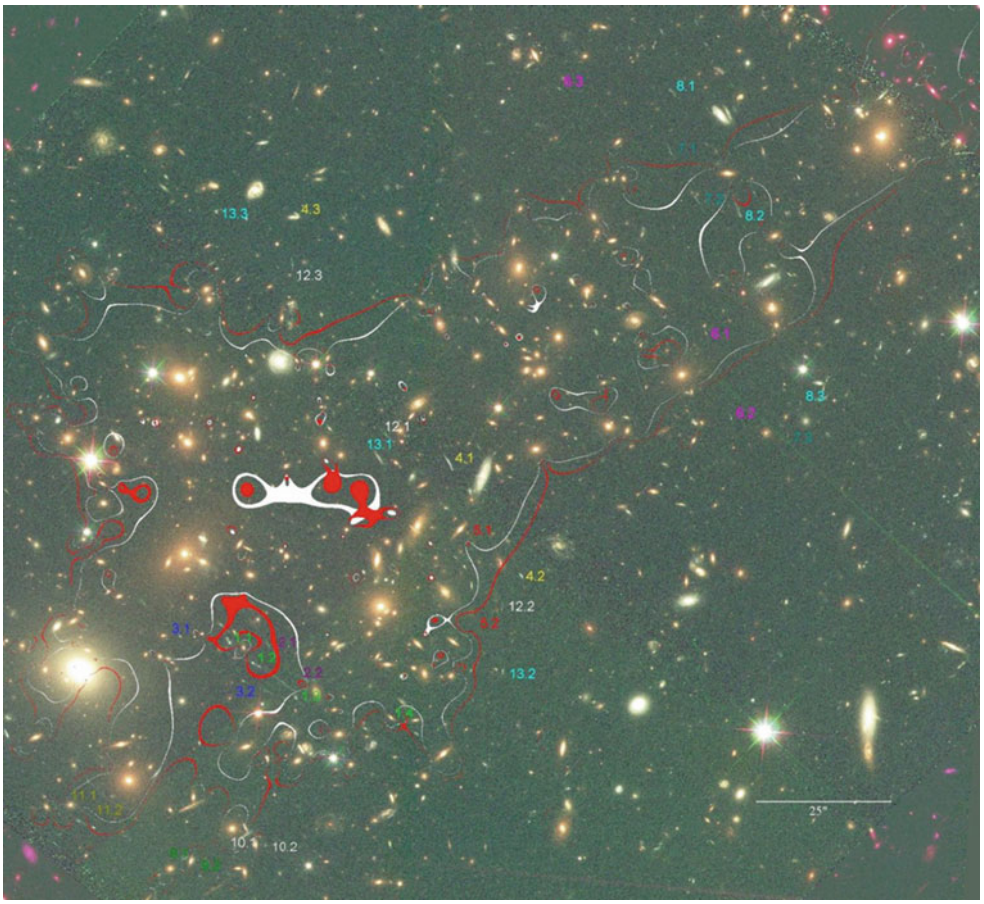


Fig. 8.20 The cluster MACS J0717.5 + 3745 at $z = 0.546$ is one of the strongest lensing clusters found at $z > 0.5$. Shown here is the inner region of that cluster, as imaged by the HST, with several multiple image systems indicated. The *white* (*red*) curve is the critical curve of this cluster for sources at $z = 2.5$ ($z = 4$). As can be seen, the critical curve is highly elongated, which renders a determination of 'the' Einstein radius somewhat ambiguous. The virial mass of this cluster is estimated from weak and strong lensing to be $(2.8 \pm 0.4) \times 10^{15} M_\odot$; clusters of this mass should be extremely rare at $z > 0.5$. The white scale bar near the lower right corner has a length of 25″. Source: A. Zitrin et al. 2009, *The Largest Gravitational Lens: MACS J0717.5 + 3745 ($z = 0.546$)*, ApJ 707, L102, p. L104, Fig. 1. ©AAS. Reproduced with permission

8.2.2 Mass-to-light ratio

On average, the mass-to-light ratio of cosmic objects seems to be an increasing function of their mass. In Chap. 3 we saw that M/L is smaller for spirals than for ellipticals, and furthermore that for ellipticals M/L increases with mass. In Chap. 6, we argued that galaxy groups like the Local Group have $M/L \sim 100h$, and that for galaxy clusters M/L is several hundreds, where all these values are quoted in Solar units. We conclude from this sequence that M/L increases with the length- or mass-scale of objects. Going to even larger scales—superclusters, for instance—M/L seems not to increase any further, rather it seems to approach a saturation value (see Fig. 8.21).

Thus, if we assume the M/L ratio of clusters to be characteristic of the average M/L ratio in the Universe, the average mass density of the Universe ρ_m can be calculated from the measured luminosity density $\mathcal{L}$ and the M/L ratio for clusters,

$$\rho_\mathrm{m} = \left\langle \frac{M}{L} \right\rangle \mathcal{L} \, .$$

Here, L and $\mathcal{L}$ refer to a fixed frequency interval, e.g., to radiation in the B-band; $\mathcal{L}$ can be measured, for instance, by determining the local luminosity function of galaxies, yielding

$$\Omega_\mathrm{m} \approx \frac{\langle M/L \rangle_\mathrm{B}}{1200\, h} \, . \tag{8.22}$$

Since several methods for determining cluster masses now exist (see Sect. 6), and since their luminosity L is measurable as well, (8.22) can be applied to clusters in order to estimate Ω_m. Typically, this results in $\Omega_\mathrm{m} \sim 0.2$, a value for Ω_m which is slightly smaller than that obtained by other methods. However, this method is presumably less reliable than the other ones described in this section: $\mathcal{L}$ is not easily determined (e.g., the normalization of the Schechter luminosity function has been revised considerably in recent years, and its accuracy is probably not better than $\sim 20\,\%$), and the M/L ratio in clusters is not necessarily representative. For instance, the evolution of galaxies in a cluster is different from that of a 'mean galaxy', as best seen in the large fraction of red galaxies in clusters.

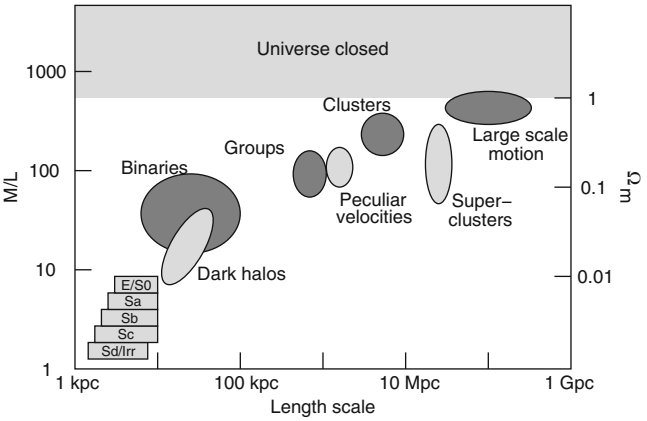


Fig. 8.21 The mass-to-light ratio M/L seems to be a function of the length- or mass-scale of cosmic objects. The luminous region in spirals has $M/L \sim 3$ (all values in Solar units of $M_\odot/L_\odot$), whereas that of ellipticals has $M/L \sim 10$. However, galaxies have a dark matter halo, so that the true mass of galaxies, and thus their M/L, is much larger than that which is measured in their visible region. Masses can also be estimated from the dynamics of galaxy pairs, typically yielding $M/L \sim 50$ for galaxies, including their dark halo. Galaxy groups and clusters have an even higher M/L ratio, hence they are particularly strongly dominated by dark matter, reaching $M/L \sim 250$. If the M/L ratio in clusters corresponds to the average M/L in the Universe, it is possible to determine the matter density in the Universe from the luminosity density, and to obtain a value of $\Omega_m \sim 0.2$. Only some early investigations of large-scale peculiar motions in the Universe have indicated even higher M/L, but these values seem not to be confirmed by more recent measurements. Adapted from J. Schombert's web page http://abyss.uoregon.edu/~js/lectures/cosmo_101.html

8.2.3 Baryon content

Clusters of galaxies largely consist of dark matter. Only about 15 % of their mass is baryonic, the major part of which is contributed by hot intergalactic gas, visible through its X-ray emission and SZ-effect. Given that clusters have a large spatial extent—of order $2h^{-1}$ Mpc which was formed by the gravitational contraction of a comoving volume with a linear extent of about ~ 10 Mpc—it is difficult to imagine how the mixture of baryons and dark matter would strongly differ from the cosmic average. Effects like feedback from supernova explosions or AGN, or other outflow phenomena that are occurring in galaxies and which may reduce their baryonic mass, are probably not effective in galaxy clusters due to their size and their strong gravitational potential. Hence, one might expect that the baryon fraction of clusters is very close to the universal baryon fraction in the Universe, Ω_b/Ω_m. In particular, one would infer that the baryon fraction is pretty much the same for all clusters.

In fact, observations of local clusters show that their baryon fraction f_b is almost constant (see Fig. 8.22). There is a clear tendency that the stellar mass fraction decreases towards higher cluster masses, parallel to the increase of the gas-mass fraction. In Fig. 8.22, the contribution by intergalactic stars was not accounted for; their contribution would (slightly) increase the stellar mass fraction, and thus f_b. Furthermore, the gas mass in clusters is determined from the X-ray luminosity. As the emissivity is $\propto n_e^2$, any clumping of gas would affect the estimated M_{gas}. Although the figure may indicate that the baryon fraction is not exactly constant, for massive clusters this seems to be true in very good approximation.

Assuming the baryon fraction in clusters to be representative of the Universe, the matter density parameter of the Universe can be determined, because the cosmic baryon density is known from primordial nucleosynthesis (see Sect. 4.4.5). This yields

$$\Omega_m = \frac{\Omega_m}{\Omega_b}\,\Omega_b \approx \frac{\Omega_b}{f_b} \approx 0.3\,. \qquad (8.23)$$

The presumed constant gas-mass fraction in clusters can also be turned into a different cosmological probe. This is related to the fact that the determination of cluster gas mass and total mass is dependent on cosmology. More precisely, to turn observables, such as X-ray flux and angular extent of a cluster into physical quantities like (gas) mass, the distance to the cluster needs to be known. For a measured cluster redshift, this is given by the luminosity distance $D_L(z_{cl})$ or the angular-diameter distance $D_A(z_{cl})$. But besides the dependence on redshift (and the simple $\propto H_0^{-1}$ dependence on the Hubble constant), these distances also depend on the density parameters of the cosmological model. In fact, one can show that the estimate of the baryon fraction $f_b \approx M_{gas}/M \propto D_A^{3/2}$ (see Problem 8.2). As a consequence, if one assumes the wrong cosmological parameters, one would find that the baryon fraction systematically varies with cluster redshift. Only for the correct model will the inferred baryon fraction be redshift-independent. Hence, this offers an alternative way of probing the density parameter. Current estimates yield $\Omega_m \sim 0.30 \pm 0.06$ for flat models. In particular, models with $\Omega_\Lambda = 0$ are again excluded with high confidence.

8.2.4 The LSS of clusters of galaxies

Under the assumption that the galaxy distribution follows that of dark matter, it enables us to draw conclusions about the statistical properties of the dark matter distribution, e.g., its power spectrum. At least on large scales, where structure evolution still proceeds almost linearly today, this assumption seems to be justified if an additional bias factor is allowed for. Hence, it is obvious to also examine the large-scale distribution of galaxy clusters, which should follow the

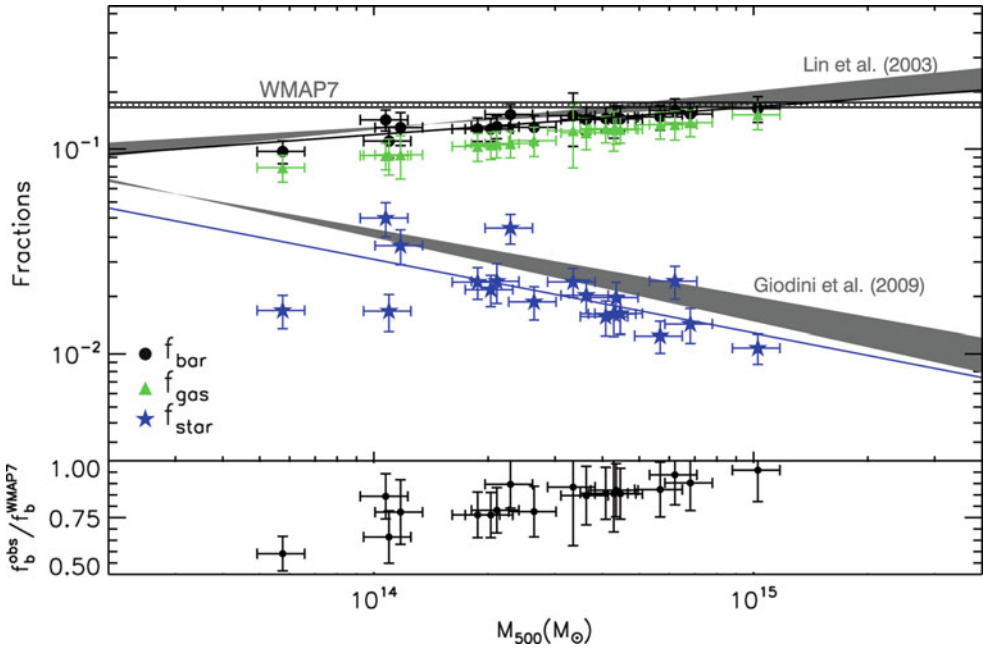


Fig. 8.22 Using a sample of 19 HIFLUGCS clusters situated in the region of the SDSS survey, the stellar mass was determined from the optical data of cluster galaxies. The mass of the intracluster gas was obtained from the X-ray emission, and total mass was estimated from scaling relations. All masses are related to r_{500}. In the *upper panel*, the total baryon-mass fraction (*black circles*), gas-mass fraction (*green triangles*) and stellar mass fraction (*blue stars*) as functions of the total mass are plotted. The *black* and the *blue solid lines* represent power-law fits for the total baryon and stellar mass fractions as a function of total mass, respectively, whereas the *two grey bands* show earlier estimates for the baryon- and stellar-mass fractions. The *horizontal line* indicates the mean baryon fraction in the Universe, i.e., $\Omega_b/\Omega_m \approx 0.17$, as obtained from measurements of the CMB anisotropies. The *lower panel* shows the ratio of baryon fraction as obtained for these clusters and the cosmic average. Source: T.F. Lagana et al. 2011, *XMM-Newton/Sloan Digital Sky Survey: Star Formation Efficiency in Galaxy Clusters and Constraints on the Matter-density Parameter*, ApJ 743, 13, p. 6, Fig. 2. ©AAS. Reproduced with permission

distribution of dark matter on linear scales as well, although probably with a different bias factor.

The ROSAT All-Sky Survey (see Sect. 6.4.5) allowed the compilation of a homogeneous sample of galaxy clusters with which the analysis of the large-scale distribution of clusters became possible for the first time. Figure 8.23 shows that the power spectrum of clusters has the same shape as that of galaxies, however with a considerably larger normalization. The ratio of the two power spectra displayed in this figure is due to different bias factors for galaxies and clusters, $b_{cl} \approx 2.6 b_g$. For this reason the power spectrum for clusters has an amplitude that is larger by a factor of about $(2.6)^2$ than that for galaxies. Since clusters of galaxies are much less abundant than galaxies, one expects them to have a considerably larger bias than galaxies; in fact, Fig. 7.23 shows that dark matter halos corresponding to cluster masses have a few times larger bias than galaxy-mass halos.

The analysis of the power spectrum by means of clusters is interesting, particularly on large scales, yielding an additional data point for the shape parameter $\Gamma = \Omega_m h$. Together with the cluster abundance, their correlation properties yield values of $\Omega_m \approx 0.34$ and $\sigma_8 \approx 0.71$.

8.3 High-redshift supernovae and the cosmological constant

In Sect. 3.9.4 we have seen that SNe Ia can be used to infer distances, due to the fact that they are standardizable candles, i.e., their intrinsic peak luminosity can be determined from measuring the width of their light curves. Furthermore, the explosions are very luminous, so they can be observed out to large distances. Hence, these sources can be used to explore the distance-redshift relation $D_L(z)$ which, as we have seen, depends on the density parameters of the universe. As we will see next, observations of high-redshift SNe Ia have indeed yielded very significant results concerning the composition of our Universe.

8.3.1 Observing SNe Ia at high redshifts

To apply this method, it is necessary to detect and observe SNe Ia at appreciable redshifts, where deviations from the linear Hubble law become visible.

Supernovae are found through the appearance of a point-like source on the sky. This remains true also for SNe at high

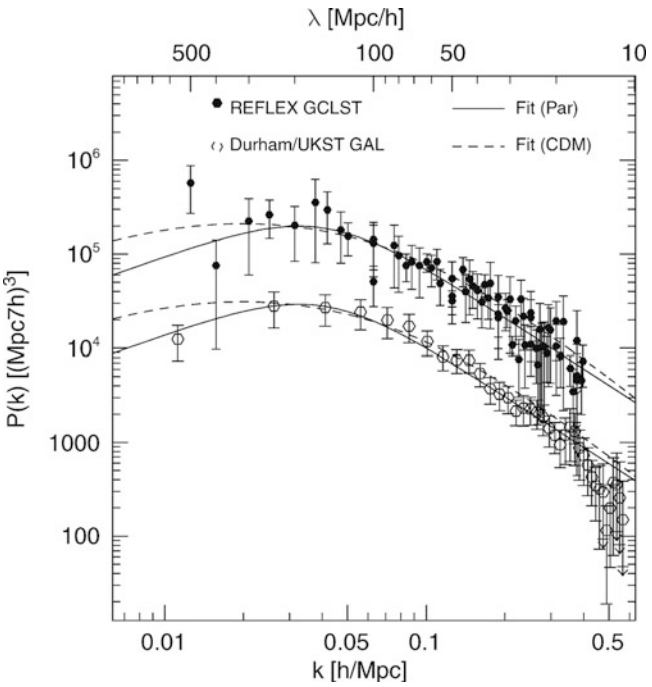


Fig. 8.23 The power spectrum of galaxies (*open symbols*) and of galaxy clusters from the REFLEX survey (*filled symbols*). The two power spectra have basically the same shape, but they differ by a multiplicative factor. This factor specifies the square of the ratio of the bias factors of optically-selected galaxies and of X-ray clusters, respectively. Particularly on large scales, mapping the power spectrum from clusters is of substantial importance. Source: P. Schuecker et al. 2001, *The ROSAT-ESO Flux-Limited X-Ray (REFLEX) galaxy cluster survey. III. The power spectrum*, A&A 368, 86, p. 101, Fig. 16. ©ESO. Reproduced with permission

redshift, although searching for this 'appearance' becomes much more challenging. An image of a field is compared to earlier images, e.g., by subtracting the old image from the new one, and a SNe shows up in this difference image as a point-like source. The development of wide-field cameras have enabled surveys with a high SNe yield. Those which are found in an early stage of the explosion can then be followed-up in detail to measure their light curves in several bands and with good time sampling. Furthermore, spectra of the SNe need to be taken in order to classify them as Type Ia and to determine their redshift. For this observation strategy to be feasible, the availability of observing time for both spectroscopy and subsequent photometry needs to be secured well before the search for candidates begins. Hence, this kind of survey requires a very well-planned strategy and coordination involving several telescopes. Since SNe Ia at high redshift are very faint, the new 8-m class telescopes need to be used for the spectroscopic observations.

In the mid-1990, two large international teams developed such an efficient strategy for the discovery and follow-up of supernovae at large distances. Both teams were very successful in detecting distant SNe Ia. In their first large campaigns, the results of which were published in 1998/1999, they detected and analyzed sources out to redshifts of $z \lesssim 0.8$. Of special relevance is that the conclusions of both teams were in extraordinary agreement. Since they use slightly different methods in the correction of the maximum luminosity (see Fig. 3.48), this agreement serves as a significant test of the systematic uncertainties intrinsic to this method. Since then, many more SNe Ia have been studied, many with redshifts ≥ 1. Wide-field imaging surveys have produced a large number of SN candidates, so the bottleneck of this method lies in the spectroscopic follow-up. Substantial advances have also been made by observing with the HST, including the detection of several SNe Ia at redshift $z > 1$.

8.3.2 Results

As a first result, we mention that the width of the light curve is larger for SNe Ia at higher redshift than it is for local objects. This is expected because, due to redshift, the *observed* width evolves by a factor $(1 + z)$. This dependence has been convincingly confirmed, showing in a direct way the transformation of the intrinsic to the observed time interval as a function of redshift.

Plotting the observed and corrected magnitudes in a Hubble diagram, one can look for the set of cosmological parameters which best describes the dependence of observed magnitudes $m_{\rm obs}$ on redshift. In the left panel of Fig. 8.24, we show the results of the two teams mentioned before, which were published at the end of the 1990s. Shown is the distance modulus as a function of redshift, for a set of low-redshift ($z \leq 0.15$) SNe Ia, and the then newly discovered and studied high-redshift sources. Furthermore, the distance-redshift relation is plotted for three different cosmological models, an Einstein–de Sitter model, a low density open model, and a flat low-density model. As can be seen, the latter fits the data well, whereas the other two models without a cosmological constant provide bad fits, and can be ruled out based on these data.

This discovery of a non-zero dark energy contents in our Universe was awarded the 2011 Nobel Prize in Physics to Saul Perlmutter, Adam Riess & Brian Schmidt, the leaders of the teams. Indeed, this discovery meant a turnaround in our physical world view because, until then, most physicists were convinced that the cosmological constant was zero.

There is another way of looking at these results. Comparing the magnitude at the maximum of the measured SNe Ia, or their distance modulus, respectively, with that which would be expected for an empty universe ($\Omega_{\rm m} = 0 = \Omega_\Lambda$), one finds a surprise (see Fig. 8.25). Considering at first only the supernovae with $z \lesssim 1$, one sees that these are fainter than predicted even for an empty universe. It should be mentioned that, according to (4.13), such an empty universe would expand at constant rate, $\ddot{a} = 0$. The luminosity distance in such a universe is therefore larger than in any

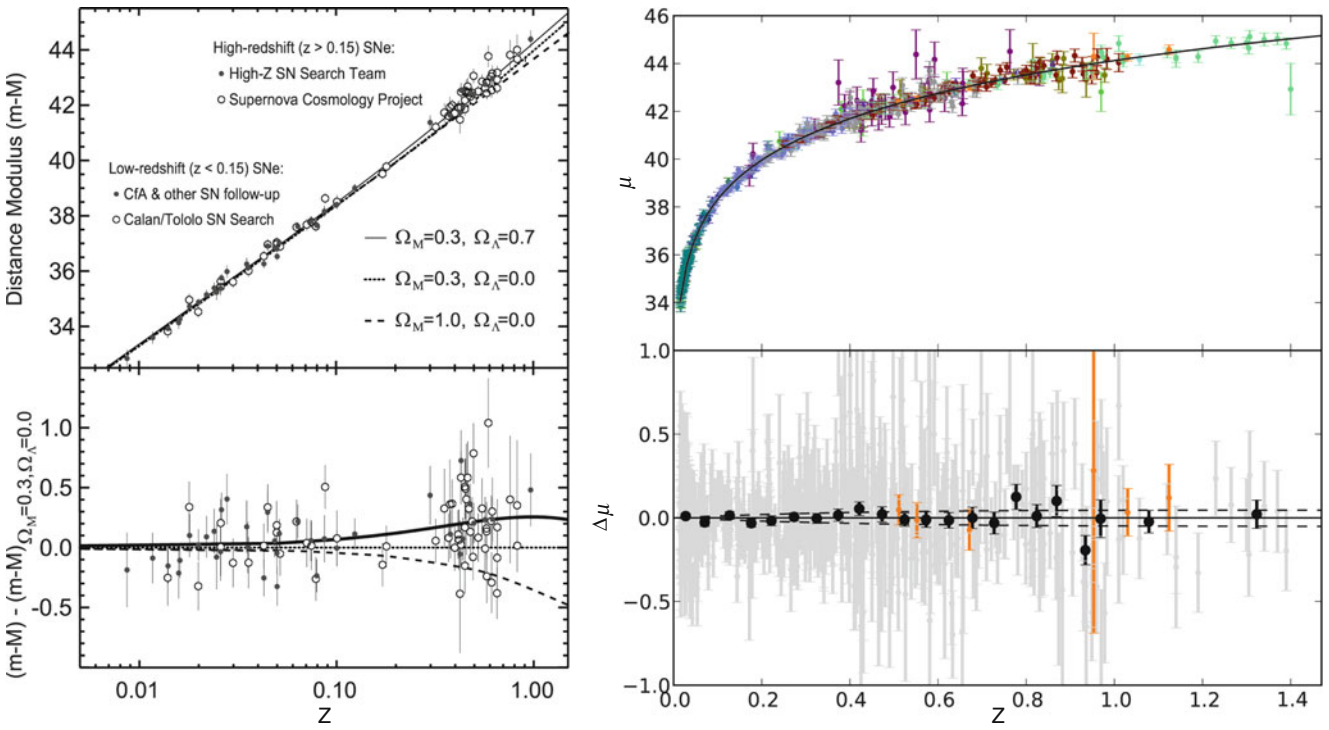


Fig. 8.24 *Left*: The discovery of the accelerated expansion of the Universe from supernova cosmology. The *upper panel* shows the distance modulus of low- and high-redshift SNe Ia, as obtained by two teams (distinguished here by the *two different symbols* for the high-z sources), as a function of redshift, together with the expected behavior from three cosmological models. The *lower panel* show the difference of the measured distance modulus and that expected in a low-density open model; obviously, the *solid curve*, corresponding to a model with $\Omega_\Lambda = 0.7$, fits the data best. *Right*: A more recent version of supernova cosmology results. The *top panel* shows the distance modulus of a set of 557 SNe Ia, determined from the corrected maximum flux of the source, where *different colors* indicate SNe discovered by different teams. The curve through the data points corresponds to the distance-redshift relation of a flat universe with $\Omega_\mathrm{m} = 0.275$. Remarkable is the small scatter of the data points around the curve which indicates that the small dispersion of corrected maximum fluxes shown in Fig. 3.47 extends to higher redshifts. In the *bottom panel*, the same data points for individual SNe are shown in *grey*, with this best-fitting model subtracted, whereas the *solid black* points show the residuals binned in redshift. Source: *Left*: S. Perlmutter & B.P. Schmidt 2003, *Measuring Cosmology with Supernovae*, astro-ph/0303428, Fig. 4. Reproduced by permission of the author. *Right*: R. Amanullah et al. 2010, *Spectra and Hubble Space Telescope Light Curves of Six Type Ia Supernovae at* $0.511 < z < 1.12$ *and the Union2 Compilation*, ApJ 716, 712, p. 727, Fig. 9. ©AAS. Reproduced with permission

other universe with a vanishing cosmological constant. The luminosity distance can only be increased by assuming that the Universe expanded *more slowly* in the past than it does today, hence that *the expansion has accelerated over time.* From (4.19) it follows that such an accelerated expansion is possible only if $\Omega_\Lambda > 0$.

Since then, this result has been confirmed by ever more detailed investigations. In particular, the sample of SNe Ia was enlarged and (by employing the HST) extended to higher redshifts; for example, the right-hand panel of Fig. 8.24 shows the Hubble diagram of 557 SN Ia, the so-called Union 2 compilation. From the high-redshift objects, it was shown that for $z \gtrsim 1$ the trend is reversed and SN Ia become brighter than they would be in an empty universe (see Fig. 8.25). At these high redshifts the matter density dominates the Universe, evolving as $(1 + z)^3$, in contrast to the constant vacuum energy.

The corresponding constraints on the density parameters Ω_m and Ω_Λ are plotted in Fig. 8.26, in comparison to those that were obtained in 1998. As becomes clear from the confidence contours, the SN Ia data are not compatible with a universe without a cosmological constant. An Einstein–de Sitter model is definitely excluded, but also a model with $\Omega_\mathrm{m} = 0.3$ (a value derived from galaxy redshift surveys) and $\Omega_\Lambda = 0$ is incompatible with these data. More recent results from supernova cosmology will be discussed in connection with other cosmological probes in Sect. 8.7.

We conclude from these results that a non-vanishing dark energy component exists in the Universe, causing an accelerated expansion through its negative pressure. The simplest form of this dark energy is the vacuum energy or the cosmological constant.

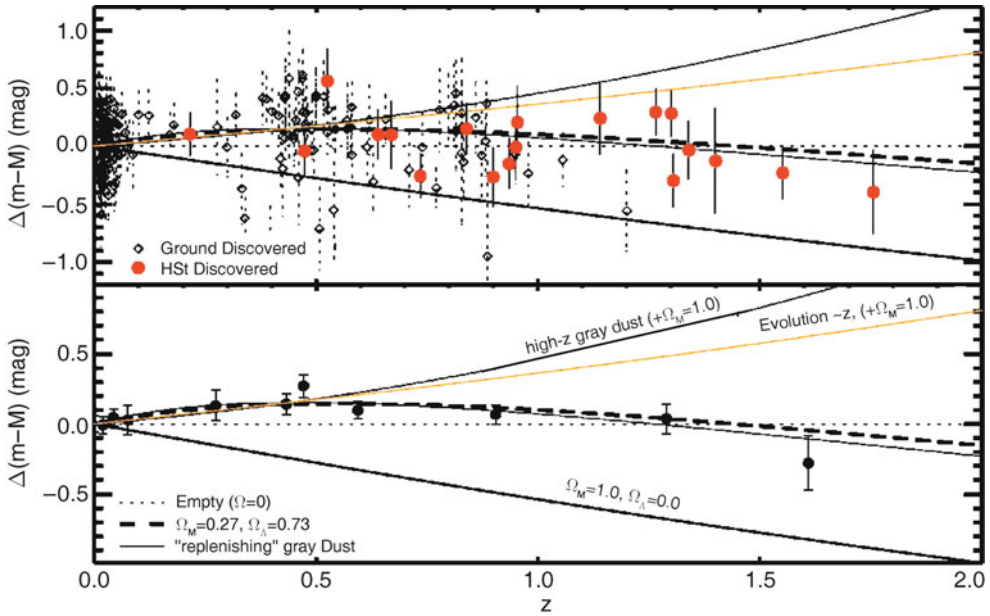


Fig. 8.25 Difference between the maximum brightness of SNe Ia and that expected in an empty universe ($\Omega_\mathrm{m} = 0 = \Omega_\Lambda$). *Diamond symbols* represent events that were detected from the ground, *circles* the ones discovered by the HST. The HST data are essential for the discovery of high-redshift SN Ia, since due to the redshift, observations in near-IR wavebands are needed. In the *top panel*, the individual SNe Ia are presented, whereas in the *bottom panel* they are averaged in redshift bins. An empty universe would correspond to the *dotted straight line*, $\Delta(m - M) = 0$. The *dashed curve* corresponds to a cosmological model with $\Omega_\mathrm{m} = 0.27$, $\Omega_\Lambda = 0.73$. Furthermore, model curves for universes with constant acceleration are drawn; these models, which are not well-motivated from physics, and models including 'gray dust' (in which the extinction is assumed to be independent of the wavelength), can be excluded. Source: A. Riess et al. 2004, *Type Ia Supernova Discoveries at $z > 1$ from the Hubble Space Telescope: Evidence for Past Deceleration and Constraints on Dark Energy Evolution*, ApJ 607, 665, p. 677, Fig. 7. ©AAS. Reproduced with permission

8.3.3 Discussion

The discovery of the Hubble diagram of SNe Ia being incompatible with a universe having a vanishing vacuum energy came as a surprise. It was the first direct evidence for the existence of dark energy. The cosmological constant, first introduced by Einstein, then later discarded again, seems to indeed have a non-vanishing value.

This far-reaching conclusion, with its consequences for fundamental physics, obviously needs to be critically examined. Are there options to explain the observations without demanding an accelerated expansion of the Universe?

Evolutionary effects. The above analysis is based on the implicit assumption that, on average, SNe Ia all have the same maximum (corrected) luminosity, independent of their redshift. As for other kinds of sources for which a Hubble diagram can be constructed and from which cosmological parameters can be derived, the major difficulty lies in distinguishing the effects of spacetime curvature from evolutionary effects. A z-dependent evolution of SNe Ia, in such a way that they become less luminous with increasing redshift, could have a similar effect on a Hubble diagram as an accelerated expansion.

At first sight, such an evolution seems improbable since, according to our current understanding, the explosion of a white dwarf close to the Chandrasekhar mass limit is responsible for these events, and this mass threshold solely depends on fundamental physical constants. On the other hand, the exact mass at which the explosion is triggered may well depend on the chemical composition of the white dwarf, and this in turn may depend on redshift. Although it is presumably impossible to prove that such evolutionary effects are not involved or that their effect is at least smaller than cosmological effects, one can search for differences between SNe Ia at low and at high redshift. For instance, it has been impressively demonstrated that the spectra of high-redshift SNe Ia are very similar to those of nearby ones. Hence, no evidence for evolutionary effects has been found from these spectral studies. Furthermore, the time until the maximum is reached is independent of z, if one accounts for the time dilation $(1 + z)$.

However, it was found that the distribution of SN Ia brightness and the average shape of the light curves depends on properties of the host galaxy, in that early-type galaxies contain a larger fraction of SNe Ia with narrower light curves. Whereas these differences are to first order accounted for by the magnitude correction according to light-curve shape,

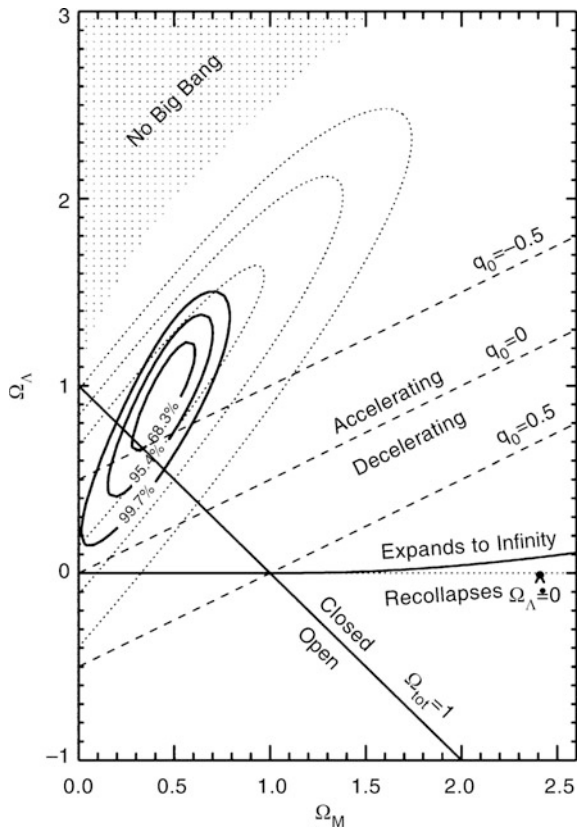


Fig. 8.26 From the measured magnitudes of the SN Ia displayed in Fig. 8.25 and the correspondingly implied values for the luminosity distances, confidence regions are plotted in the Ω_m–Ω_Λ-plane which classifies the cosmological models (see Fig. 4.7). The *solid* contours result from the 157 SNe Ia that are also plotted in Fig. 8.25, whereas the *dotted* contours represent the early results from 1998. *Dashed lines* represent cosmological models with the same deceleration parameter q_0. Source: A. Riess et al. 2004, *Type Ia Supernova Discoveries at* $z > 1$ *from the Hubble Space Telescope: Evidence for Past Deceleration and Constraints on Dark Energy Evolution*, ApJ 607, 665, p. 678, Fig. 8. ©AAS. Reproduced with permission

there seems to remain a small residual difference in the distribution of corrected peak magnitudes, with galaxies of high stellar masses hosting slightly brighter SN Ia on average. This effect is currently smaller than the statistical uncertainties and does not affect the conclusions concerning cosmological parameters.

Extinction. The correction of the luminosity for extinction in the host galaxy and in the Milky Way is determined from reddening. The relation between extinction and reddening depends on the properties of the dust—if these evolve with z the correction may become systematically wrong. Therefore, it is important to employ multi-band data with which the reddening law (i.e., the relation between reddening and extinction) can be measured. In fact, such studies revealed that the typical reddening law is described by $R_V \sim 2$, compared to the mean one of the Milky Way with $R_V \sim 3.1$. The correction for dust can also be checked by separately investigating SNe Ia that occur in early-type galaxies, in which only little dust exists, and comparing these to events in spiral galaxies.

One possibility that has been discussed is the existence of 'gray dust': dust that causes an absorption independent of wavelength. In such a case extinction would not reveal itself by reddening. However, this hypothesis lacks a theoretical explanation for the physical nature of the dust particles. In addition, the observation of SNe Ia at $z \gtrsim 1$ shows that the evolution of their magnitude at maximum is compatible with a Λ-universe. In contrast, in a scenario involving 'gray dust', a monotonic decrease of the brightness with redshift would be expected, relative to an empty universe (see Fig. 8.25). Nevertheless, there may be some (small) amount of intergalactic dust (or dust in the very outskirts of galaxies along the line-of-sight to the SNe). Whereas this dust contribution is partly corrected for using reddening, the fact that this extinction can occur anywhere along the line-of-sight implies that it does not follow the reddening law at the redshift of the source or the local reddening law of the Galaxy.

Other systematics. Figure 3.48 suggests that the light-curve fitting method to correct for the peak brightness works very well. However, when it is used for precision cosmology, it needs to be seen how well this purely empirical method corrects for differences of the explosion. Since several independent light-curve fitting methods have been developed, one can compare their performance. Indeed, small differences in the estimated peak brightness are obtained by these methods, but they yield smaller effects than the current statistical accuracy.

Selection effects may also play some role, in that brighter objects are more easily detected (and spectroscopically followed-up) than fainter ones. Such effects need to be quantified with simulations.

At present, the combination of the various systematic effects is constrained to be sub-dominant compared to the statistical uncertainty which is given by the number of SNe Ia which have sufficiently good data to be included in supernova cosmology projects. However, the ever larger imaging surveys will produce a large rate of newly detected SNe in the upcoming years, which can be used to reduce the statistical uncertainty substantially. This, however, will lead to tighter constraints on cosmological parameters only if the potential systematic effects are better understood, either through empirical studies, or by a better physical understanding of the explosion process.

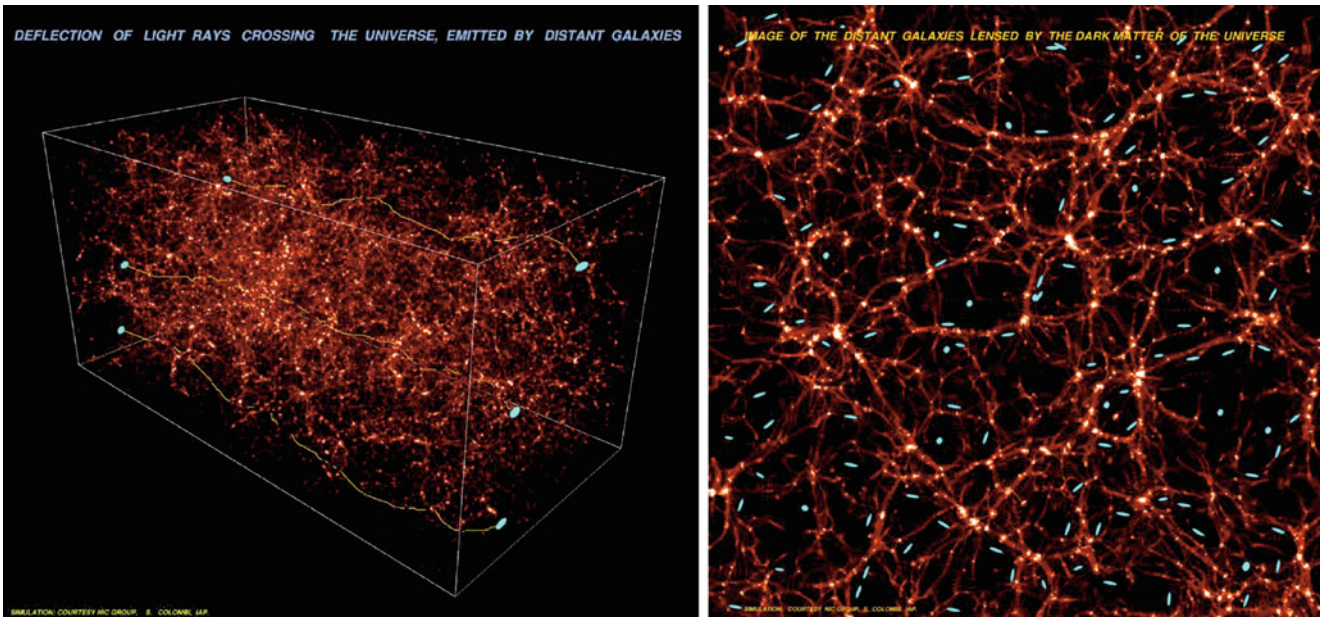


Fig. 8.27 As light beams propagate through the Universe they are affected by the inhomogeneous matter distribution; they are deflected, and the shape and size of their cross section changes. This effect is displayed schematically in the *left panel*—light beams from sources at the far side of the cube are propagating through the large-scale distribution of matter in the Universe, and we observe the distorted images of the sources. In particular, the image of a circular source is elliptical to a first approximation. Since the distribution of matter is highly structured on large scales, the image distortion caused by light deflection is coherent: the distortion of two neighboring light beams is very similar, so that the observed ellipticities of neighboring galaxies are correlated. From a statistical analysis of the shapes of galaxy images, conclusions about the statistical properties of the matter distribution in the Universe can be drawn. Hence, the ellipticities of images of distant sources are closely related to the (projected) matter distribution, as displayed schematically in the *right panel*. Credit: Y. Mellier, L. van Waerbeke, S. Colombi et al./Canada-France-Hawaii Telescope Corporation

8.4 Cosmic shear

The principle. On traversing the inhomogeneous matter distribution in the Universe, light beams are deflected and distorted, where the distortion is caused by the tidal gravitational field of the inhomogeneously distributed matter. As was already discussed in the context of the reconstruction of the matter distribution in galaxy clusters (see Sect. 6.6.2), by measuring the shapes of images of distant galaxies this tidal field can be mapped. The distortion of light bundles, and thus of the images of distant galaxies, by the light deflection of the large-scale structure is called *cosmic shear*.

There are two major differences between cosmic shear and the previously discussed weak lensing by clusters. First, in cosmic shear the light deflection and distortion is caused by the three-dimensional mass distribution in the Universe between the distant sources and us, and not (mainly) by a single mass concentration along the line-of-sight. Of course, massive clusters are part of the large-scale structure, but they are very rare; for most lines-of-sight, the lensing effects are much weaker than near clusters, which is the second major difference. A typical value for this shear is about 1 %, meaning that the image of an intrinsically circular source attains an axis ratio of 0.99 : 1. This induced ellipticity is *much* smaller than the width of the intrinsic ellipticity distribution of sources. Thus, in order to measure cosmic shear, one needs a *large* number of galaxies.

The shear field results from the projection of the three-dimensional tidal field along the line-of-sight. The shear in the direction of any single galaxy is not only unmeasurable, but it yields no interesting information. If two sources are located closely on the sky, their light bundles propagate through nearly the same gravitational field, and hence their distortion is expected to be very similar, as illustrated in Fig. 8.27. In other words, the distortions of pairs of galaxy images are expected to be correlated. The larger the separation of image pairs, the more can small-scale inhomogeneities of the matter density field distort just one of the two light bundles and leave the other almost unaffected. This implies that the shear correlation is expected to decrease with increasing angular separation, and that this decrease depends on the density fluctuations in the Universe as a function of scale. For example, if there are high-amplitude fluctuations with a large wavelength, the correlated shear will drop less quickly with separation than in the case where these large-scale fluctuations are much weaker. The strength of the density fluctuations as a function of wavelength is described by the power spectrum $P(k, z)$. Thus, the shear two-point correlation function $\xi_+(\theta)$ probes the density fluctuations

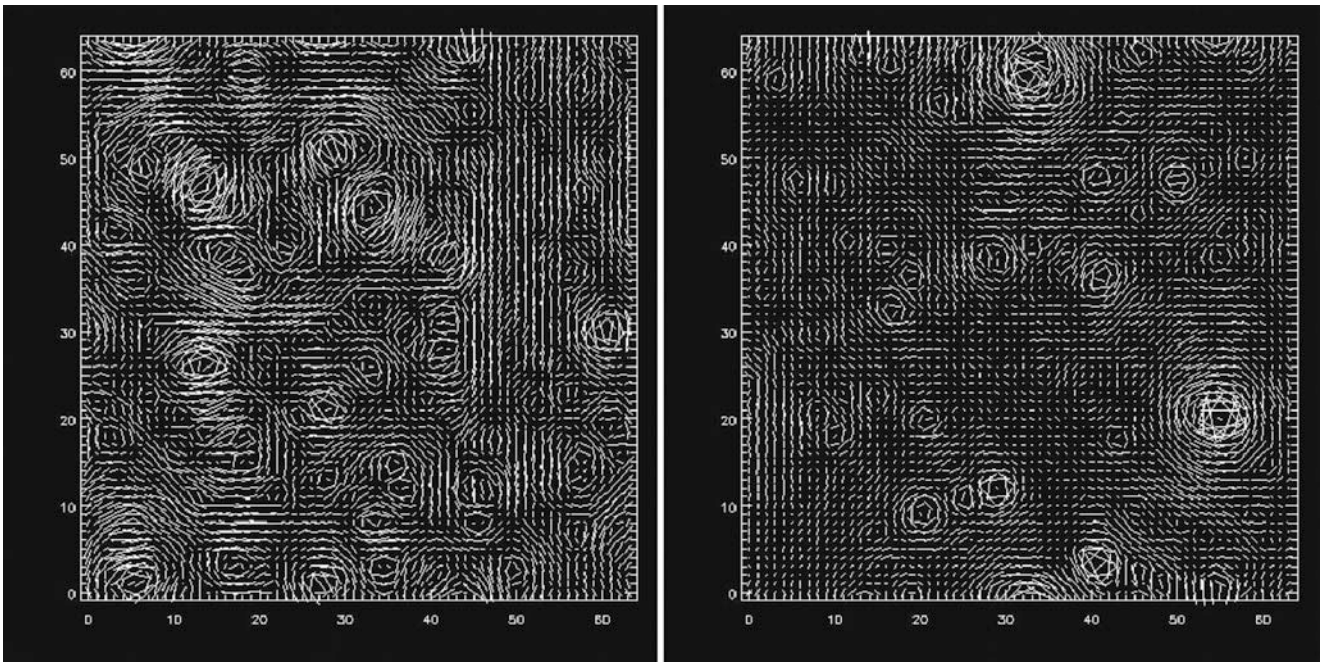


Fig. 8.28 The shear field as obtained by numerical ray-tracing simulations, for two different cosmological models: an Einstein–de Sitter model (*left panel*) and a low-density open model (*right panel*). As can be clearly seen, the statistical properties of these two shear fields are significantly different. In the low-density model, the circular shear patterns (which are generated by mass concentrations along the line-of-sight to the source population) are fairly isolated, whereas they are densely together in the high-density model. Hence, if these statistical properties can be measured, one can distinguish between such cosmological models. The shear correlation function is one of the quantities used to characterize the statistical properties of shear fields. In both panels, the field-of-view is $1\,\mathrm{deg}^2$. Source: B. Jain et al. 2000, *Ray-tracing Simulations of Weak Lensing by Large-Scale Structure*, ApJ 530, 547, p. 558, 559, Fig. 6. ©AAS. Reproduced with permission

on a comoving scale $f_k(z)\,\theta$ at redshift z, or the power spectrum at wave number $k = 2\pi/[f_k(z)\,\theta]$, integrated over the redshift from 0 to the redshift of the source galaxies.

In other words, the shear correlation function is an integral over the matter power spectrum, with a weight function that depends on the redshift distribution of the source galaxies. Hence, measuring the shear correlation function $\xi_+(\theta)$ yields information on the power spectrum; in addition, $\xi_+(\theta)$ is sensitive to the expansion history of the Universe, due to the dependence of the lensing strength on the distance-redshift relation [cf. the definition (3.67) of the critical surface mass density]. As is illustrated in Fig. 8.28, by comparing measurements of cosmic shear with cosmological models we obtain constraints on the cosmological parameters, without the need to make any assumptions about the relation between luminous matter (galaxies) and dark matter.

First detections. In March 2000, four research groups published, quasi simultaneously, the first measurements of cosmic shear, and in the fall of 2000 another measurement was obtained from VLT observations. After that, several teams worldwide have successfully performed measurements of cosmic shear, for which a large number of different telescopes have been used, including the HST. Some of the early results are compiled in Fig. 8.29.

These early results were encouraging and demonstrated the feasibility of such measurements. In order to make progress, the survey size needed to be increased to reduce the error bars in the measurements. The development of wide-field cameras and of special software for data analysis are mainly responsible for the achievements in the past years. It was also clear that cosmic shear requires the best possible imaging quality, since the faint background galaxies are small, with a size comparable with, or smaller than the seeing obtainable at the best observing sites from the ground. In parallel, theoretical studies underlined the large cosmological potential of cosmic shear, so that it is now seen as a most valuable tool for observational cosmology.

Systematics. With increased statistical power of cosmic shear surveys, and hence reduced statistical error, systematic effects become more important. There are quite a number of effects which need to be accounted for in order to obtain cosmologically reliable results.

The first is the impact of the point-spread function (PSF). In principle, the problems are the same as in the mass reconstruction of galaxy clusters with the weak lensing effect (Sect. 6.6.2), but they are substantially more difficult to deal with since the measurable signal is considerably smaller. The PSF smears the true image, and thus makes

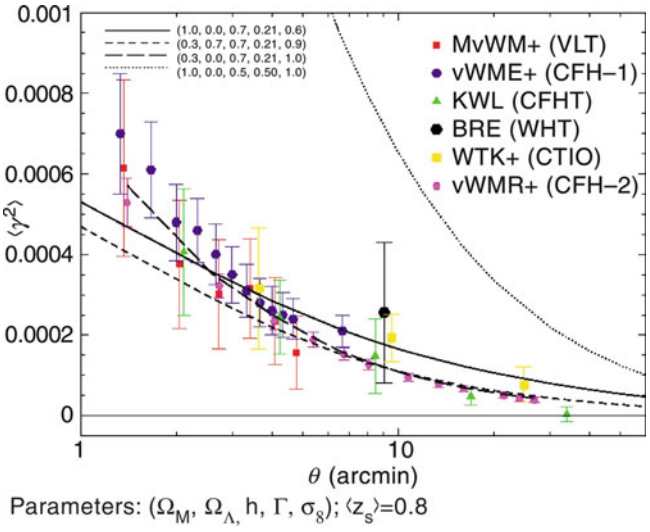


Fig. 8.29 Early measurements of cosmic shear. Plotted is the shear dispersion, an integral over the shear correlation function, measured from the ellipticities of faint and small galaxy images on deep CCD exposures, as a function of angular scale. Data from different teams are represented by *different symbols*. For instance, MvWM+ resulted from a VLT-project, vWMR+ from a large survey (VIRMOS-Descartes) at the CFHT. For this latter project, the images of about 450 000 galaxies were analyzed; the corresponding error bars from this survey are significantly smaller than those of the earlier surveys. The curves indicate cosmic shear predictions in different cosmological models, where the curves are labeled by the cosmological parameters Ω_m, Ω_Λ, h, Γ and σ_8. Credit: L. van Waerbeke & Y. Mellier/Institute d'Astrophysique de Paris

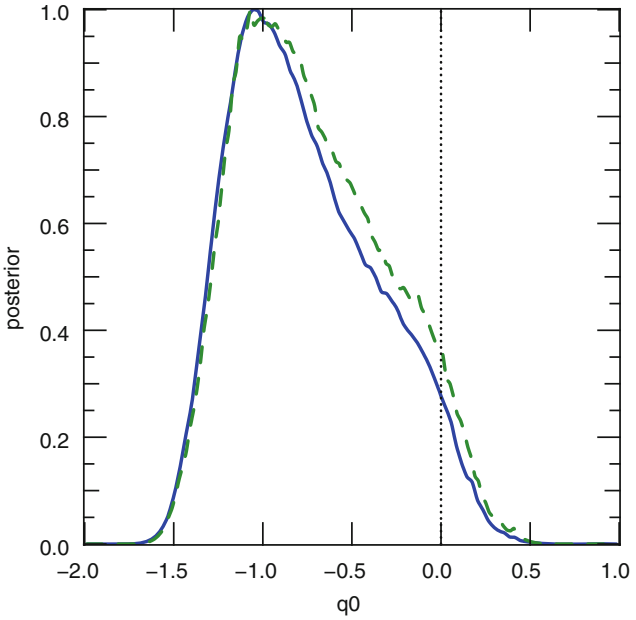


Fig. 8.30 The probability distribution for the deceleration parameter q_0 as obtained from the cosmic shear analysis of the COSMOS survey. For this analysis, it was assumed that the Hubble constant lies in the range determined by the Hubble Key Project (i.e., $h = 0.72 \pm 0.08$), and the baryon density was fixed to its Big Bang Nucleosynthesis value (*dashed curve*); the *solid curve* assumes a slightly narrower range of h. Note that the probability for q_0 to be positive is only about 5 %, which means that one can exclude with 95 % probability that the Universe is not accelerating today. Source: T. Schrabback et al. 2010, *Evidence of the accelerated expansion of the Universe from weak lensing tomography with COSMOS*, A&A 516, A63, p. 14, Fig. 13. ©ESO. Reproduced with permission

the observed images rounder than they would be without the blurring by the atmosphere. In addition, the PSF is not necessarily circular, but due to a combination of effects (e.g., tracking errors, or wind shake of the telescope, effects of the telescope/detector optical system), it can be elliptical as well. Therefore, even if the true image was round, the observed image could carry an ellipticity from the PSF. These effects need to be corrected for, otherwise the observed 'shear correlation function' would be totally dominated by PSF effects. Fortunately, the PSF can be directly measured from the images, since stars are point-like sources. Hence, the size and shape of the PSF can be measured from the observed images of stars. After this measurement, the galaxy ellipticities need to be corrected for the influence of the PSF. Several sophisticated methods for this have been developed over the years, and they are still improving.

There are also astronomical systematics, the most important being intrinsic alignments of galaxies. To understand this effect, we recall our discussion how dark matter halos can attain an angular momentum: If a non-spherical mass distribution is located in a tidal gravitational field, it experiences a torque which tends to align the body with the direction of the tidal field. Thus, a close pair of galaxies may have a correlated intrinsic ellipticity, since they are subject to the same (or similar) tidal field. Of course, this argument assumes that the shape and orientation of the observable galaxy (i.e., the stars, or the light distribution) is related to the shape, orientation, or spin of its dark matter halo. This intrinsic alignment could mimic a shear correlation and needs to be considered. Fortunately, there is a fairly straightforward way how this can be treated, since such alignments can only occur if the two galaxies are close together in three-dimensional space, i.e., if they have the same redshift. Excluding pairs of galaxies with the same redshift from the measurement of the shear correlation function eliminates this effect. However, to be able to do so, one needs redshift information about the individual source galaxies. Estimates are that, without a correction, these intrinsic alignments can cause a spurious signal that is a few percent of the true shear correlation.

There is a second intrinsic alignment effect that is slightly more subtle. Suppose a galaxy at redshift z_1 lies in a tidal field and its orientation has been affected by it. A source nearby on the sky, but at redshift $z_2 > z_1$, experiences a shear which is the integral of the tidal gravitational field along its line-of-sight, including the tidal field from the redshift regime around z_1. Thus, the shear of the galaxy at z_2 can

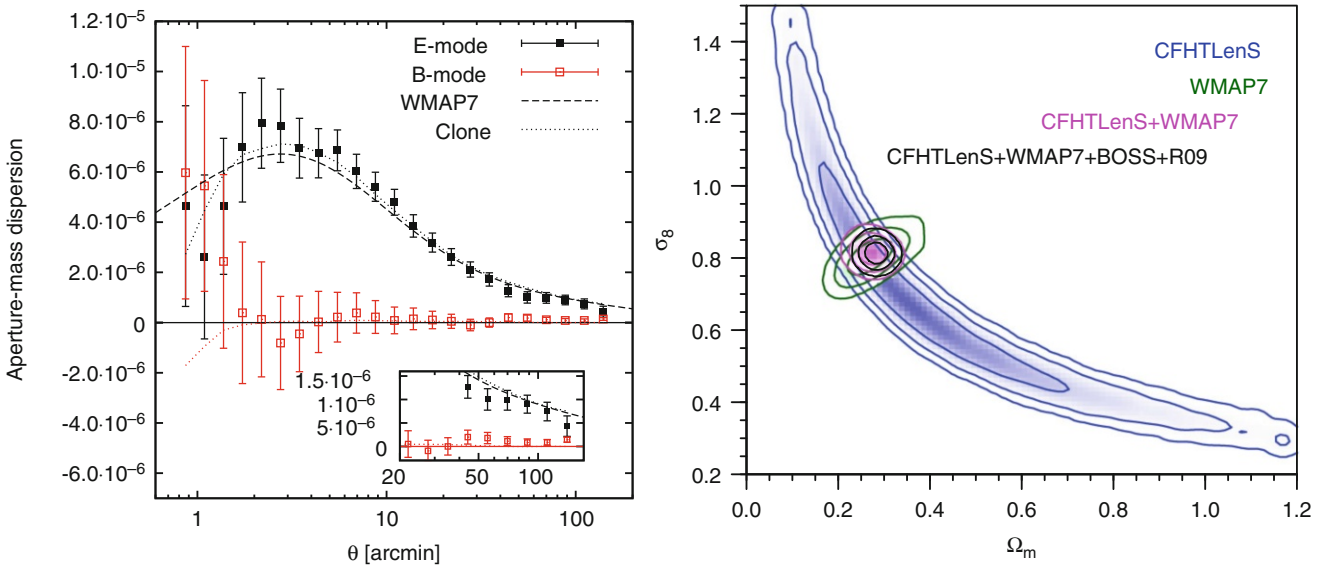


Fig. 8.31 Results from the CFHTLenS project. The *left panel* displays the cosmic shear measurement in the survey, here expressed as a particular integral over the shear two-point correlation function. The *black symbols* show the cosmic shear signal (here called 'E-mode'), whereas the *red symbols* ('B-mode') indicate a different integral over the shear correlations which is expected to be zero—it thus serves as a diagnostic for remaining systematic effects in the data set. The B-mode signal is compatible with zero on all angular scales, where the E-mode signal is detected with very high significance. In particular, the measured cosmic shear is compatible with the expectation from the standard cosmological model, shown by the *solid curve*. The *right panel* displays in *blue* the allowed region in σ_8-Ω_m parameter space, where a flat ΛCDM model is assumed. Whereas the confidence region is considerable larger than obtained from CMB anisotropy measurements (Sect. 8.6), shown in green, they are essentially orthogonal to them. Source: M. Kilbinger et al. 2013, *CFHTLenS: combined probe cosmological model comparison using 2D weak gravitational lensing*, MNRAS 430, 2200, p. 2208, 2212, Figs. 8, 10. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

be correlated to the intrinsic ellipticity of the galaxy at z_1. This again is a spurious signal that adds to the true cosmic shear effect and needs to be accounted for. One can show that with redshift information of galaxies, also this 'shape-shear alignment' can be eliminated.

Results. We want to mention here two of the cosmic shear surveys that have been conducted. The first one is the COSMOS survey, the largest contiguous field ever imaged by HST. It consists of 579 fields of the ACS camera and comprises a total area of 1.64 deg^2. The exquisite image quality of HST and its low sky background makes that a particularly valuable data set for cosmic shear. Furthermore, the COSMOS field has been reobserved in many other wavebands, and this additional information can be used to obtain redshift information about the faint galaxies in the field (see Sect. 9.1.2). Comparing the observed shear correlation function with theoretical predictions, and assuming a spatially flat universe, a relation between the normalization of the power spectrum, σ_8, and the matter density parameter Ω_m of the form

$$\sigma_8 \left(\frac{\Omega_\mathrm{m}}{0.3}\right)^{0.51} = 0.75 \pm 0.08 \qquad (8.24)$$

was obtained. Note that this constraint is very similar in form, and compatible with, the constraint (8.18) obtained from cluster abundance. The reason for that is that the spatial scale where the cosmic shear in COSMOS is most sensitive to is around 5 Mpc, i.e., a scale comparable to that probed by clusters (see Problem 7.3). Individual constraints on these two parameters from COSMOS are

$$\Omega_\mathrm{m} = 0.27 \pm 0.03\,, \quad \sigma_8 = 0.80 \pm 0.03\,. \qquad (8.25)$$

Dropping the assumption on flatness, the COSMOS survey yields independent evidence for the accelerated current expansion of the Universe. Figure 8.30 shows the probability distribution of the deceleration parameter q_0 [see (4.34)] whose sign agrees with the sign of $\ddot{a}$ at redshift zero. With about 95 % probability, the COSMOS data require an accelerating universe today.

The state-of-the-art in cosmic shear research is defined by the CFHTLenS survey, which is based on 170 deg^2 of deep five-band optical imaging with the CFHT (see Fig. 1.35). The combination of image quality, wavelength coverage, and depth allowed a very detailed study of weak lensing effects in this survey. The multi-color property of the survey allows an estimate of redshift information for the galaxies, and thus a

good calibration of the lensing strength. Also for this survey, the strongest constraint is obtained for a combination of σ_8 and Ω_m, yielding (see Fig. 8.31)

$$\sigma_8 \left(\frac{\Omega_\mathrm{m}}{0.27}\right)^{0.6} = 0.79 \pm 0.03 \,, \tag{8.26}$$

if no other information is used in the analysis. The degeneracy between these two parameters can be broken if information from CMB anisotropies, BAOs, and an estimate of the Hubble constant is used, yielding $\Omega_\mathrm{m} = 0.283 \pm 0.010$ and $\sigma_8 = 0.813 \pm 0.014$.

8.5 Origin of the Lyman-α forest

We have seen in Sect. 5.7 that in the spectrum of any QSO a large number of absorption lines at wavelengths shorter than the Lyα emission line of the QSO are found. The major fraction of these absorption lines originate from the Lyα transition of neutral hydrogen located along the line-of-sight to the source (see Fig. 8.32). Since the absorption is found in the form of a line spectrum, the absorbing hydrogen cannot be distributed homogeneously. A homogeneous intergalactic medium containing neutral hydrogen would be visible in continuum absorption. In this section, we will first examine this continuum absorption. We will then summarize some observational results on the Lyα forest and explain why studying this provides us with valuable information about the cosmological parameters.

8.5.1 The homogeneous intergalactic medium

We first ask whether part of the baryons in the Universe may be contained in a homogeneous intergalactic medium. This question can be answered by means of the *Gunn–Peterson test*. Neutral hydrogen absorbs photons at a rest wavelength of $\lambda = \lambda_{\mathrm{Ly}\alpha} = 1216\,\text{Å}$. Photons from a QSO at redshift z_QSO attain this wavelength $\lambda_{\mathrm{Ly}\alpha}$ somewhere along the line-of-sight between us and the QSO, if they are emitted by the QSO at $\lambda_{\mathrm{Ly}\alpha}\left(1+z_\mathrm{QSO}\right)^{-1} < \lambda < \lambda_{\mathrm{Ly}\alpha}$. However, if the wavelength at emission is larger than $\lambda_{\mathrm{Ly}\alpha}$, the radiation can nowhere on its way to us be absorbed by neutral hydrogen. Hence, a jump in the observed continuum radiation should occur between the red and the blue side of the Lyα emission line of the QSO: this is the Gunn–Peterson effect. The optical depth for absorption is, for models with $\Omega_\Lambda = 0$, given by

$$\tau = 4.14 \times 10^{10}\, h^{-1} \frac{n_\mathrm{HI}(z)/\mathrm{cm}^{-3}}{E(z)} \,, \tag{8.27}$$

where $n_\mathrm{HI}(z)$ is the density of neutral hydrogen at the absorption redshift z, with $(1+z) = \lambda/\lambda_{\mathrm{Ly}\alpha} < (1+z_\mathrm{QSO})$. If we express the neutral hydrogen density in terms of the neutral baryon fraction x_HI and the baryon density of the Universe, this expression can be transformed as

$$\tau = 1.41 \times 10^5\, h^{-1} \left(\frac{\Omega_\mathrm{b} h^2}{0.02}\right) \Omega_\mathrm{m}^{-1/2} \left(\frac{1+z}{7}\right)^{3/2} x_\mathrm{HI} \,, \tag{8.28}$$

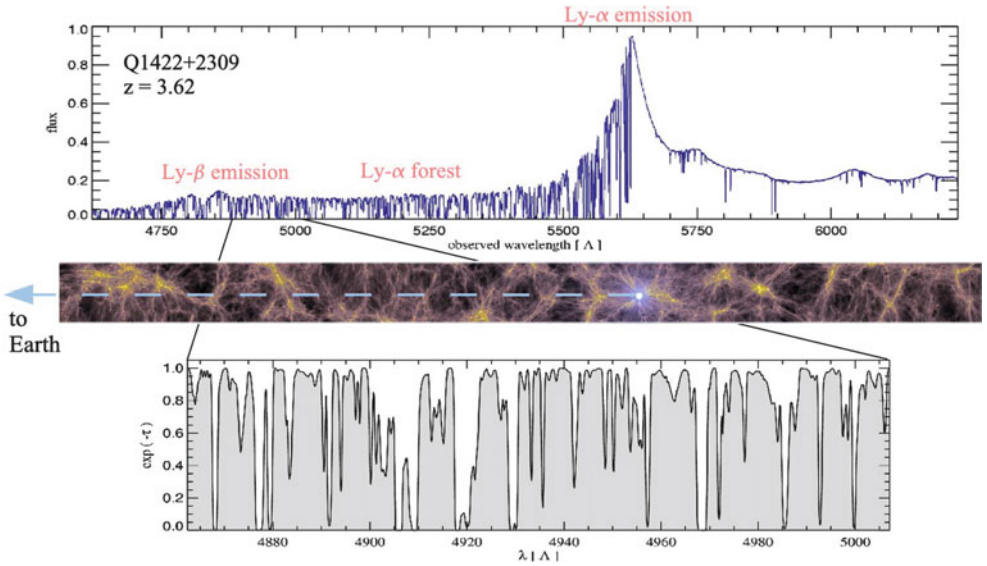


Fig. 8.32 The *top panel* shows the optical spectrum of the QSO 1422 + 2309, with a zoom of a small portion of it shown *at the bottom*. Due to the redshifting of photons, the absorption at a given wavelength is due to neutral hydrogen along the line-of-sight at a redshift determined by the observed wavelength. By studying the characteristics of the absorption spectrum, one obtains a tomographic view through the neutral hydrogen in the Universe. Source: V. Springel et al. 2006, *The large-scale structure of the Universe*, astro-ph/0604561, Fig. 2. Reproduced by permission of the authors

which shows that the Gunn–Peterson optical depth is large, unless the neutral fraction $x_{\rm HI}$ is very low.

Such a jump in the continuum radiation of QSOs across their Lyα emission line, with an amplitude $S(\text{blue})/S(\text{red}) = \mathrm{e}^{-\tau}$, has not been observed for QSOs at $z \lesssim 5$. Tight limits for the optical depth were obtained by detailed spectroscopic observations, yielding $\tau < 0.05$ for $z \lesssim 3$ and $\tau < 0.1$ for $z \sim 5$. At even higher redshift observations become increasingly difficult, because the Lyα forest then becomes so dense that hardly any continuum radiation is visible between the individual absorption lines (see, e.g., Fig. 5.55 for a QSO at $z_{\rm QSO} = 3.62$). From the upper limit for the optical depth, one obtains bounds for the density of uniformly distributed neutral hydrogen,

$$n_{\rm HI}(\text{comoving}) \lesssim 2{\times}10^{-13}\, h\,\mathrm{cm}^{-3} \quad \text{or} \quad \Omega_{\rm HI} \lesssim 2{\times}10^{-8}\, h^{-1}\ .$$

From this we conclude that hardly any homogeneously distributed baryonic matter exists in the intergalactic medium, or that hydrogen in the intergalactic medium is virtually fully ionized. However, from primordial nucleosynthesis we know that the average density of hydrogen is much higher than the above limits, so that hydrogen must be present in essentially fully ionized form. We will discuss in Sect. 10.3 how this reionization of the intergalactic medium presumably happened.

After 2000, QSOs at redshifts >6 were discovered, not least by careful color selection in data from the Sloan Digital Sky Survey (see Sect. 8.1.2). The spectrum of one of these QSOs is displayed in Fig. 5.58. For this QSO, we can see that virtually no radiation bluewards of the Lyα emission line is detected. After this discovery, it was speculated whether the redshift had been identified at which the Universe was reionized. The situation is more complicated, though. First, the Lyα forest is so dense at these redshifts that lines blend together, making it very difficult to draw conclusions about a homogeneous absorption. Second, in spectra of QSOs at even higher redshift, radiation bluewards of the Lyα emission line has been found. As we will soon see, the reionization of the Universe probably took place at a redshift somewhat higher than $z \sim 6$.

8.5.2 Phenomenology of the Lyα forest

Neutral hydrogen in the intergalactic medium *is* being observed in the Lyα forest. For the observation of this Lyα forest, spectra of QSOs with high spectral resolution are required because the typical width of the lines is very small, corresponding to a velocity dispersion of $\sim 20\,\mathrm{km/s}$. To obtain spectra of high resolution and of good signal-to-noise ratio, very bright QSOs are selected. In this field, enormous progress has been made since the emergence of 10 m-class telescopes.

As mentioned before, the line density in the Lyα forest is a strong function of the absorption redshift. The number density of Lyα absorption lines with equivalent width (in the rest frame of the absorber) $W \geq 0.32$ Å at $z \gtrsim 2$ is found to follow

$$\frac{\mathrm{d}N}{\mathrm{d}z} \sim k(1+z)^{\gamma}\ , \tag{8.29}$$

with $\gamma \sim 2.5$ and $k \sim 4$, which implies a strong redshift evolution. At lower redshift, where the Lyα forest is located in the UV part of the spectrum and therefore is considerably more difficult to observe (only by UV-sensitive satellites), the evolution is slower and the number density deviates from the power law given above.

From the line strength and width, the HI column density $N_{\rm HI}$ of a line can be measured. The number density of lines as a function of $N_{\rm HI}$ is

$$\frac{\mathrm{d}N}{\mathrm{d}N_{\rm HI}} \propto N_{\rm HI}^{-\beta}\ , \tag{8.30}$$

with $\beta \sim 1.6$. This power law approximately describes the distribution over a wide range of column densities, $10^{12}\,\mathrm{cm}^{-2} \lesssim N_{\rm HI} \lesssim 10^{22}\,\mathrm{cm}^{-2}$, including Ly-limit systems and damped Lyα systems.

The temperature of the absorbing gas can be estimated from the line width as well, by identifying the width with the thermal line broadening. As typical values, one obtains $\sim 10^4$ K to 2×10^4 K which, however, are somewhat model-dependent.

The proximity effect. The statistical properties of the Lyα forest depend only on the redshift of the absorption lines, and not on the redshift of the QSO in the spectrum of which they are measured. This is as expected if the absorption is not physically linked to the QSO, and this observational fact is one of the most important indicators for an intergalactic origin of the absorption.

However, there is one effect in the statistics of Lyα absorption lines which is directly linked to the QSO. One finds that the number density of Lyα absorption lines at those redshifts which are only slightly smaller than the emission line redshift of the QSO itself, is lower than the mean absorption line density at this redshift (averaged over many different QSO lines-of-sight). This effect indicates that the QSO has some effect on the absorption lines, if only in its immediate vicinity; for this reason, it is named the *proximity effect*. An explanation of this effect follows directly from considering the ionization stages of hydrogen. The gas is ionized by energetic photons which originate from hot stars

and AGNs and which form an ionizing background. On the other hand, ionized hydrogen can recombine. The degree of ionization results from the equilibrium between these two processes.

The number of photoionizations of hydrogen atoms per volume element and unit time is proportional to the density of neutral hydrogen atoms and given by

$$\dot{n}_{\rm ion} = \Gamma_{\rm HI}\, n_{\rm HI} \ , \tag{8.31}$$

where $\Gamma_{\rm HI}$, the photoionization rate, is proportional to the density of ionizing photons. The corresponding number of recombinations per volume and time is proportional to the density of free protons and electrons,

$$\dot{n}_{\rm rec} = \alpha\, n_{\rm p}\, n_{\rm e} \ , \tag{8.32}$$

where the recombination coefficient α depends on the gas temperature. The Gunn–Peterson test tells us that the intergalactic medium is essentially fully ionized, and thus $n_{\rm HI} \ll n_{\rm p} = n_{\rm e} \approx n_{\rm b}$ (we disregard the contribution of helium in this consideration). We then obtain for the density of neutral hydrogen in an equilibrium of ionization and recombination

$$n_{\rm HI} = \frac{\alpha}{\Gamma_{\rm HI}}\, n_{\rm p}^2 \ . \tag{8.33}$$

This results shows that $n_{\rm HI}$ is inversely proportional to the number density of ionizing photons. However, the intergalactic medium in the vicinity of the QSO does not only experience the ionizing background radiation field but, in addition, the energetic radiation from the QSO itself. Therefore, the degree of ionization of hydrogen in the immediate vicinity of the QSO is higher, and consequently less Lyα absorption can take place there.

Since the contribution of the QSO to the ionizing radiation depends on the distance of the gas from the QSO ($\propto r^{-2}$), and since the spectrum and ionizing flux of the QSO is observable, examining the proximity effect provides an estimate of the intensity of the ionizing background radiation as a function of redshift. This value can then be compared to the total ionizing radiation which is emitted by QSOs and young stellar populations at the respective redshift. This comparison, in which the luminosity function of AGNs and the star-formation rate in the Universe are taken into account, yields good agreement, thus confirming our model for the proximity effect.

Transverse proximity effect. A similar effect should occur if there is a QSO situated near the line-of-sight to a more distant QSO. The ionizing radiation from the foreground QSO should lead to a decrease of the neutral hydrogen on the line-of-sight to the distant QSO at the redshift of the foreground QSO. Hence, by studying the Lyman-α forest of QSOs which have a lower-redshift neighbor at small angular separation, one would expect to see this transverse proximity effect.

As it turns out, evidence for the occurrence of the transverse proximity effect is sparse. Even in cases where the impact of the ionizing radiation from the foreground QSO is estimated to be several times larger than the ionizing background radiation, the effect has not been detected.

However, there are several caveats in the estimate of the effect. Perhaps the most important one is that it depends on the luminosity of the foreground QSO. This we can determine from the observed flux and its redshift—but we determine the luminosity at the time when the light of the distant QSO passed the foreground QSO. In order to affect the intergalactic medium at the sight-line towards the background QSO, one would need the luminosity of the foreground QSO at an earlier time—namely a time shifted by the light travel time from the foreground QSO to the line-of-sight towards the background QSO. With characteristic transverse separations of several Mpc, the time shift is of the order of 10^7 yr. We have seen that QSOs are variable on short time-scales—and from studies of the QSO population one expects ever stronger variability on the time-scales relevant here. A second possibility is that the ionizing radiation from the foreground QSO is highly anisotropic, so that only a small fraction of the luminosity we see actually arrives at the line-of-sight to the distant QSO.

Overall, it is likely that a study of the transverse proximity effect will tell us more about QSOs than about the intergalactic medium.

8.5.3 Models of the Lyman-α forest

Since the discovery of the Lyα forest, various models have been developed in order to explain its nature. Since about the mid-1990s, one model has been established that is directly linked to the evolution of large-scale structure in the Universe.

The 'old' model of the Lyman-α forest. Prior to this time, models were designed in which the Lyα forest was caused by quasi-static hydrogen clouds. These clouds (Lyα clouds) were postulated and were initially seen as a natural picture given the discrete nature of the absorption lines. From the statistics of the number density of lines, the cloud properties (such as radius and density) could then be constrained. If the line width represented a thermal velocity distribution of the atoms, the temperature and, together with the radius, also the mass of the clouds could be derived (e.g., by utilizing the density profile of an isothermal sphere). The conclusion from

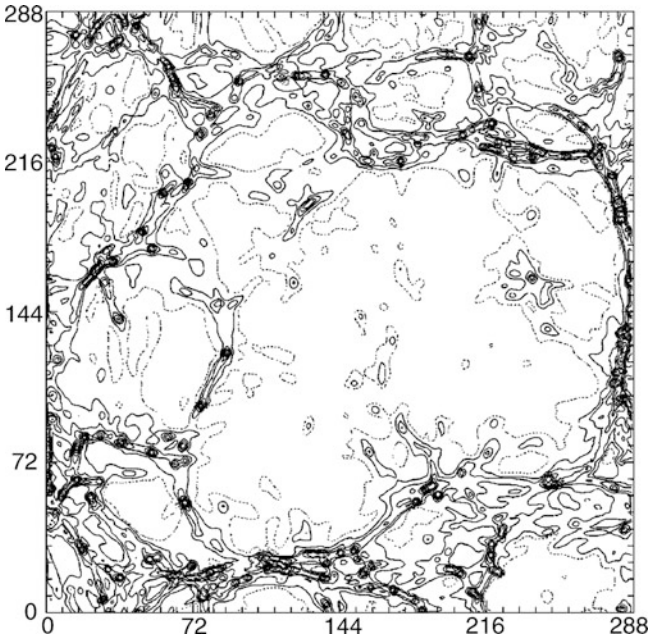


Fig. 8.33 Column density of neutral hydrogen, computed in a joint simulation of dark matter and gas. The size of the cube displayed here is $10h^{-1}$ Mpc (comoving). By computing the Lyα absorption of photons crossing a simulated cube like this, simulated spectra of the Lyα forest are obtained, which can then be compared statistically with observed spectra. Source: J. Miralda-Escudé et al. 1996, *The Ly alpha Forest from Gravitational Collapse in the Cold Dark Matter + Lambda Model*, ApJ 471, 582, p. 587, Fig. 2. ©AAS. Reproduced with permission

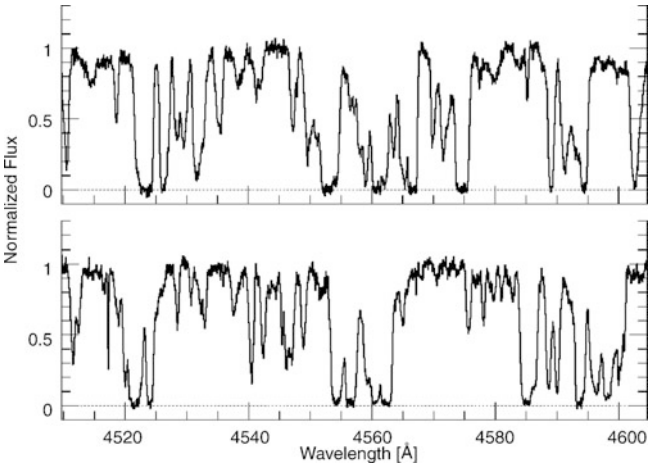


Fig. 8.34 One of the spectra is a section of the Lyα forest towards the QSO 1422 + 231 (see also Fig. 5.55), the other is a simulated spectrum; both are statistically so similar that it is impossible to distinguish them—which one is which? Source: R. Davé 2001, *The Evolution of the Lyman Alpha Forest From Redshift 3 to 0*, astro-ph/0105085, Fig. 1

these considerations was that such clouds would evaporate immediately unless they were gravitationally bound in a dark matter halo (mini-halo model), or confined by the pressure of a hot intergalactic medium.[10]

The new picture of the Lyman-α forest. Since the mid 1990s, a new paradigm has existed for the nature of the Lyα forest. Its establishment became possible through advances in hydrodynamic cosmological simulations.

We discussed structure formation in Chap. 7, where we concentrated mainly on dark matter. After recombination at $z \sim 1100$ when the Universe became neutral and therefore the baryonic matter no longer experienced pressure by the photons, baryons were, just like dark matter, only subject to gravitational forces. Hence the behavior of baryons and dark matter became very similar up to the time when baryons began to experience significant pressure forces by heating (e.g., due to photoionization) and compression. The spatial distribution of baryons in the intergalactic medium thus followed that of dark matter, as is also confirmed by numerical simulations. In these simulations, the intensity of ionizing radiation is accounted for—it is estimated, e.g., from the proximity effect. Figure 8.33 shows the column density distribution of neutral hydrogen which results from such a simulation. It shows a structure similar to the distribution of dark matter, however with a higher density contrast due to the quadratic dependence of the HI density on the baryon density—see (8.33).

From the distribution of neutral gas simulated this way, synthetic absorption line spectra can be computed. For these, the temperature of the gas and its peculiar velocity are used, the latter obtained from the simulation as well. Such a synthetic spectrum is displayed in Fig. 8.34, together with a measured Lyα spectrum. These two spectra are, from a statistical point of view, virtually identical, i.e., their density of lines, the width and optical depth distributions, and their correlation properties are equal. For this reason, the evolution of cosmic structure provides a natural explanation for the Lyα forest, without the necessity of additional free parameters or assumptions. In this model, the evolution of $\mathrm{d}N/\mathrm{d}z$ is driven mainly by the Hubble expansion and the resulting change in the degree of ionization in the intergalactic medium. In contrast to the 'old' model, no gas clouds need to be postulated, and the absorption occurs mainly in sheets and filaments, instead of isolated 'clouds'.

Besides the correlation properties of the Lyα lines in an individual QSO spectrum, we can also consider the correlation between absorption line spectra of QSOs which have a small angular separation on the sky. In this case, the corresponding light rays are close together, probing neighboring spatial regions of the intergalactic medium. If the neutral hydrogen is correlated on scales larger than the

[10]The latter assumption was excluded at last by the COBE measurements of the CMB spectrum, because such a hot intergalactic medium would cause deviations of the CMB spectrum from its Planck shape, by Compton scattering of the CMB photons.

transverse separation of the two lines-of-sight towards the QSOs, correlated Lyα absorption lines should be observable in the two spectra. As a matter of fact, it is found that the absorption line spectra of QSOs show correlations, provided that the angular separation is sufficiently small. The correlation lengths derived from these studies are $\gtrsim 100h^{-1}\,\mathrm{kpc}$, in agreement with the results from numerical simulations. In particular, the lines-of-sight corresponding to different images of multiple-imaged QSOs in gravitational lens systems are very close together, so that the correlation of the absorption lines in these spectra can be very well verified.

Where are the baryons located? As another result of these investigations it is found that at $2 \lesssim z \lesssim 4$ the majority ($\sim 85\,\%$) of baryonic matter is contained in the Lyα forest, mainly in systems with column densities of $10^{14}\,\mathrm{cm}^{-2} \lesssim N_{\mathrm{HI}} \lesssim 3 \times 10^{15}\,\mathrm{cm}^{-2}$. Thus, at these high redshifts we observe nearly the full inventory of baryons. At lower redshift, this is no longer the case. Indeed, only a fraction of the baryons can be observed in the local Universe, for instance in stars or in the intergalactic gas in clusters of galaxies. From theoretical arguments, we expect that the majority of baryons today should be found in the form of intergalactic gas, for example in galaxy groups and large-scale filaments that are seen in simulations of structure formation. This gas is expected to have a temperature between $\sim 10^5\,\mathrm{K}$ and $\sim 10^7\,\mathrm{K}$ and is therefore very difficult to detect; it is called the warm-hot intergalactic medium. At these temperatures, the gas is essentially fully ionized so that it cannot be detected in absorption line spectra, whereas the temperature and density are too low to expect significant X-ray emission from this gas.[11]

8.5.4 The Lyα forest as cosmological tool

The aforementioned simulations of the Lyα forest predict that most of the lines originate in regions of the intergalactic medium where the gas density is $\rho_{\mathrm{g}} \lesssim 10\bar{\rho}_{\mathrm{g}}$. Hence, the density of the absorbing gas is relatively low, compared, e.g., to the average gas density in a galaxy. The temperature of the gas causing the absorption is about $\sim 10^4\,\mathrm{K}$. At these densities and temperatures, pressure forces are small compared to gravitational forces, so that the gas follows the density distribution of dark matter very closely. From the absorption line statistics, it is therefore possible to derive the statistical properties of the dark matter distribution. More precisely, the two-point correlation function of the Lyα lines reflects the spectrum of density fluctuations in the Universe, and hence it can be used to measure the power spectrum $P(k)$.

The relation between density and absorption optical depth. We will consider some aspects of this method in more detail. The temperature of the intergalactic gas is not homogeneous because gas heats up by compression. Thus at a fixed redshift dense gas is hotter than the average baryon temperature T_0. As long as the compression proceeds adiabatically, T basically depends on the density, $T = T_0(\rho_{\mathrm{g}}/\bar{\rho}_{\mathrm{g}})^{\alpha}$, where T_0 and the exponent α depend on the ionization history and on the spectrum of the ionizing photons. Typical values are $4000\,\mathrm{K} \lesssim T_0 \lesssim 10\,000\,\mathrm{K}$ and $0.3 \lesssim \alpha \lesssim 0.6$. The density of neutral hydrogen is specified by (8.33), $n_{\mathrm{HI}} \propto \rho_{\mathrm{g}}^2 T^{-0.7}/\Gamma_{\mathrm{HI}}$, where the temperature dependence of the recombination rate was taken into account. Since the temperature depends on the density, one obtains for the optical depth of Lyα absorption

$$\tau = A\left(\frac{\rho_{\mathrm{g}}}{\bar{\rho}_{\mathrm{g}}}\right)^{\beta}\,, \tag{8.34}$$

where $\beta = 2 - 0.7\alpha \approx 1.6$, with the prefactor depending on the observed redshift, the ionization rate Γ_{HI}, and the average temperature T_0.

In Fig. 8.35, the distribution of the optical depth and gas density at redshift $z = 3$ is plotted, obtained from a hydrodynamical simulation. As is seen from the right-hand panel, the distribution follows the relation (8.34) very closely, which means that a major fraction of the gas was not heated by shock fronts, but rather by adiabatic compression. Even with peculiar motion of the gas and thermal broadening taken into account, as is the case in the panel on the left, the average distribution still follows the analytical relation very closely. One caveat of this method is that (8.34) neglects possible spatial fluctuations of the ionizing background which would show up as fluctuations in Γ_{HI} and thus in the prefactor A—hence, a spatially varying ionizing background causes additional structure in the Lyman-α forest.

Measuring the power spectrum. From the observed distribution of τ, it is thus possible to draw conclusions about the distribution of the gas overdensity $\rho_{\mathrm{g}}/\bar{\rho}_{\mathrm{g}}$. As argued above, the latter is closely related to the corresponding overdensity of dark matter. From an absorption line spectrum, $\tau(\lambda)$ can be determined (wavelength-)pixel by pixel, where λ corresponds, according to $\lambda = (1+z)\,1216\,\text{Å}$,

[11] Although hydrogen is not detectable in this intergalactic medium due to its complete ionization, lines from metal ions at a high ionization stage can be observed in UV absorption line spectra, for instance the lines of OVI, five times ionized oxygen. To derive a baryon density from observations of these lines, assumptions about the temperature of the gas and about its metallicity are required. The latest results, which have mainly been obtained using the UV satellite FUSE, are compatible with the idea that today the major fraction of baryons is contained in this warm-hot intergalactic medium.

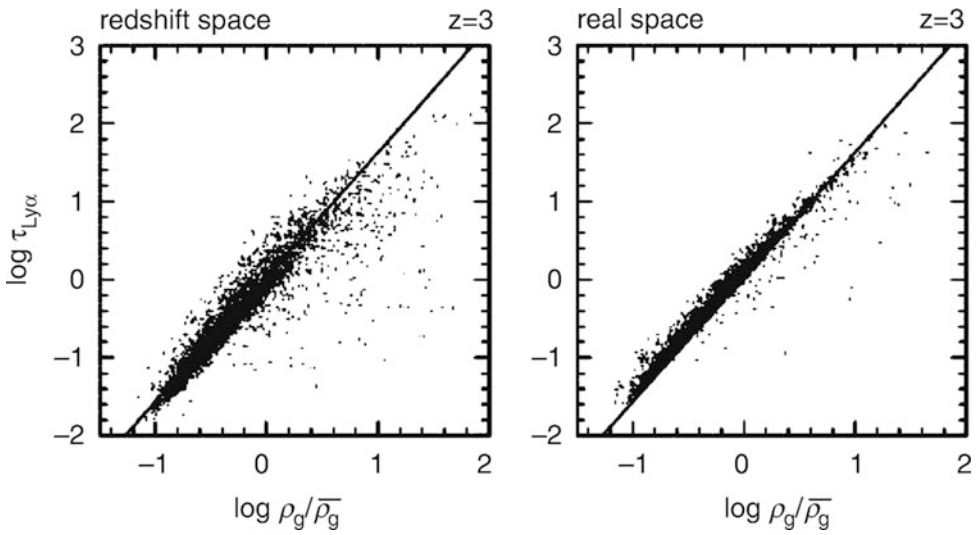


Fig. 8.35 Optical depth for Lyα absorption versus gas density, obtained from a cosmological simulation. Each data point represents a line-of-sight through a gas distribution like the one presented in Fig. 8.33. For the *panel on the right*, peculiar motion of the gas was neglected; in this case, the points follow the relation (8.34) very accurately. With the peculiar motions and thermal line broadening taken into account (*left panel*), the points also follow this relation on average. Source: D. Weinberg et al. 1998, *Cosmology with the Lyman-alpha Forest*, astro-ph/9810142, Fig. 1

to a distance along the line-of-sight, at least if peculiar velocities are disregarded. From $\tau(\lambda)$, the overdensity as a function of this distance follows with (8.34), and thus a one-dimensional cut through the density fluctuations is obtained. The correlation properties of this density are determined by the power spectrum of the matter distribution, which can be measured in this way.

This probe of the density fluctuations is applied at redshifts $2 \lesssim z \lesssim 4$, where, on the one hand, the Lyα forest is in the optical region of the observed spectrum, and on the other hand, the forest is not too dense for this analysis to be feasible. This technique therefore probes the large-scale structure at significantly earlier epochs than is the case for the other cosmological probes described earlier. At such earlier epochs the density fluctuations are linear down to smaller scales than they are today. For this reason, the Lyα forest method yields invaluable information about the power spectrum on smaller scales than can be probed with, say, galaxy redshift surveys.

With the spectra from the SDSS and its follow-up projects, the number of high-redshift QSOs with well-measured spectra increased tremendously. For each QSO in the appropriate redshift range $2 \lesssim z \lesssim 4$, the density can be measured over an appreciable redshift range. Due to the high density of these QSO sight-lines, a quasi three-dimensional distribution of the gas is obtained, which yields significantly better information about the power spectrum of the density fluctuations than one-dimensional density fields obtain from individual QSOs.

Results. Two results from such studies should be mentioned here. The fact that this technique can measure the power spectrum of density fluctuations to much smaller scales than possible from redshift surveys gives it a particular sensitivity to the properties of dark matter. In Sect. 7.8, we discussed the possibility that the dark matter is not cold, but instead 'warm'; this would erase small-scale fluctuations due to the free-streaming of these particles, and has been suggested as a possible solution for the 'substructure problem'. The strongest constraints on the properties of warm dark matter indeed comes from studies of the Lyman-α forest at small length-scales and high redshift. Provided the warm dark matter particle is a thermal relic from the Big Bang, a lower bound of its mass of $\sim$ 3 keV was derived. Given this mass bound, the formation of small-mass halos, and thus the substructure in galactic halos, is suppressed only for masses lower than $\sim 2 \times 10^8 h^{-1}\, M_\odot$. Given that the abundance mismatch between predicted subhalos and the observed satellites in the Milky Way starts at much larger masses, the warm dark matter model essentially has lost all its potential appeal.

A second result to be mentioned here is that from $\sim$ 50 000 QSOs with $2.1 \leq z \leq 3.5$, baryonic acoustic oscillations in the intergalactic gas were detected and investigated quantitatively. Measuring the BAOs at such redshifts thus allows one to determine the distance-redshift relation out to $z \sim 2.5$. Since $\mathrm{d}x = c\,\mathrm{d}z/H(z)$, the measurement of the BAOs in the radial direction can be used to determine $H(z)$ directly. Furthermore, the (luminosity) distance which is determined in supernova cosmology is an integral over the Hubble function, so that by differentiation, $H(z)$ can be determined from these results as well. An estimate of $\dot{a} = H\,a = H(z)/(1+z)$ from these various techniques is shown in Fig. 8.36. This way of presenting the results yields perhaps the clearest view of the fact that the Universe changed from a decelerating expansion for $z \gtrsim 0.8$ to an accelerated one at lower redshifts, showing the dominance of dark energy at the later stages of cosmic evolution.

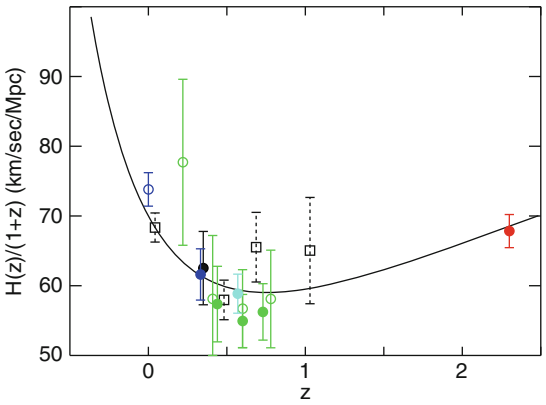


Fig. 8.36 The derivative of the scale factor, $\dot{a} = aH(a) = H(z)/(1+z)$, as a function of redshift, as determined from various methods. The *filled circles* are from BAO measurements, and most of them correspond to data displayed in Fig. 8.8, except the *red* point which is the measurement from BAOs in the Lyman-α forest. The *blue open circle* is a local determination of the Hubble constant, whereas the other points are from supernova studies. The plot clearly indicates that $\dot{a}$ has a minimum at $z \sim 0.8$; at larger redshifts, it is a decreasing function of time, i.e., $\ddot{a} < 0$, whereas for $z \lesssim 0.8$, $\ddot{a} > 0$, i.e., the Universe is accelerating. The *solid curve* corresponds to a model with $\Omega_{\rm m} = 0.27$, $\Omega_\Lambda = 0.73$ and $h = 0.7$. Source: N.G. Busca et al. 2013, *Baryon acoustic oscillations in the Lyα forest of BOSS quasars*, A&A 552, A96, p. 14, Fig. 21. ©ESO. Reproduced with permission

8.6 Angular fluctuations of the CMB

The cosmic microwave background consists of photons that last interacted with matter at $z \sim 1100$. Since the Universe must already have been inhomogeneous at this time, in order for the structures present in the current Universe to be able to form, it is expected that these spatial inhomogeneities are visible as a (small) anisotropy of the CMB: the angular distribution of the CMB temperature reflects the matter inhomogeneities at the redshift of decoupling of radiation and matter.

Since the discovery of the CMB in 1965, such anisotropies have been searched for. Under the assumption that the matter in the Universe solely consists of baryons, the expectation was that we would find relative fluctuations in the CMB temperature of amplitude $\Delta T/T \sim 10^{-3}$ on scales of a few arcminutes. This expectation is based on the theory of gravitational instability for structure growth: to account for the density fluctuations observed today where the density contrast $\delta \sim 1$ on scales of $\sim 10h^{-1}$ Mpc, one needs relative density fluctuations at $z \sim 1000$ of order $D_+(z = 1000) \sim 10^{-3}$. Despite increasingly more sensitive observations, such fluctuations were not detected. The upper limits resulting from these searches for anisotropies provided one of the arguments that, in the mid-1980s, caused the idea of the existence of dark matter on cosmic scales to increasingly enter the minds of cosmologists. As we will see soon, in a universe which is dominated by dark matter the expected CMB fluctuations on small angular scales are considerably smaller than in a purely baryonic universe. It was the COBE satellite with which temperature fluctuations in the CMB were finally observed in 1992 (Fig. 1.21), a discovery which was awarded the Physics Nobel Prize in 2006. Over the following years, sensitive and significant measurements of the CMB anisotropy were carried out using balloons and ground-based telescopes. Two more satellites have observed the full microwave sky, WMAP and Planck; their results, together with ground-based measurements at smaller angular scales, have yielded the most stringent constraints on cosmological parameters yet.

We will first describe the physics of CMB anisotropies, before turning to the observational results and their interpretation. As we will see, the CMB anisotropies depend on nearly all cosmological parameters, such as $\Omega_{\rm m}$, $\Omega_{\rm b}$, Ω_Λ, $\Omega_{\rm HDM}$, H_0, the normalization σ_8, the primordial slope $n_{\rm s}$, and the shape parameter Γ of the power spectrum. Therefore, from an accurate mapping of the angular distribution of the CMB and by comparison with theoretical expectations, all these parameters can, in principle, be determined.

8.6.1 Origin of the anisotropy: Overview

The CMB anisotropies reflect the conditions in the Universe at the epoch of recombination, thus at $z \sim 1100$. Temperature fluctuations originating at this time are called *primary anisotropies*. Later, as the CMB photons propagate through the Universe, they may experience a number of distortions along their way which, again, may change their temperature distribution on the sky. These effects then lead to *secondary anisotropies*.

Primary anisotropies. The most basic mechanisms causing primary anisotropies can be divided into those which occur on scales larger than the horizon size at recombination, i.e., which can not have been affected by physical interactions up to the time of last scattering, and those on smaller scales. The effects on superhorizon scales are the following:

- Inhomogeneities in the gravitational potential cause photons which originate in regions of higher density to climb out of a potential well. As a result of this, they lose energy and are redshifted (gravitational redshift). This effect is partly compensated for by the fact that, besides the gravitational redshift, a gravitational time delay also occurs: a photon that originates in an overdense region will be scattered at a slightly earlier time, and thus at a slightly higher temperature of the Universe, compared to a photon from a region of average density. Both effects always occur side by side. They are combined under the term *Sachs–Wolfe effect*. Its separation into two processes

is necessary only in a simplified description; a general relativistic treatment of the Sachs–Wolfe effect jointly yields both processes.

- We have seen that density fluctuations are always related to peculiar velocities of matter. Hence, the electrons that scatter the CMB photons for the last time do not follow exactly the Hubble expansion, but have an additional velocity that is closely linked to the density fluctuations (compare Sect. 7.2.3). This results in a Doppler effect: if photons are scattered by gas receding from us with a speed larger than that corresponding to the Hubble expansion, these photons experience an additional redshift which reduces the temperature measured in that direction.
- On scales larger than the horizon scale at recombination (see Sect. 4.5.2), the distribution of baryons follows that of the dark matter, so that in regions of a higher dark matter density, the baryon density is also enhanced. This leads to an increased temperature of the baryons in overdense regions.

These three effects are relevant on scales larger than the (sound) horizon scale at the epoch of recombination. Obviously, they are closely coupled to each other. In particular, on scales $> r_{\mathrm{H,com}}(z_{\mathrm{rec}})$ the first two effects can partially compensate each other, though the Sachs–Wolfe effect is the dominant one at superhorizon scales. Inside the (sound) horizon,[12] two other effects dominate the primary anisotropy signal:

- On subhorizon scales, the pressure of the baryon-photon fluid is effective because, prior to recombination, these two components had been closely coupled by Compton scattering. As we discussed in Sect. 7.4.3, this leads to sound waves in the baryon-photon fluid, the baryonic acoustic oscillations. In the density peaks of these sound waves, the baryon-photon fluid is adiabatically compressed and thus hotter than the average. The CMB sky yields a two-dimensional cut through this three-dimensional density (and temperature) field of these sound waves, and thus reflect these fluctuations, yielding temperature anisotropies with characteristic length (or angular) scales (see also Fig. 7.7).
- The coupling of baryons and photons is not perfect since, owing to the finite mean-free path of photons, the two components are decoupled on small spatial scales. This implies that on small length-scales, the temperature fluctuations can be smeared out by the diffusion of photons. This process is known as *Silk damping*, and it implies that on angular scales below about $\sim 5'$, only very small primary fluctuations exist.

Secondary anisotropies result, among other things, from the following effects:

- Thomson scattering of CMB photons. Since the Universe is currently transparent for optical photons (since we are able to observe UV-radiation from objects at $z > 6$), it must have been reionized between $z \sim 1000$ and $z \sim 6$, presumably by radiation from the very first generation of stars and/or by the first QSOs. After this reionization, free electrons are available again, which may then scatter the CMB photons. Since Thomson scattering is essentially isotropic, the direction of a photon after scattering is nearly independent of its incoming direction. This means that scattered photons no longer carry information about the CMB temperature fluctuations. Hence, the scattered photons form an isotropic radiation component whose temperature is the average CMB temperature. The radiation we observe in any direction therefore consists of a fraction f_{sc} which has undergone scattering, and a fraction $(1-f_{\mathrm{sc}})$ of unscattered radiation. The main effect resulting from this scattering is a reduction of the amplitude of the measured temperature anisotropies by a factor $(1 - f_{\mathrm{sc}})$.
- Photons propagating towards us are traversing a Universe in which structure formation takes place. Due to this evolution of the large-scale structure, the gravitational potential is changing over time. If it was time-independent, photons would enter and leave a potential well with their frequency being unaffected, compared to photons that are propagating in a homogeneous universe: the blueshift they experience when falling into a potential well is exactly balanced by the redshift they suffer when climbing out. However, this 'conservation' of photon energy no longer applies if the potential varies with time. One can show that for an Einstein–de Sitter model, the peculiar gravitational potential ϕ (7.10) is constant over time,[13] and hence, the light propagation in the evolving universe yields no net frequency shift. For other cosmological models this effect does occur; it is called the *integrated Sachs–Wolfe (ISW) effect*.
- The gravitational deflection of CMB photons, caused by the gravitational field of the cosmic density fluctuations, leads to a change in the photon direction. This means that two lines-of-sight separated by an angle θ at the observer have a physical separation at recombination which may be different from $D_{\mathrm{A}}(z_{\mathrm{rec}})\theta$, due to the gravitational light deflection. Because of this, the correlation function of the temperature fluctuations is slightly smeared out. This effect is relevant on small angular scales.
- The Sunyaev–Zeldovich effect, which we discussed in Sect. 6.4.4 in the context of galaxy clusters, also affects the temperature distribution of the CMB. Some of the photons propagating along lines-of-sight passing through clusters of galaxies or other regions of dense and hot

[12] We recall that the sound horizon is of the same order as the (event) horizon, since the sound velocity in the baryon-photon fluid is $\sim c/\sqrt{3}$.

[13] This is seen with (7.10) due to the dependence $\bar{\rho} \propto a^{-3}$ and $\delta \propto D_+ = a$ for an EdS model.

gas are scattered by the hot electrons, resulting in a temperature change in these directions. We recall that in the direction of clusters the measured intensity of the CMB radiation is reduced at low frequencies, whereas it is increased at high frequencies. Hence, the SZ effect can be identified in the CMB data if measurements are conducted over a sufficiently large frequency range, whereas the ISW and gravitational lensing effects preserve the Planck spectrum.

8.6.2 Description of the CMB anisotropy

Correlation function and power spectrum. In order to characterize the statistical properties of the angular distribution of the CMB temperature, the two-point correlation function of the temperature on the sphere can be employed, in the same way as it is used for describing the density fluctuations or the angular correlation function of galaxies. To do this, the relative temperature fluctuations $\mathcal{T}(\boldsymbol{n}) = [T(\boldsymbol{n}) - T_0]/T_0$ are defined, where $\boldsymbol{n}$ is a unit vector describing the direction on the sphere, and T_0 is the average temperature of the CMB. The correlation function of the temperature fluctuations is then defined as

$$C(\theta) = \left\langle \mathcal{T}(\boldsymbol{n})\, \mathcal{T}(\boldsymbol{n}') \right\rangle \,, \tag{8.35}$$

where the average extends over all pairs of directions $\boldsymbol{n}$ and $\boldsymbol{n}'$ with angular separation θ. As for the description of the density fluctuations in the Universe, it is also common for the CMB to consider the power spectrum of the temperature fluctuations, instead of the correlation function.

We recall (see Sect. 7.3.2) that the power spectrum $P(k)$ of the density fluctuations is defined as the Fourier transform of the correlation function. However, exactly the same definition cannot be applied to the CMB. The difference here is that the density fluctuations $\delta(\boldsymbol{x})$ are defined on a flat space (approximately, at the relevant length-scales). In this space, the individual Fourier modes (plane waves) are orthogonal, which enables a decomposition of the field $\delta(\boldsymbol{x})$ into Fourier modes in an unambiguous way. In contrast to this, the temperature fluctuations $\mathcal{T}$ are defined on the sphere. The analog to the Fourier modes in a flat space are spherical harmonics on the sphere, a complete orthogonal set of functions into which $\mathcal{T}(\boldsymbol{n})$ can be decomposed.[14] On small angular scales, where a sphere can be considered locally flat, spherical harmonics approximately behave like plane waves. The power spectrum of temperature fluctuations, in most cases written as $\ell(\ell+1)C_\ell$, then describes the amplitude of the fluctuations on an angular scale $\theta \sim \pi/\ell = 180°/\ell$. $\ell = 1$ describes the dipole anisotropy, $\ell = 2$ the quadrupole anisotropy, and so on.

Line-of-sight projection. The CMB temperature fluctuations on the sphere result from projection, i.e., the integration along the line-of-sight of the three-dimensional temperature fluctuations which we discussed above. This integration also needs to account for the secondary effects, those in the propagation of photons from $z \sim 1100$ to us. Overall, this is a relatively complicated task that, moreover, requires the explicit consideration of some aspects of General Relativity. The necessity for this can clearly be seen by considering the fact that two directions which are separated by more than $\sim 1°$ have a spatial separation at recombination which is larger than the horizon size at that time—so spacetime curvature explicitly plays a role. Fortunately, the physical phenomena that need to be accounted for are (nearly) all of a linear nature. This means that, although the corresponding system of coupled equations is complicated, it can nevertheless straightforwardly be solved, since the solution of a system of linear equations is not a difficult mathematical problem. Generally accessible software packages exist (e.g., CMBFAST or CAMB[15]), which compute the power spectrum C_ℓ for any combination of cosmological parameters.

8.6.3 The fluctuation spectrum

Horizon scale. To explain the basic features of CMB fluctuations, we first point out that a characteristic length-scale exists at $z_{\rm rec}$, namely the horizon length. It is specified by (4.77). For cosmological models with $\Omega_\Lambda = 0$, the horizon spans an angle of—see (4.78)—

$$\theta_{\rm H,rec} \approx 1.8° \sqrt{\Omega_{\rm m}} \,.$$

This angle is modified for models with a cosmological constant; if the Universe is flat ($\Omega_{\rm m} + \Omega_\Lambda = 1$), one finds

$$\theta_{\rm H,rec} \approx 1.8° \,, \tag{8.36}$$

with a very weak dependence on the matter density, about $\propto \Omega_{\rm m}^{-0.1}$. As we will demonstrate in the following, this angular scale of the horizon is directly observable.

Fluctuations on large scales. On scales $\gg \theta_{\rm H,rec}$ the Sachs–Wolfe effect dominates, since sound waves in the baryon-photon fluid can occur only on scales below the (sound)

[14] Spherical harmonics are encountered in many problems in mathematical physics, for instance in the quantum mechanical treatment of the hydrogen atom or, more generally, in all spherically symmetric problems in physics.

[15] You can try it out at http://lambda.gsfc.nasa.gov/toolbox/tb_camb_form.cfm!

horizon length. For this reason, the CMB angular spectrum directly reflects the fluctuation spectrum $P(k)$ of matter. In particular, for a Harrison–Zeldovich spectrum, $P(k) \propto k$, one expects that

$$\ell(\ell+1)C_\ell \approx \text{const} \quad \text{for} \quad \ell \ll \frac{180^\circ}{\theta_{\text{H,rec}}} \simeq 100 \,,$$

and the amplitude of the fluctuations immediately yields the amplitude of $P(k)$. This flat behavior of the fluctuation spectrum for $n_\text{s} = 1$ is modified by the integrated Sachs–Wolfe effect, except for an Einstein–de Sitter model.

Sound horizon and acoustic peaks. On angular scales $<\theta_{\text{H,rec}}$, fluctuations are observed that were inside the horizon prior to recombination, hence physical effects may act on these scales. As already mentioned, the fluid of baryons and photons is dominated by the energy density of the photons. Their pressure prevents the baryons from falling into the potential wells of dark matter. Instead, this fluid oscillates. Since the energy density is dominated by photons, i.e., by relativistic particles, this fluid is relativistic and its sound speed is $c_\text{s} \approx c/\sqrt{3}$. Therefore, the maximum wavelength at which a wave may establish a full oscillation prior to recombination is the sound horizon (see Sect. 7.4.3)

$$r_\text{s} \simeq r_\text{H}(t_\text{rec})/\sqrt{3} \,, \tag{8.37}$$

where the estimate neglects the deviation of the sound speed from $c/\sqrt{3}$, i.e., a non-zero value of the parameter $\mathcal{R}$ which describes the ratio of baryons to photons [see (7.43)]. It corresponds to an angular scale of $\theta_1 = r_\text{s}/f_k(z_\text{rec}) \approx \theta_{\text{H,rec}}/\sqrt{3} \sim 1^\circ$, or $\ell_1 \sim 200$ for a flat cosmological model with $\Omega_\text{m} + \Omega_\Lambda = 1$. By the Doppler effect and by adiabatic compression, these oscillations generate temperature fluctuations that should be visible in the temperature fluctuation spectrum C_ℓ. Hence, $\ell(\ell+1)C_\ell$ should have a maximum at $\ell_1 \sim 200$; additional maxima are expected at integer multiples of ℓ_1. These maxima in the angular fluctuation spectrum are termed *acoustic peaks* (or Doppler peaks); their ℓ-values and their amplitudes are the most important diagnostics of the CMB anisotropies.

Silk damping. Since recombination is not instantaneous but extends over a finite range in redshift, CMB photons are last scattered within a shell of finite thickness. Considering a length-scale that is much smaller than the thickness of this shell, several maxima and minima of T are located within this shell along a line-of-sight. For this reason, the temperature fluctuations on these small scales are largely averaged out in the integration along the line-of-sight. The thickness of the recombination shell is roughly equal to the diffusion length of the photons, therefore this effect is relevant on the same length-scales as the aforementioned Silk damping. This means that on scales $\lesssim 5'$ ($\ell \gtrsim 2500$), one expects a damping of the anisotropy spectrum and, as a consequence, only very small (primary) temperature fluctuations on such small scales.

Model dependence of the fluctuation spectrum. Figure 8.37 shows the power spectra of CMB fluctuations where, starting from some reference model, individual cosmological parameters are varied. First we note that the spectrum is basically characterized by three distinct regions in ℓ (or in the angular scale). For $\ell \lesssim 100$, $\ell(\ell+1)C_\ell$ is a relatively flat function if—as in the figure—a Harrison–Zeldovich spectrum ($n_\text{s} = 1$) is assumed. In the range $\ell \gtrsim 100$, local maxima and minima can be seen that originate from the acoustic oscillations. For $\ell \gtrsim 2000$, the amplitude of the power spectrum strongly decreases due to Silk damping.

Figure 8.37a shows the dependence of the power spectrum on the curvature of the Universe, thus on $\Omega_\text{tot} = \Omega_\text{m} + \Omega_\Lambda$. We see that the curvature has two fundamental effects on the spectrum: first, the locations of the minima and maxima of the Doppler peaks are shifted, and second, the spectral shape at $\ell \lesssim 100$ depends strongly on Ω_tot. The latter is a consequence of the integrated Sachs–Wolfe (ISW) effect because the more the world model is curved, the stronger the time variations of the peculiar gravitational potential ϕ. The shift in the acoustic peaks is essentially a consequence of the change in the geometry of the Universe: the physical size of the sound horizon depends only weakly on the curvature (since the curvature term in the Friedmann equation is totally negligible in the pre-recombination era), but the angular diameter distance $D_\text{A}(z_\text{rec})$ is a very sensitive function of this curvature, so that the angular scale that corresponds to the sound horizon changes accordingly.

The dependence on the cosmological constant for flat models is displayed in Fig. 8.37b. Here one can see that the effect of Ω_Λ on the locations of the acoustic peaks is comparatively small, in accordance to what we said earlier: the peak locations depend most strongly on the curvature of the Universe. The most important influence of Ω_Λ is seen for small ℓ. For $\Omega_\Lambda = 0$, the ISW effect vanishes and the power spectrum is flat (for $n_\text{s} = 1$), whereas larger Ω_Λ always produce a pronounced ISW effect.

The influence of the baryon density is presented in Fig. 8.37c. An increase in the baryon density causes the amplitude of the first Doppler peak to rise, whereas that of the second peak decreases. In general, the amplitudes of the odd-numbered Doppler peaks increase, and those of the even-numbered peaks decrease with increasing $\Omega_\text{b}h^2$. Furthermore, the damping of fluctuations sets in at smaller ℓ (hence, larger angular scales) if Ω_b is reduced, since in this case the mean free path of photons increases, and so the fluctuations are smeared out over larger scales. We also

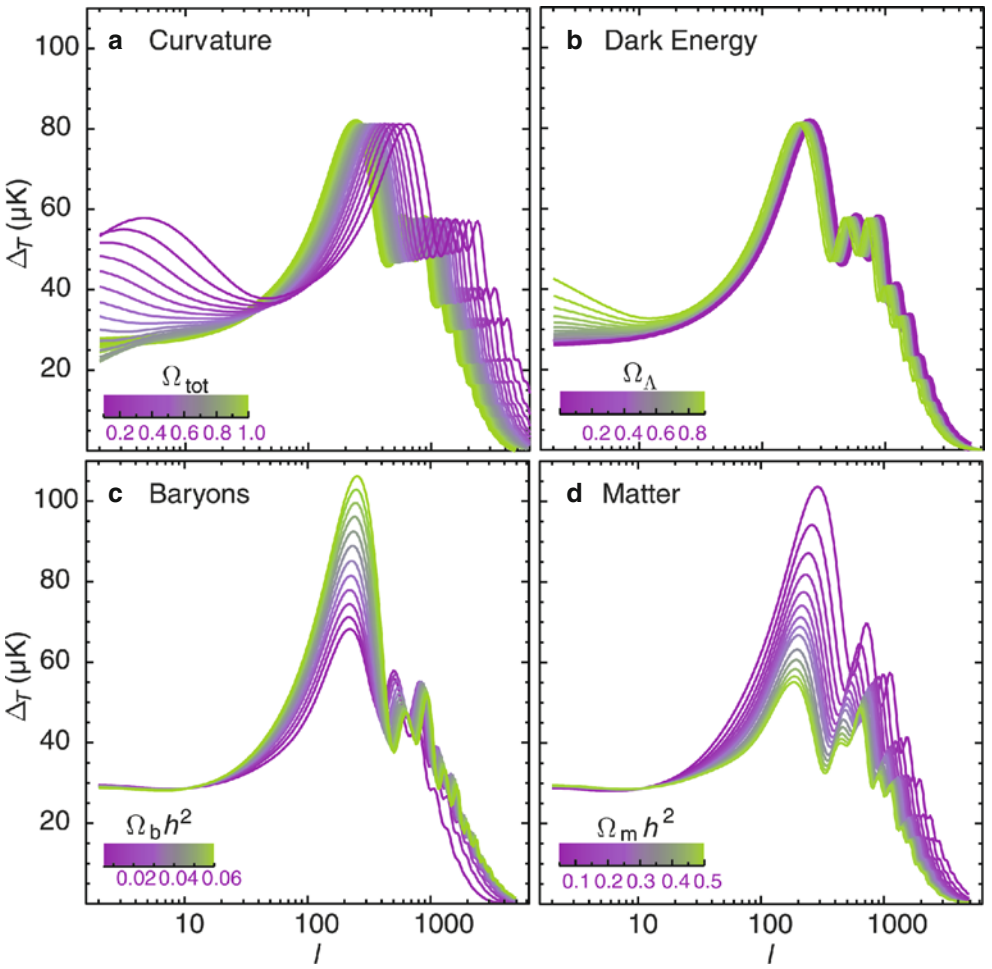


Fig. 8.37 Dependence of the CMB fluctuation spectrum on cosmological parameters. Plotted is the square root of the power per logarithmic interval in ℓ, $\Delta_T = \sqrt{\ell(\ell+1)C_\ell/(2\pi)}\,T_0$. These power spectra were obtained from an accurate calculation, taking into account all the processes previously discussed in the framework of perturbation theory in General Relativity. In all cases, the reference model is defined by $\Omega_m + \Omega_\Lambda = 1$, $\Omega_\Lambda = 0.65$, $\Omega_b h^2 = 0.02$, $\Omega_m h^2 = 0.147$, and a slope in the primordial density fluctuation spectrum of $n_s = 1$, corresponding to the Harrison–Zeldovich spectrum. In each of the four panels, one of these parameters is varied, as indicated, and the other three remain fixed. The various dependences are discussed in detail in the main text. Source: W. Hu & S. Dodelson 2002, *Cosmic Microwave Background Anisotropies*, ARA&A 40, 171, Fig. 4, PLATE. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 40 ©2002 by Annual Reviews www.annualreviews.org

see that with increasing baryon density, the peak location move towards larger ℓ, i.e., smaller angular scale. This is caused by the impact of baryons on the sound velocity in the baryon-photon fluid (7.42): for larger Ω_b, r_s decreases.

Finally, Fig. 8.37d demonstrates the dependence of the temperature fluctuations on the density parameter $\Omega_m h^2$. Changes in this parameter affect the epoch of matter-radiation equality a_{eq}, resulting in both, a shift of the locations of the Doppler peaks and in changes of their amplitudes. Furthermore, a reduction in the matter density implies a larger ratio of baryon to dark matter density, thus increasing the importance of the baryon-photon fluid.

From this discussion, it becomes obvious that the CMB temperature fluctuations contain an enormous amount of information about the cosmological parameters. Thus, from an accurate measurement of the fluctuation spectrum, tight constraints on these parameters can be obtained.

Secondary anisotropies. In Fig. 8.38, the secondary effects in the CMB anisotropies are displayed and compared to the reference model used above. Besides the already extensively discussed integrated Sachs–Wolfe effect, the influence of free electrons after reionization of the Universe has to be mentioned in particular. Scattering of CMB photons on these electrons essentially reduces the fluctuation amplitude on all scales, by a factor $e^{-\tau}$, where τ is the optical depth with respect to Thomson scattering. Here, $e^{-\tau} = 1 - f_{sc}$ is the probability that a photon from the CMB is *not* scattered by electrons after reionization, which depends on the reionization redshift: the earlier

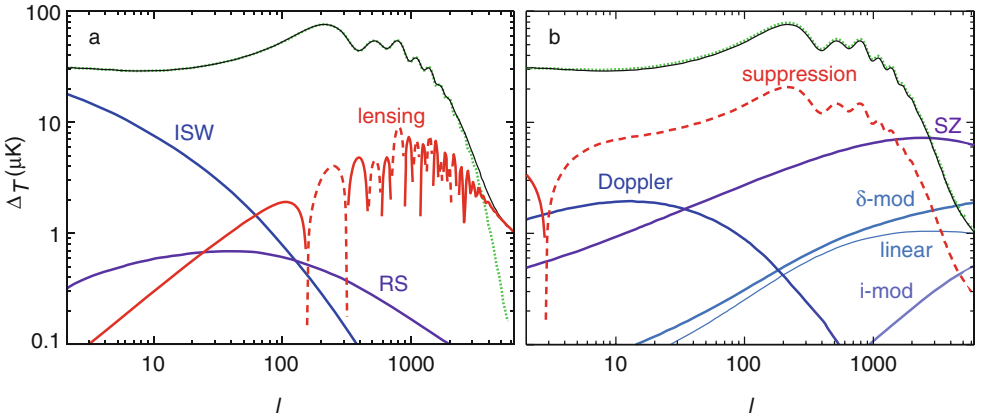


Fig. 8.38 The *uppermost curve* in each of the two panels shows the spectrum of primary temperature fluctuations for the same reference model as used in Fig. 8.37, whereas the other curves represent the effect of secondary anisotropies. In the *left panel*, secondary anisotropies due to purely gravitational effects are shown, whereas the *right panel* displays secondary anisotropies which are created by the interaction of photons with the electron/baryon component. Solid curves indicate an increase of the temperature fluctuations, dashed curves correspond to a decrease. Note that in contrast to Fig. 8.37, here the fluctuations are shown on a logarithmic scale. On large angular scales (small ℓ), the integrated Sachs–Wolfe effect dominates, whereas the effects of gravitational light deflection (lensing) and of the Sunyaev–Zeldovich effect (SZ) dominate at large ℓ. On intermediate angular scales, the scattering of photons by free electrons which are present in the intergalactic medium after reionization (curve labeled 'suppression') is the most efficient secondary process. Other secondary effects which are included in these plots are considerably smaller than the ones mentioned above and are thus of little interest here. Source: W. Hu & S. Dodelson 2002, *Cosmic Microwave Background Anisotropies*, ARA&A 40, 171, Fig. 7, PLATE. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 40 ©2002 by Annual Reviews www.annualreviews.org

the Universe was reionized, the larger is τ. Also visible in Fig. 8.38 is the fact that, on small angular scales, gravitational light deflection and the Sunyaev–Zeldovich effect become dominant. The identification of the latter is possible by its characteristic frequency dependence, whereas distinguishing the lens effect from other sources of anisotropies is possible since lensing changes the statistical properties of the CMB temperature field. Whereas the primary temperature fluctuations are assumed to obey Gaussian statistics, which implies that the four-point correlation function of the temperature field is a sum of products of two-point correlation functions, lensing changes this property, generating non-trivial fourth-order temperature correlations. By measuring those, the lensing effect can be identified.

Polarization of the CMB. The cosmic background radiation is blackbody radiation, and one would expect that it is therefore unpolarized. Whereas this is true to leading order, the CMB indeed is partly polarized, and this polarized component of the CMB has been measured. The origin of the polarization shall be explained in the following.

The scattering of photons on free electrons not only changes the direction of the photons, but also produces a linear polarization of the scattered radiation. The direction of this polarization is perpendicular to the plane spanned by the incoming and the scattered photons. If the radiation field is isotropic, then the net polarization of all scattered photons would add up to zero. However, if the radiation field as seen by the scattering electrons is anisotropic, a finite net polarization may occur. For example, if in the frame of the scattering electrons, there are more photons from 'the left and the right' than from 'above and below', a net polarization in the up-down direction would result. Such an anisotropy pattern is exactly what occurs in the presence of a quadrupole anisotropy. Not only we experience a quadrupole anisotropy of the CMB, but also the electrons in the last scattering surface, and in particular, the electrons present after reionization of the Universe. For that reason, the CMB is partially polarized. On large angular scales, this polarization is caused mainly by scattering at low redshifts, i.e., after reionization, and the degree of polarization depends on the fraction of CMB photons which undergo a scattering. Therefore, polarization of the CMB at large angles (small ℓ) allows us to measure the scattering optical through the post-reionization era, and thus to estimate the redshift at which the Universe became reionized.

8.6.4 Observations of the CMB anisotropy

To understand why so much time lies between the discovery of the CMB in 1965 and the first measurement of CMB fluctuations in 1992, we note that these fluctuations have a relative amplitude of $\sim 2 \times 10^{-5}$. The smallness of this effect means that in order to observe it, very high

precision is required. The main difficulty with ground-based measurements is emission by the atmosphere. To avoid this, or at least to minimize it, satellite experiments or balloon-based observations are strongly preferred. Hence, it is not surprising that the first detection of CMB fluctuations was made by the COBE satellite.[16] Besides mapping the temperature distribution on the sphere (see Fig. 1.21) with an angular resolution of $\sim 7°$, COBE also found that the CMB is the most perfect blackbody that had ever been measured. It found that the power spectrum for $\ell \lesssim 20$, the angular range accessible by COBE, is almost flat, and therefore compatible with the Harrison–Zeldovich spectrum.

Galactic foreground. The measured temperature distribution of the microwave radiation is a superposition of the CMB and of emission from Galactic (and extragalactic) sources. In the vicinity of the Galactic disk, this foreground emission dominates, which is clearly visible in Fig. 1.21, whereas it seems to be considerably weaker at higher Galactic latitudes. Due to its different spectral behavior, the foreground emission can be identified and subtracted. We note that the Galactic foreground basically consists of three components: synchrotron radiation from relativistic electrons in the Galaxy, thermal radiation by dust, and bremsstrahlung from hot gas. The synchrotron component defines a spectrum of about $I_\nu \propto \nu^{-0.8}$, whereas the dust is much warmer than 3 K and thus shows a spectral distribution of about $I_\nu \propto \nu^{3.5}$ in the spectral range of interest for CMB measurements. Bremsstrahlung has a flat spectrum in the relevant spectral region, $I_\nu \approx$ const. This can be compared to the spectrum of the CMB, which has a form $I_\nu \propto \nu^2$ in the Rayleigh–Jeans region.

There are two ways to extract the foreground emission from the measured intensity distribution. First, by observing at several frequencies the spectrum of the microwave radiation can be examined at any position, and the three aforementioned foreground components can be identified by their spectral signature and subtracted. As a second option, external datasets may be taken into account. At longer wavelengths, the synchrotron radiation is significantly more intense and dominates. From a sky map at radio frequencies, the distribution of synchrotron radiation can be obtained and its intensity at the frequencies used in the CMB measurements can be extrapolated. In a similar way, the infrared emission from dust, as measured, e.g., by the IRAS satellite (see Fig. 2.14), can be used to estimate the dust emission of the Galaxy in the microwave domain. Finally, one expects that gas that is emitting bremsstrahlung also shows strong Balmer emission of hydrogen, so that the bremsstrahlung pattern can be predicted from an Hα map of the sky. Both options, the determination of the foregrounds from multi-frequency data in the CMB experiment and the inclusion of external data, are utilized in order to obtain a map of the CMB which is as free from foreground emission as possible—which indeed seems to have been accomplished in the bottom panel of Fig. 1.21.

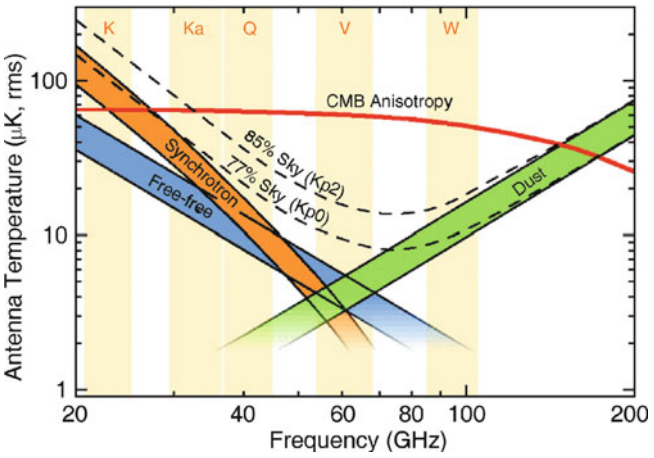


Fig. 8.39 The antenna temperature ($\propto I_\nu \nu^{-2}$) of the CMB and of the three foreground components discussed in the text, as a function of frequency. The five frequency bands of WMAP are marked. The *dashed curves* specify the average antenna temperature of the foreground radiation in the 77 and 85 % of the sky, respectively, in which the CMB analysis was conducted. We see that the three high-frequency channels are not dominated by foreground emission. Source: C.L. Bennett et al. 2003, *First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Foreground Emission*, ApJS 148, 97, p. 109, Fig. 10. ©AAS. Reproduced with permission

Besides these Galactic foregrounds, also extragalactic sources produce emission which affect the microwave sky. Among them are extragalactic radio sources—mainly AGN—whose synchrotron emission contributes to the low-frequency foreground, and dusty star-forming galaxies, relevant at higher frequencies. Known sources are masked out before a CMB analysis is conducted, whereas the unidentified sources need to be accounted for in a statistical way.

The optimal frequency for measuring the CMB anisotropies is where the foreground emission has a minimum; this is the case at about 70 GHz (see Fig. 8.39). Unfortunately, this frequency lies in a spectral region that is difficult to access from the ground.

From COBE to WMAP. In the years after the COBE mission, different experiments performed measurements of the anisotropy from the ground, focusing mainly on smaller angular scales. In around 1997, evidence was accumulating for the presence of the first Doppler peak, but the error bars of individual experimental results were too large at that time to clearly localize this peak. The breakthrough was then achieved in March 2000, when two groups published their

[16] with the exception of the dipole anisotropy, caused by the peculiar velocity of the Sun, which has an amplitude of $\sim 10^{-3}$; this was identified earlier

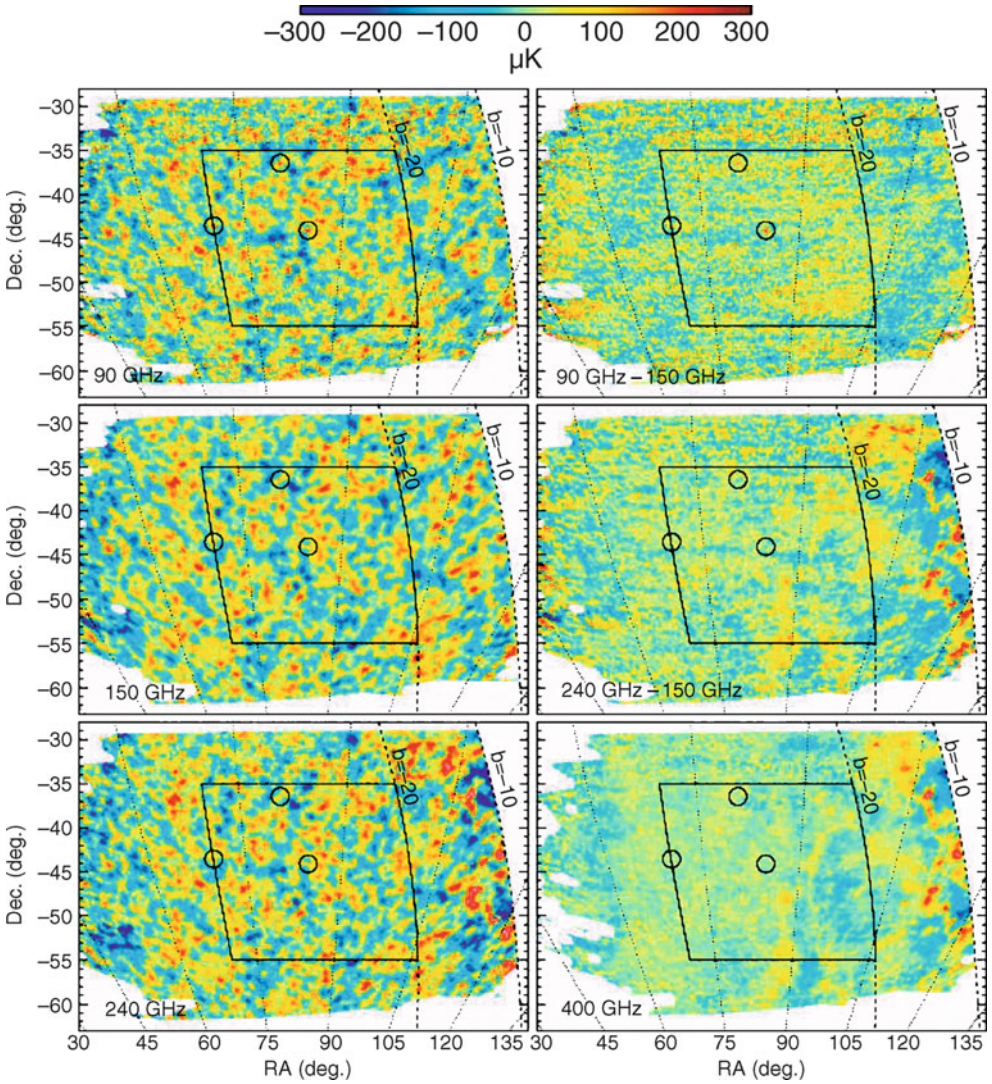


Fig. 8.40 In 2000, two groups published the results of their CMB observations, BOOMERANG and MAXIMA. This figure shows the BOOMERANG data. *On the left*, the temperature distributions at 90 GHz, 150 GHz, and 240 GHz are displayed, while the *lower right panel* shows that at 400 GHz. The *three small circles* in each panel denote the location of known strong point sources. The *two upper panels on the right* show the differences of temperature maps obtained at two different frequencies, e.g., the temperature map obtained with the 90 GHz data minus that obtained from the 150 GHz data. These difference maps feature considerably smaller fluctuations than the individual maps. This is compatible with the idea that the major fraction of the radiation originates in the CMB and not, e.g., in Galactic radiation which has a different spectral distribution and would thus be more prominent in the difference maps. Only the region within the *dashed rectangle* was used in the original analysis of the temperature fluctuations, in order to avoid boundary effects. The fluctuation spectrum computed from the difference maps is compatible with pure noise. Source: P. de Bernardis et al. 2000, *A Flat Universe from High-Resolution Maps of the Cosmic Microwave Background Radiation*, astro-ph/0004404, Fig. 1

CMB anisotropy results: BOOMERANG and MAXIMA. Both are balloon-based experiments, each observing a large region of the sky at different frequencies. In Fig. 8.40, the maps from the BOOMERANG experiment are presented. Both experiments have unambiguously measured the first Doppler peak, localizing it at $\ell \approx 200$. From this, it was concluded that we live in a nearly flat universe—the quantitative analysis of the data yielded $\Omega_m + \Omega_\Lambda \approx 1 \pm 0.1$. Furthermore, clear indications of the presence of the second Doppler peak were found.

In April 2001, refined CMB anisotropy measurements from three experiments were released, BOOMERANG, MAXIMA, and DASI. For the former two, the observational data were the same as published the year before, but improved analysis methods were applied; in particular, a better instrumental calibration was obtained. The resulting temperature fluctuation spectrum is presented in Fig. 8.41, demonstrating that it was now possible to identify and to determine the locations of the first three Doppler peaks.

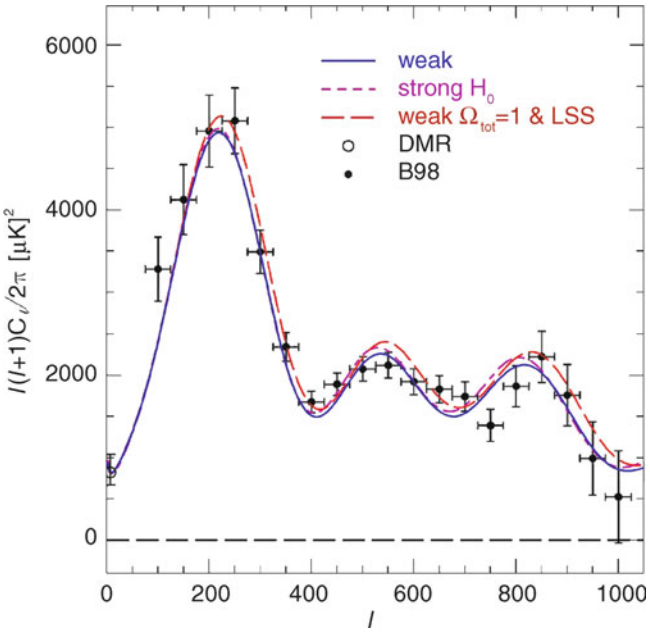


Fig. 8.41 Power spectrum of the CMB angular fluctuations, measured with the BOOMERANG experiment. These results were published in 2001, based on the same data as the previously released results, but using an improved analysis. Plotted are the coefficients $\ell(\ell+1)C_\ell/(2\pi)$ as a function of wave number or the multipole order $\ell \sim 180^\circ/\theta$, respectively. The first three peaks can clearly be distinguished; they originate from oscillations in the photon-baryon fluid at the time of recombination. Curves show the fluctuation spectra of several cosmological models which provide good fits to the CMB data. The model denoted "weak" (*solid curve*) uses the constraints $0.45 \leq h \leq 0.90$, $t_0 > 10\,\mathrm{Gyr}$, and it has $\Omega_\Lambda = 0.51$, $\Omega_\mathrm{m} = 0.51$, $\Omega_\mathrm{b}h^2 = 0.022$, $h = 0.56$, and accordingly $t_0 = 15.2\,\mathrm{Gyr}$. The *short-dashed curve* ("strong H_0") uses a stronger constraint $h = 0.71 \pm 0.08$, and yields $\Omega_\Lambda = 0.62$, $\Omega_\mathrm{m} = 0.40$, $\Omega_\mathrm{b}h^2 = 0.022$, $h = 0.65$, and accordingly $t_0 = 13.7\,\mathrm{Gyr}$. Source: D. Netterfield et al. 2002, *A Measurement by BOOMERANG of Multiple Peaks in the Angular Power Spectrum of the Cosmic Microwave Background*, ApJ 571, 604, p. 611, Fig. 3. ©AAS. Reproduced with permission

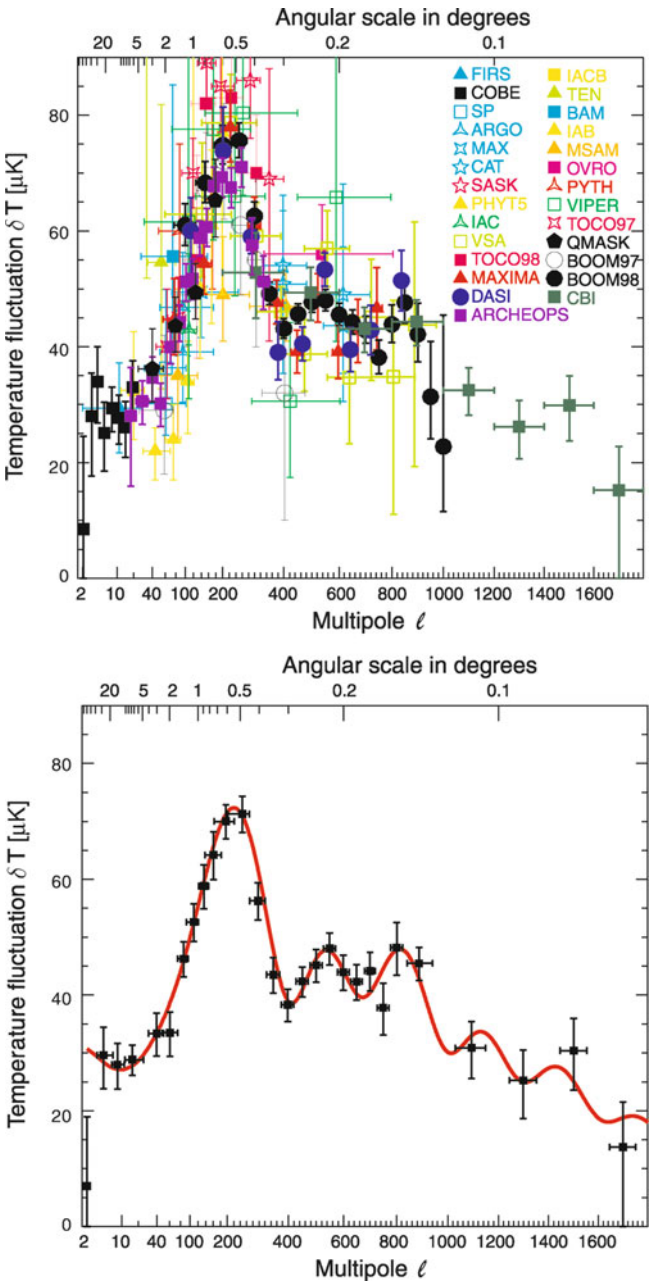


Fig. 8.42 This figure summarizes the status of the CMB anisotropy measurements as of the end of 2002. *On the top*, the results from a large number of individual experiments are shown. *On the bottom*, the 'best' spectrum of the fluctuations is plotted, obtained by a weighted mean of the individual results where the corresponding error bars have been taken into account for the weighting. The *red curve* shows the fluctuation spectrum of the best-fitting cosmological model, with parameters given in the text. Source: X. Wang et al. 2002, *Last stand before WMAP: Cosmological parameters from lensing, CMB, and galaxy clustering*, Phys. Rev. D 68, 123001, Figs. 1, 2

The status of measurements of the CMB anisotropy as of the end of 2002 is shown in Fig. 8.42. In the top panel, the results of numerous experiments are plotted individually. The panel at the bottom shows the weighted mean of these experiments. Although it might not be suspected at first sight, the results of all experiments shown on the top are compatible with each other. With that we mean that the individual measurements, given their error bars, are statistically compatible with the power spectrum that results from the weighted mean.

With the optimally averaged power spectrum, we can now determine the cosmological model which best describes these data. Under the assumption of a flat model, we obtain $\Omega_\Lambda = 0.71 \pm 0.11$ and a baryon density of $\Omega_\mathrm{b}h^2 = 0.023 \pm 0.003$, in excellent agreement with the value obtained from primordial nucleosynthesis [see (4.68)]. Furthermore, the spectral index of the primordial density fluctuations is constrained to $n_\mathrm{s} = 0.99 \pm 0.06$, which is very close to the Harrison–Zeldovich value of 1. In addition, the Hubble

constant is estimated to be $h = 0.71 \pm 0.13$, again in extraordinarily good agreement with the value obtained from local investigations using the distance ladder, which is a completely independent measurement. These agreements are truly impressive if one recalls the assumptions our cosmological model is based upon.

8.6.5 WMAP: Precision measurements of the CMB anisotropy

In June 2001, the Wilkinson Microwave Anisotropy Probe satellite was launched, named in honor of David Wilkinson, one of the pioneers of CMB research. WMAP was, after COBE, only the second experiment to obtain an all-sky map in the microwave regime. Compared to COBE, WMAP observed over a wider frequency range, using five (instead of three) frequencies, with a much improved angular resolution (which is frequency-dependent; about $20'$, compared to $\sim 7°$ for COBE), and in addition, WMAP was able to measure the polarization of the CMB. Results from the first year of observation with WMAP were published in 2003. These accurate results were in full agreement with the expectations from a model which is spatially flat, i.e., $\Omega_\mathrm{m} + \Omega_\Lambda = 1$, dominated by cold dark matter with a baryon fraction as determined from Big Bang nucleosynthesis and a primordial spectral index with $n_\mathrm{s} \approx 1$. These findings therefore justify to call this model *the* standard model of cosmology. Some of the most important results from WMAP will be discussed in the following.

Comparison to COBE. Since WMAP was the first satellite after COBE to map the full sky in the relevant frequency range, its first year results allowed the first verification of the COBE measurements. In Fig. 8.43, sky maps by COBE and by WMAP are displayed. The dramatically improved angular resolution of the WMAP map is obvious. In addition, it can clearly be seen that both maps are very similar if one compares them at a common angular resolution. This comparison can be performed quantitatively by 'blurring' the WMAP map to the COBE resolution using a smoothing algorithm. Since WMAP did not observe at exactly the same frequencies as COBE, it is necessary to interpolate between two frequencies in the WMAP maps to match the frequency of the COBE map. The comparison then shows that, when accounting for the noise, the two maps are completely identical, with the exception of a single location in the Galactic disk. This can be explained, e.g., by a deviation of the spectral behavior of this source from the 2.73 K blackbody spectrum that was implicitly assumed for the aforementioned interpolation between two WMAP frequencies. The confirmation of the COBE measurements is indeed highly impressive and presumably contributed to the Nobel Prize in physics awarded to John C. Mather and George F. Smoot in 2006, the principal investigators of the two leading experiments on COBE.

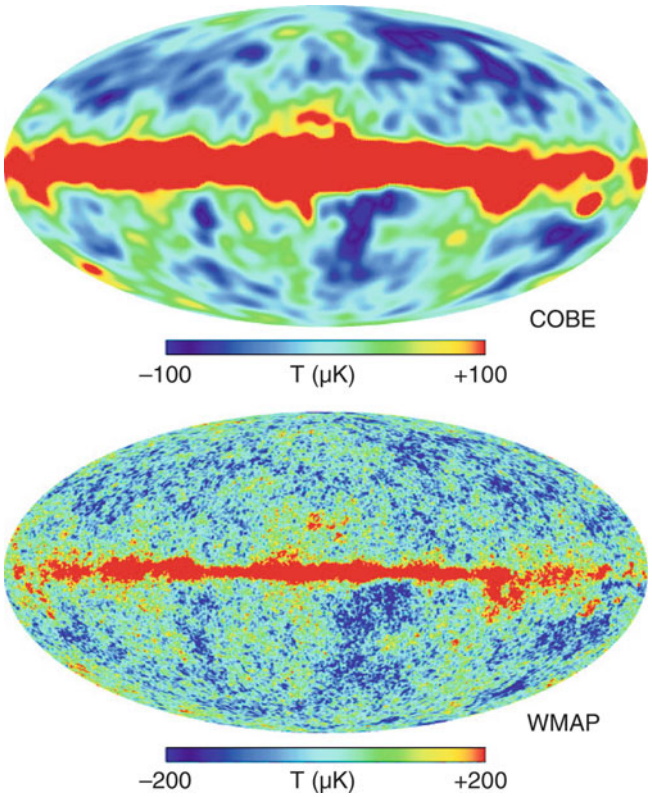


Fig. 8.43 Comparison of the CMB anisotropy measurements by COBE (*top*) and the first-year measurement of WMAP (*bottom*), after subtraction of the dipole originating from the motion of the Sun relative to the CMB rest frame. The enormously improved angular resolution of WMAP is easily seen. Although these maps were recorded at different frequencies, the similarity in the temperature distribution is clearly visible; apart from the different resolution and noise properties, these two maps are essentially indistinguishable. Thus, the COBE results were, for the first time, confirmed independently. Source: C.L. Bennett et al. 2003, *First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Maps and Basic Results*, ApJS 148, 1, p. 15, Fig. 7. ©AAS. Reproduced with permission

Cosmic variance. Before we continue discussing the WMAP results we need to explain the concept of cosmic variance. The angular fluctuation spectrum of CMB anisotropies is quantified by the multipole coefficients C_ℓ. For instance, C_1 describes the strength of the dipole. The dipole has three components; these can be described, for example, by an amplitude and two angles which specify a direction on the sphere. Accordingly, the quadrupole has five independent components, and in general, C_ℓ is defined by $(2\ell + 1)$ independent components.

Cosmological models of the CMB anisotropies predict the *expectation value* of the amplitude of the individual components C_ℓ. In order to compare measurements of the

CMB with these models one needs to understand that we will never measure the expectation value, but instead we measure only the mean value of the components contributing to the C_ℓ on *our* microwave sky. Since the quadrupole has only five independent components, the expected statistical deviation of the average from the expectation value is $C_2/\sqrt{5}$. In general, the statistical deviation of the average of C_ℓ from the expectation value is

$$\Delta C_\ell = \frac{C_\ell}{\sqrt{2\ell+1}} \,. \tag{8.38}$$

In contrast to many other situations, in which the statistical uncertainties can be reduced by analyzing a larger sample, this is not possible in the case of the CMB: there is only one microwave sky that we can observe. Hence, we cannot compile a sample of microwave maps, but instead depend on the one map of our sky. Observers at another location in the Universe will see a different CMB sky, and thus will measure different values C_ℓ, since their CMB sky corresponds to a different realization of the random field which is specified by the power spectrum $P(k)$ of the density fluctuations. This means that (8.38) is a fundamental limit to the statistical accuracy, which cannot be overcome by any improvements in instrumentation. This effect is called *cosmic variance*. The precision of the first-year WMAP measurements is, for all $\ell \lesssim 350$, better than the cosmic variance (8.38). Therefore, the fluctuation spectrum for $\ell \lesssim 350$ measured by WMAP in its first year is 'definite', i.e., further improvements of the accuracy in this angular range will not provide additional cosmological information (however, additional measurements can test for potential systematic effects, such as calibration issues).

The fluctuation spectrum. Since WMAP observed at five different frequencies, the Galactic foreground radiation can, in principle, be separated from the CMB due to the different spectral behavior. Alternatively, external datasets may be utilized for this, as described in Sect. 8.6.4. This second method is preferred because, by using multi-frequency data in the foreground subtraction, the noise properties of the resulting CMB map would get very complicated. The sky regions in which the foreground emission is particularly strong—mainly in the Galactic disk—are disregarded in the determination of C_ℓ. Furthermore, known Galactic and extragalactic point sources are also masked in the map.

The resulting fluctuation spectrum is presented in Fig. 8.44. In this figure, instead of plotting the individual C_ℓ, the fluctuation amplitudes are averaged in ℓ-bins. The solid curve indicates the expected fluctuation spectrum in a ΛCDM-Universe whose parameters are quantitatively discussed further below. The gray region surrounding the model spectrum specifies the width of the cosmic variance, according to (8.38) and modified with respect to the applied binning.

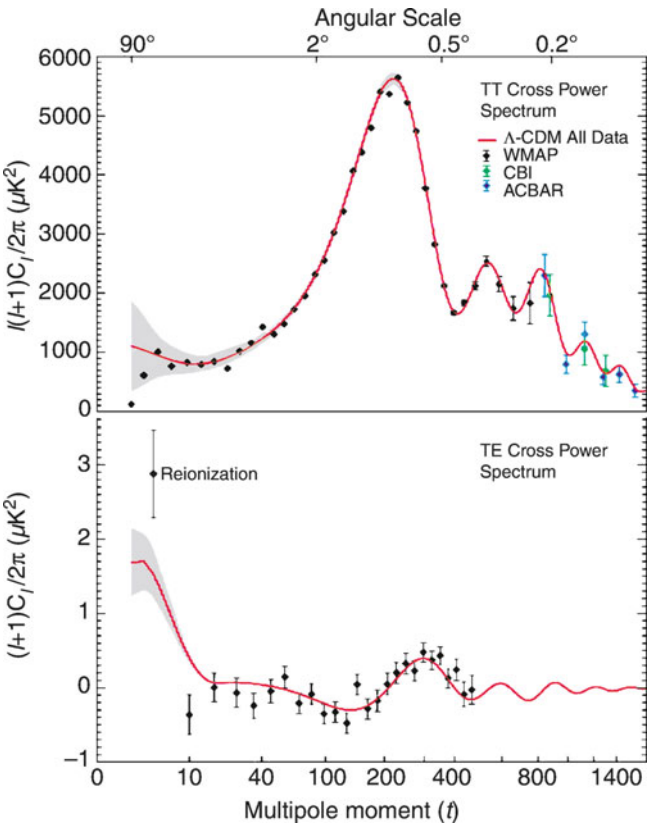


Fig. 8.44 As the central result from the first-year WMAP measurements, the *top panel* shows the fluctuation spectrum of the CMB temperature (TT), whereas the *bottom panel* displays the power spectrum of the correlation between the temperature distribution and polarization amplitude (TE). Besides the data points from WMAP, which are plotted here in ℓ-bins, the results from two other (ground-based) CMB experiments (CBI and ACBAR) are also plotted, at larger ℓ. The curve in each panel shows the best-fitting ΛCDM model, and the *gray region* surrounding it indicates the cosmic variance. The large amplitude of the point in the TE spectrum at small ℓ indicates an unexpectedly high polarization on large angular scales, which suggests an early reionization of the Universe. Source: C.L. Bennett et al. 2003, *First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Maps and Basic Results*, ApJS 148, 1, p. 22, Fig. 12. ©AAS. Reproduced with permission

The first conclusion is that the measured fluctuation spectrum agrees with the model extraordinarily well. Virtually no statistically significant deviations of the data points from the model are found, except for $\ell = 2$ (the quadrupole) and in the region around $\ell \sim 30$, where the fluctuations are somewhat smaller than predicted. However, it must be kept in mind that some deviations are expected to occur as statistical outliers.[17] The agreement of the data with the model is in fact

[17] Recall that a 1σ error bar means that there is a $\sim 68\,\%$ probability that the true value lies within the 1σ regime. Conversely this implies that

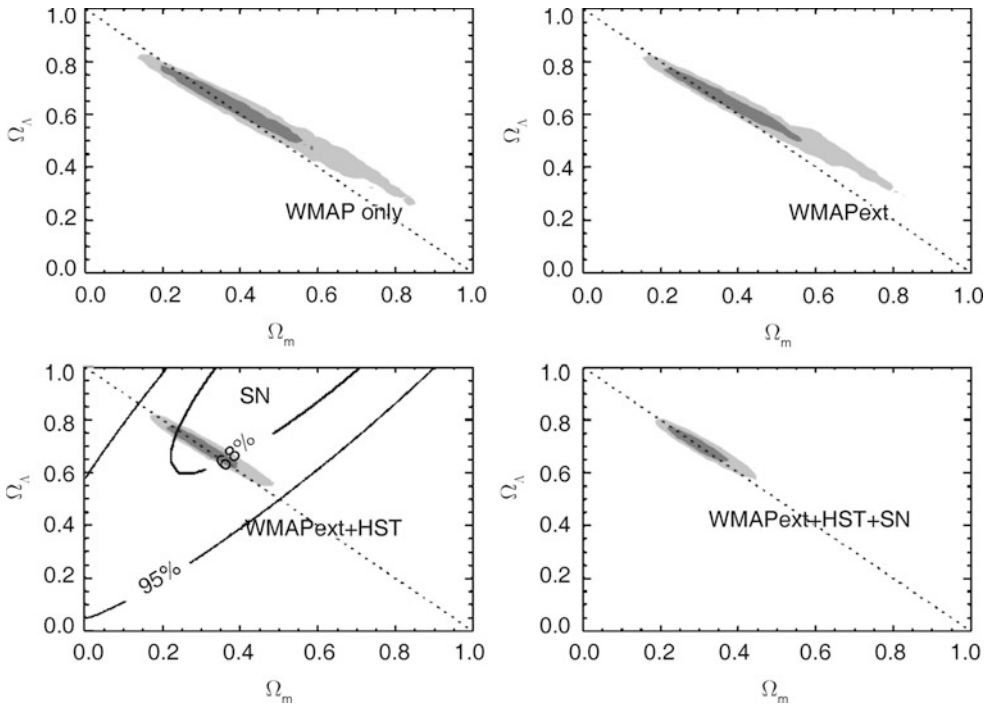


Fig. 8.45 1σ and 2σ confidence regions (dark gray and light gray areas, respectively) in the Ω_m–Ω_Λ plane. In the *upper left panel*, only the WMAP data were used. In the *upper right panel*, the WMAP data were combined with CMB measurements on smaller angular scales (WMAPext). In the *lower left panel*, the WMAPext data were combined with the determination of the Hubble constant from the HST Key Project, and the confidence region which is obtained from SN Ia measurements is included only for comparison. In the *lower right panel*, the SN Ia data are included in addition. The *dashed line* indicates models of vanishing curvature, $\Omega_m + \Omega_\Lambda = 1$. Source: D. Spergel et al. 2003, *First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters*, ApJS 148, 175, p. 189, Fig. 13. ©AAS. Reproduced with permission

spectacular: despite its enormous potential for new discoveries, the first-year data of WMAP 'only' confirmed what had already been concluded from earlier measurements. Hence, the results from WMAP confirmed the cosmological model in an impressive way and, at the same time, considerably improved the accuracy of the parameter values.

The data point which apparently deviates most from the model is that of the quadrupole, $\ell = 2$. In the COBE measurements, the amplitude of the quadrupole was also smaller than expected, as can be seen in Fig. 8.42. If one assigns physical significance to this deviation, this discrepancy may provide the key to possible extensions of the standard model of cosmology. Indeed, shortly after publication of the WMAP results, a number of papers were published in which an explanation for the low quadrupole amplitude was sought. However, the deviation of the measured quadrupole is less than 2σ away from the expectation, i.e., there is about a 5 % probability that such a deviation occurs, for any value of ℓ. Hence, the significance of the anomalously small quadrupole is far too low for drawing any far-reaching conclusions.

almost every third data point will deviate from the underlying model by more than the size of the error bar.

Polarization measurement. In the lower part of Fig. 8.44, the power spectrum of the correlation between the temperature distribution and the polarization is plotted. One finds a surprisingly large value of this cross-power for small ℓ. This measurement is probably the most unexpected discovery in the WMAP data from the first year of observation, because it requires a very early reionization of the Universe, $z_{ion} \sim 15$, hence much earlier than might be expected from, e.g., the spectra of QSOs at $z \gtrsim 6$.

To highlight one of the results from the early WMAP data, Fig. 8.45 displays the allowed regions in the Ω_m–Ω_Λ parameter plane. The CMB data alone show that the Universe is very close to being flat. However, significant deviations from flatness are not ruled out, but in order to deviate substantially from $\Omega_m + \Omega_\Lambda = 1$, the Hubble constant must be much smaller than allowed by direct measurements. Using the constraint on H_0 from the Hubble Key Project, the deviations from flatness are confined to $|\Omega_m + \Omega_\Lambda - 1| \lesssim 0.02$.

Verification of the ISW effect. The detection of the integrated Sachs–Wolfe effect in the fluctuation spectrum by itself is a verification of the value for Ω_Λ being different

from zero, fully independent of the supernovae results. As a matter of fact, the physical origin of this effect can be proven directly because the integrated Sachs–Wolfe effect is produced at relatively low redshifts ($z \lesssim 1$), where the influence of a cosmological constant on the expansion rate $H(z)$ and the growth factor $D_+(z)$ is noticeable, as a result of the time evolution of the gravitational potential. Therefore, it should be directly correlated with the large-scale matter fluctuations which are traced by the distribution of galaxies, AGNs, and clusters of galaxies, assuming a bias model. Indeed, correlations between CMB temperature fluctuations measured by WMAP with luminous red galaxies and QSOs from the SDSS, infrared sources from the 2MASS catalog, and radio sources from the NRAO VLA Sky Survey were found, clearly indicating a rather local origin of some fraction of the CMB anisotropies. In particular, no such correlations were found for source populations at redshifts $\gtrsim 1$, as also expected from the ISW. Hence, the ISW effect in the CMB data is robustly detected.

8.6.6 From WMAP to Planck

WMAP observed the microwave sky for a total of 9 years, and four more data releases after the first-year release of 2003 were made, the last one in December 2012 (termed WMAP-9). Increased photon statistics, and a better understanding of the instrumental properties yielded improved results in these later releases. Furthermore, the fluctuation spectrum of the CMB polarization was measured. In parallel, new balloon and ground-based experiments were conducted; for example, the results from a 2003 flight of the BOOMERANG experiment were published in 2005, measuring the polarization fluctuation spectrum. We mention in particular the Atacama Cosmology Telescope (ACT) and the South Pole Telescope (SPT); these ground-based experiments use larger aperture telescopes and thus have better angular resolution than space missions. Furthermore, these two experiments observed at higher frequencies than WMAP, up to 218 GHz, the frequency at which the thermal Sunyaev–Zeldovich (SZ) effect does have no impact on the temperature measurements (see Fig. 6.35). These higher-resolution observations are particularly useful in constraining the contributions by foreground emission, including point sources, sources of the SZ-effect, and dust radiation from the Galaxy and star-forming galaxies.

The final results from WMAP confirmed the first-year results and significantly tightened the allowed parameter ranges. Any curvature of the Universe is constrained to $|\Omega_\mathrm{m} + \Omega_\Lambda - 1| \lesssim 0.007$. The CDM power spectrum fits the CMB data over the full angular range, and the primordial spectral index n_s deviates slightly, but significantly from the Harrison–Zeldovich value of unity. Together, these two findings confirm essential predictions from an early inflationary phase of the cosmic expansion.

The WMAP data are fully compatible with a six parameter model of a flat ΛCDM model. The six parameters are: physical baryon density $\Omega_\mathrm{b}h^2$, physical cold dark matter density $\Omega_\mathrm{c}h^2$, Ω_Λ, two parameters specifying the primordial density perturbations, i.e., the slope n_s and an amplitude, and finally the optical depth τ for electron scattering due to free electrons after reionization (such that the fraction of photons which get scattered at low redshift is $f_\mathrm{sc} = 1 - \mathrm{e}^{-\tau} \approx \tau$, where in the final step $\tau \ll 1$ was assumed). The permitted volume in this six-dimensional parameter space was reduced by the WMAP CMB measurement by a factor of 68 000 relative to the CMB measurements before WMAP!

Further highlights from WMAP include an upper limit on the sum of the neutrino masses, $\sum m_\nu \lesssim 0.5\,\mathrm{eV}$, and the firm detection of helium in the pre-recombination era. To elaborate on the latter aspect, Big Bang nucleosynthesis predicts a helium abundance by mass of $\approx$25 %, which is in agreement with measurements of the helium abundance in the current Universe, once one accounts for the fact that additional helium is generated by nuclear fusion in stars. The CMB offers the opportunity to constrain the helium abundance before recombination, and directly proves that most of the helium in our current Universe was formed in the early phases, an independent confirmation of Big Bang nucleosynthesis.

The sensitivity to the helium abundance is due to the fact that helium recombines at higher temperatures, and thus the number density of electrons in the epoch preceding hydrogen recombination is lower, compared to the case where all baryons were in form of hydrogen. A lower electron density implies a somewhat larger mean-free path of photons, i.e., the coupling between photons and baryons gets weaker. Correspondingly, the diffusion length scale in increased, which affects the angular scale at which Silk-damping becomes important.

The Planck satellite was launched in May 2009, together with Herschel, and surveyed the microwave sky from August 2009 on. It carried two separate instruments, the low-frequency and high-frequency instruments (LFI and HFI, respectively). The HFI observed the full sky five times, before its supply in ^{3}He run out and the instrument could not be kept cool anymore. The LFI completed eight full sky surveys before being shut down in the fall of 2013. The LFI observed at three frequencies, 30, 44 and 70 GHz, whereas the HFI operated at six frequencies, 100, 143, 217, 353, 545 and 857 GHz; hence, Planck covers a much larger range of frequencies than WMAP. This broad frequency range is particularly useful for observing foreground emission, i.e., synchrotron and free-free emission at low frequencies, and dust emission at high frequencies. The angular resolution for

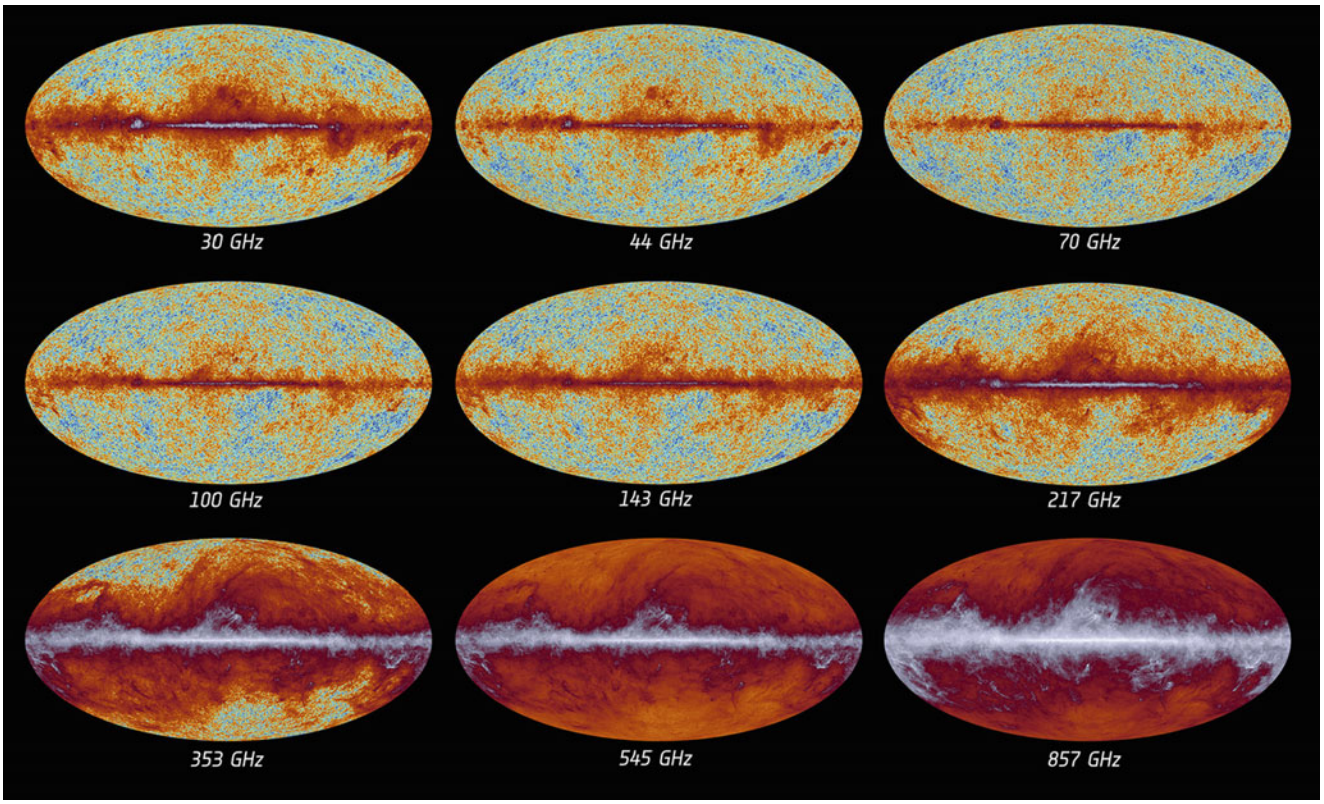


Fig. 8.46 The sky as seen by Planck: All-sky maps in the nine frequency bands of the Planck satellite. Credit: ESA and the Planck Collaboration

the four highest frequency bands is slightly better than 5′, and thus considerably higher than that of WMAP.

First results from the Planck mission were released in January 2011, the Planck's Early Release Compact Source Catalogue. It contained a list of unresolved and compact sources in all nine frequency bands extracted from the first complete Planck all-sky survey. The extragalactic part of these compact source are mainly radio sources, star-forming galaxies, and galaxy clusters. Follow-up studies of some of these sources, in particular galaxy clusters, with telescopes at other wavebands were published in a series of papers in 2012; some of these results were discussed in Sect. 6.4.5. Finally, in March 2013, the data and data products from the first 15.5 months of the Planck mission were released, together with 29 publications describing the spacecraft, the instruments, data processing, catalogs of compact and SZ sources, as well as scientific analyses of the data, including the cosmological results. For example, Planck detected considerably more than 10 000 compact sources in each of the four highest frequency bands (those with the best angular resolution), with more than 2000 at high Galactic latitudes (and thus presumably extragalactic origin), and a positional accuracy of better than 1′. In the same release, a catalog of 1227 galaxy cluster candidates was published, as detected by their SZ-signal; of those, about 850 had been confirmed as clusters.

Planck sky maps and CMB power spectrum. In Fig. 8.46, the all-sky maps from the 2013 data release of Planck at nine different frequencies are displayed. Synchrotron emission from the Galaxy is strong at the lowest frequency, whereas the Galactic dust emission, as well as radiation from the CO molecule, are very prominent at the highest frequencies. As mentioned before, these maps allow the measurements of astrophysical foreground emission, needed for the cosmological analysis of the temperature fluctuations. In addition, it must be stressed that these maps contain a wealth of astrophysical information about the foreground, in particular of our Milky Way: Planck is more than a cosmology mission, in that it provided the first all-sky maps at these high microwave frequencies. The main frequency bands for studying the CMB anisotropies are at 100, 143 and 217 GHz which are the ones where the effects of Galactic emission are minimal. We note again that the latter of these frequencies is insensitive to the thermal SZ-effect.

As a prime result from Planck, the CMB temperature fluctuation spectrum is shown in Fig. 8.47. Planck has located the first seven acoustic peaks of the CMB spectrum; its high angular resolution allowed a measurement of the C_ℓ up to $\ell \sim 2500$. The data shown in Fig. 8.47 are binned in ℓ; for individual C_ℓ, the signal-to-noise is greater than unity for all multipoles with $\ell \lesssim 1700$. In fact, the accuracy with which the Planck power spectrum can constrain models is limited

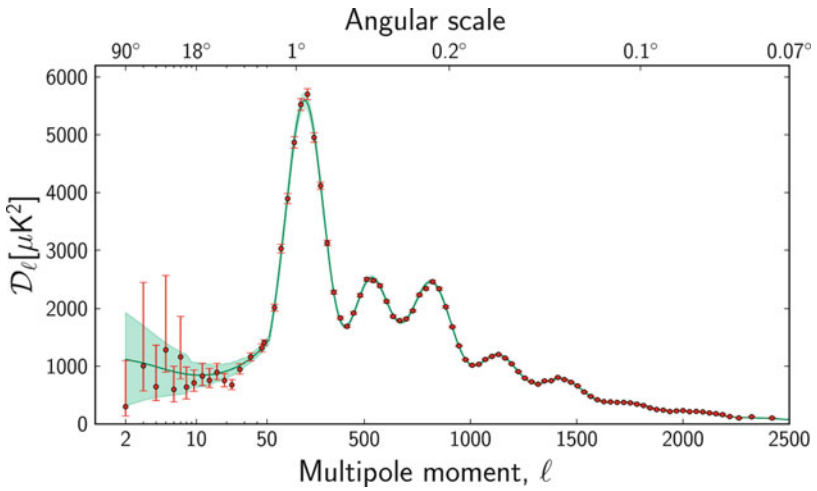


Fig. 8.47 The angular power spectrum (points with error bars) of the CMB temperature fluctuations as measured by the Planck satellite. Plotted is $\mathcal{D}_\ell = \ell(\ell+1)C_\ell/(2\pi)$ as a function of wavenumber (lower axis) and angular scale (upper axis). Note that the ℓ-scale changes from a logarithmic scale for $\ell \leq 50$ to a linear one for larger ℓ. The *solid curve* shows the prediction of a flat ΛCDM model, with the *shaded band* indicating the uncertainty due to cosmic variance, as applies for the ℓ-binning used. Source: Planck Collaboration 2013, *Planck 2013 results. I. Overview of products and scientific results*, arXiv:1303.5062, Fig. 19; original source: ESA and the Planck Collaboration

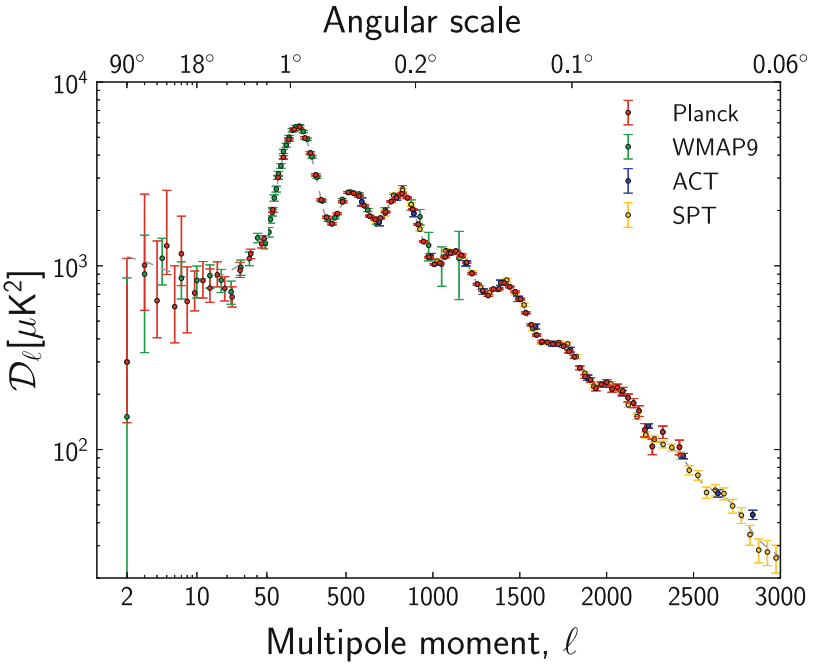


Fig. 8.48 Angular power spectrum of CMB temperature fluctuations, $\mathcal{D}_\ell = \ell(\ell+1)C_\ell/(2\pi)$, as measured by Planck (*red* points), compared to the measurements of WMAP-9 (*green*) and two high-resolution ground-based experiments, ACT and SPT (*blue* and *yellow* points, respectively). Source: Planck Collaboration 2013, *Planck 2013 results. I. Overview of products and scientific results*, arXiv:1303.5062, Fig. 25; original source: ESA and the Planck Collaboration

by cosmic variance for $\ell \lesssim 1500$, and limited by the ability to model extragalactic foreground emission for $\ell \gtrsim 1500$—hence, it is not dominated by instrumental noise at any ℓ.

Comparison with other CMB measurements. The Planck power spectrum is in excellent agreement with that obtained by WMAP, as shown in Fig. 8.48. Furthermore, the Planck power spectrum at high ℓ is smoothly connected to high-resolution measurements from the ground-based experiments ACT and SPT. Taken together, Fig. 8.48 shows that the CMB spectrum is now measured with high accuracy for $\ell \lesssim 2500$. This measured spectrum can be fitted with a simple six-parameter cosmological model, as shown by the dashed curve in Fig. 8.48. We will discuss the values of these parameters in the next section.

On closer investigation, one finds that the relative calibration of the WMAP and Planck spectra are slightly different, concerning the overall normalization. The two spectra agree at a remarkable level for all ℓ in the range $50 \leq \ell \leq 400$ if the WMAP power spectrum is multiplied by a factor of 0.976.

Whereas the origin of this slight difference is not clear at present, it may be related to a small uncertainty in the overall instrumental calibration.

Gravitational lensing of the CMB. As was mentioned in Sect. 8.6.3, gravitational light deflection by inhomogeneities in the Universe is a cause of secondary anisotropies which affect the power spectrum of CMB fluctuations. These lensing effects are included in the models which predict the power spectrum C_ℓ as a function of the cosmological parameters, since these same parameters predict the density fluctuation spectrum of matter in the universe. The prime effect of lensing is that the acoustic peaks and troughs are slightly smoothed out, making them a bit broader and decreasing their amplitude; furthermore, once the primary anisotropies decrease in amplitude due to Silk damping, lensing generates an appreciable power at large ℓ, as can be seen in Fig. 8.38.

However, gravitational lensing does not only affect the power spectrum, but induces small non-Gaussian features in the temperature sky map. Specifically, it generates a (small) non-trivial four-point correlation in the temperature distribution. These lensing effects were seen in cross-correlations of the CMB anisotropies measured by WMAP with tracers of the large-scale structure, such as high-redshift radio galaxies. The ground-based experiments ACT and SPT detected the lensing signal in the CMB temperature maps directly. With Planck, the lensing signal was measured across the whole sky, and a reconstruction of the gravitational lensing potential as a function of sky position was performed. Hence, Planck enabled the study of large-scale structure lensing, in a similar manner as the cosmic shear studies described in Sect. 8.4, except that the 'source population' is not in the form of isolated galaxies with a broad redshift distribution, but the CMB sky at a redshift $z \approx 1100$ fixed by the physics of recombination. The difference in source redshift implies a different redshift sensitivity to the mass distribution: Whereas the cosmic shear signal obtained from faint galaxies at $z_s \sim 1$ is most sensitive to the mass distribution at intermediate redshifts, $0.2 \lesssim z \lesssim 0.7$, lensing of the CMB has its maximum sensitivity for matter at $z \sim 2.5$ [this behavior follows directly from the redshift dependence of the critical surface mass density $\Sigma_{\rm cr}$—see (3.67)].

The corresponding power spectrum of the gravitational lensing potential is shown in Fig. 8.49, where the lensing signal was extracted from sky maps at three different frequencies, shown as colored boxes. An optimal combination of these three different results was constructed and is shown by the solid boxes. The black curve is the lensing power spectrum as predicted by the best-fitting flat ΛCDM model obtained by Planck, which is the one shown in Fig. 8.48. We see that the measured lensing power is in excellent agreement with the predictions based on the cosmological model, which again is an amazing result: in order to obtain this prediction, the same model that describes the density fluctuations at $z \sim 1100$ is used to also describe the properties of the density field in the later universe—the one responsible for lensing. Hence, the agreement seen in Fig. 8.49 is a very powerful consistency check for our understanding of cosmological structure growth.

One can further check the interpretation of the lensing signal by cross-correlating maps of the lensing potential with a distribution of tracers of the cosmic density field. Such a correlation is seen for a number of source populations, namely the clusters from the maxBCG catalog (see Sect. 6.2.4), luminous red galaxies from SDSS, mid-infrared sources from the WISE all-sky catalog, and radio sources from the NRAO VLA Sky Survey. In all cases, a highly significant correlation is seen, and its angular dependence follows the expectations which are based on the redshift distributions of these source populations and their assumed bias factors. Moreover, at the highest frequencies the extragalactic part of the Planck map is dominated by star-forming galaxies with typical redshifts $2 \lesssim z \lesssim 4$ (see Sect. 9.3.3). Hence, this source population is best matched to the redshift dependence of the lensing efficiency. It was found that there is a strong correlation between the temperature fluctuations in the two highest-frequency maps of Planck and the deflection angle predicted from the lensing potential. Thus, the intensity distribution at these high frequencies essentially measures the source populations of dusty galaxies which in turn are good tracers of the large-scale matter distribution responsible for the lensing effect.

Polarization from Planck. The cosmological analysis of the 2013 Planck data release does not include polarization measurements. Obtaining reliable, well-calibrated results for the polarization signal is much more difficult than for the temperature anisotropies, mainly for two reasons: First, the polarization signal is more than an order of magnitude smaller than the temperature signal, and thus more difficult to measure. Second, the calibration of the instruments for polarization measurements is more difficult, since there is no standard calibrator source on the sky for polarization available. For these reasons, no Planck polarization data were included in this first cosmological analysis.

However, Planck has measured the polarization signal with very high confidence over a broad range of angular scales, as can be seen from Fig. 8.50. These polarization measurements are in very good agreement with the predictions of the polarization, based on the ΛCDM model that yields the best fit to the measured temperature fluctuations. Thus, once the instrumental calibration of Planck is better understood, these polarization data will yield further constraints on the cosmological model.

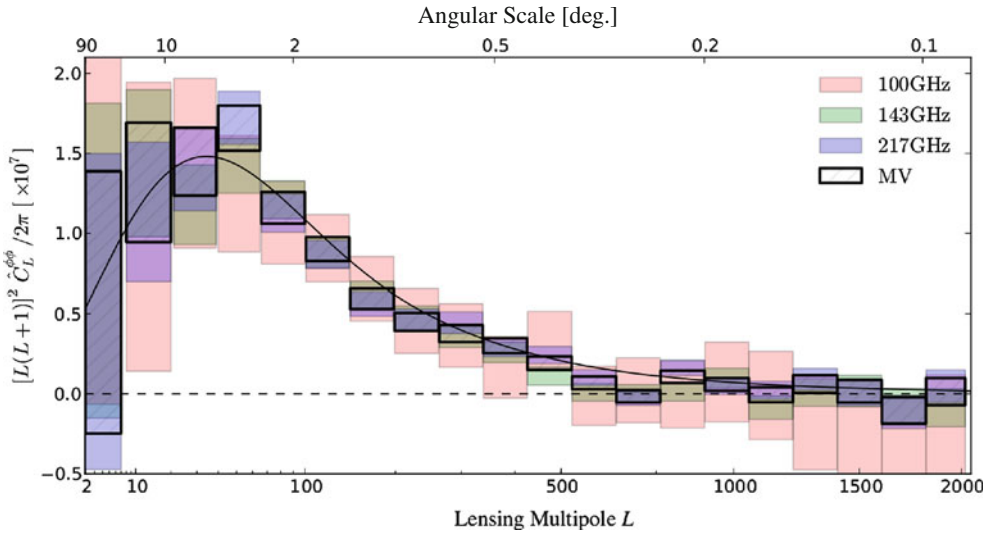


Fig. 8.49 The angular power spectrum of the lensing potential as obtained from non-Gaussian features in the Planck CMB temperature map, as a function of multipole (lower axis) and angular scale (upper axis). Results from three frequency channels are shown separately, as well as an optimal combination of these individual spectra (*solid boxes*). Source: Planck Collaboration 2013, *Planck 2013 results. XVII. Gravitational lensing by large-scale structure*, arXiv:1303.5077, Fig. 10; original source: ESA and the Planck Collaboration

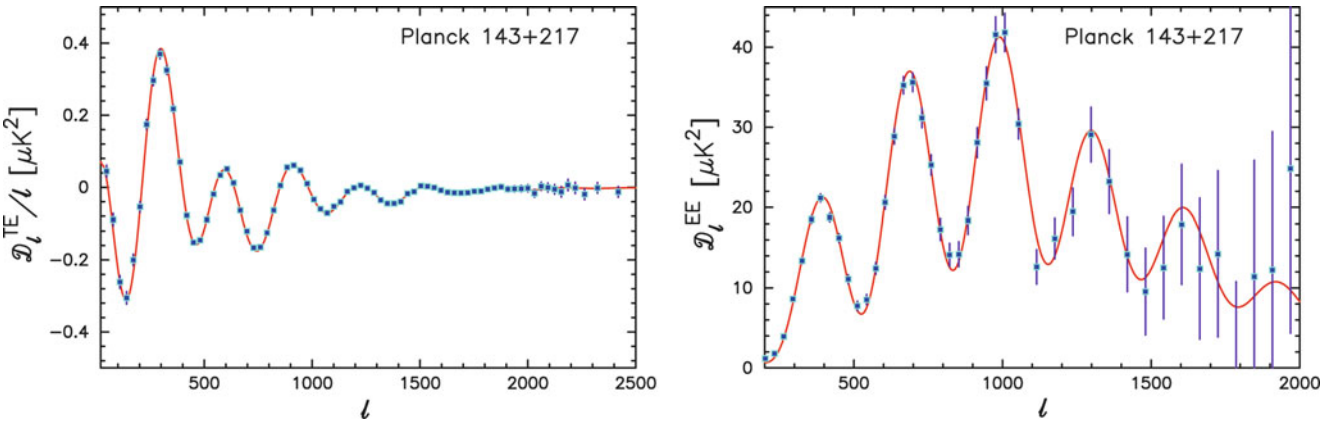


Fig. 8.50 The Planck temperature-polarization cross-power spectrum (*left*) and the polarization power spectrum (*right*), compared to the best fitting ΛCDM model (*red curve*) whose parameters were determined without using polarization data. Source: Planck Collaboration 2013, *Planck 2013 results. XVI. Cosmological parameters*, arXiv:1303.5076, Fig. 11; original source: ESA and the Planck Collaboration

8.7 Cosmological parameters

For a long time, the determination of the cosmological parameters has been one of the prime challenges in cosmology, and numerous different methods were developed and applied to determine H_0, Ω_m, Ω_Λ, etc. Until about the turn of the Millennium, these different methods yielded results with relatively large error margins, some of which did not even overlap. In recent years, the situation has fundamentally changed, as already discussed in the previous sections. The measurements by WMAP and Planck represent the current state-of-the-art in the determination of the cosmological parameters, and thus we begin this section with a presentation of the results from CMB measurements.

8.7.1 The standard cosmological model from CMB measurements

Before discussing the parameters of the cosmological standard model as obtained from the CMB measurements in more detail, we want to mention two (near) degeneracies in parameter space, and how they can be broken.

Geometric degeneracy. The first degeneracy comes about by considering the relation between the physical conditions in the Universe at recombination and the observed microwave sky. In order to translate physical scales at the last-scattering surface to observables, one needs the comoving angular diameter distance $f_k(z_\mathrm{rec})$, which depends on the density parameters Ω_m and Ω_Λ. On the other hand, the

density fluctuation spectrum at recombination depends on the epoch of matter-radiation equality, which in turn depends on $\Omega_m h^2$ [see (4.30)]. In particular, the comoving scale of the sound horizon r_s depends on a_{eq}, as seen by (7.46), and thus on $\Omega_m h^2$. Therefore, the observed temperature fluctuations are essentially preserved if the sound horizon r_s and the distance $f_k(z_{rec})$ to the last-scattering surface are both multiplied by the same factor, to keep the angular size of the sound horizon,

$$\theta_1 = r_s / f_k(z_{rec}) \tag{8.39}$$

unchanged.

The location of the acoustic scale (8.39) is the most precise measurement from CMB temperature fluctuations. With the localization of seven peaks in the power spectrum, Planck determined that scale with an accuracy of $\sim 0.1\,\%$. The aforementioned degeneracy between changes in the sound horizon at recombination and the distance to the last-scattering surface is illustrated in Fig. 8.51, where the allowed combinations of parameters are plotted as colored points in the Ω_m–Ω_Λ-plane (left panel) and the H_0–curvature plane (right panel) and where color encodes the corresponding value of the Hubble constant (left) and Ω_Λ (right). To obtain these constraints, the Planck temperature power spectrum (hereafter denoted as 'PL') was combined with ground-based measurements of the power spectrum from ACT and SPT (as shown in Fig. 8.48; these high-ℓ measurements will be denoted as 'highL' henceforth) and the polarization power spectrum for $\ell \leq 23$ as obtained from WMAP (denoted as 'WP'); we will discuss below which role the polarization results for large angular scales play for the analysis. As is easily seen, deviations away from flat models (indicated by the dashed lines) are permitted from this data set, and these deviations are strongly correlated with the value of the Hubble constant.

In fact, one can compare the left panel of Fig. 8.51 with the upper two panels in Fig. 8.45, where the same degeneracy is visible. In the lower panels of Fig. 8.45, this degeneracy is broken by using the determination of H_0 as obtained by independent measurements (here, from the Hubble Key Project). This rules out very small values of H_0 which are required for strong deviations from flatness. With the measurement of the lensing potential in the Planck data, this degeneracy can be broken without referring to external measurements of H_0, because the lensing effect probes the geometry of the universe and the amplitude of the density fluctuations in an independent way. As shown by the solid contours in Fig. 8.51, the allowed region in parameter space dramatically shrinks when this lensing information is added to the analysis.

Amplitude degeneracy. As one of the most significant effect of secondary anisotropies, we discussed in Sect. 8.6.3 the scattering of CMB photons in the epoch after the reionization of the Universe. As can be seen from Fig. 8.38, the prime effect of this scattering is to reduce the overall observed amplitude of the temperature fluctuations. This implies the presence of a degeneracy between the intrinsic density fluctuations at recombination, and the scattering optical depth τ due to free electrons after reionization. Also here, information from the lensing potential can partly break the degeneracy. However, the best signature of the scattering optical depth is provided by polarization information at large angular scales, as mentioned before. For that reasons, most of the analyses of the 2013 Planck results included the low-ℓ polarization information from the WMAP experiment.

The base model. According to the model of inflation, our Universe is expected to be spatially flat. Motivated by this theoretical prediction, and encouraged by the result shown in Fig. 8.51, one may first see whether the CMB data are compatible with a flat cosmological model.

Such a model is described by a minimum of six parameters. Two of them characterize the initial density fluctuations, namely an amplitude A and the power-law slope n_s (see Sect. 7.4.1). As a third parameter, the post-reionization scattering optical depth τ needs to be chosen. The remaining three parameters describe the energy contents and scale of the universe. One could choose them to be Ω_b, Ω_m and H_0 (recalling that the restriction to flatness then fixes $\Omega_\Lambda = 1 - \Omega_m$). But one is free to use another set of combinations of these parameters. For this choice, one should recall that the parameter uncertainties obtained from such an analysis are usually strongly correlated; an example of this is seen in Fig. 8.51, where the Hubble constant and the matter density parameter are strongly coupled. Hence, one might use parameter combinations where these correlations are expected to be reasonable small. For the analysis described below, the combinations $\omega_b \equiv \Omega_b h^2$, $\omega_c \equiv \Omega_c h^2$ (where $\Omega_c = \Omega_m - \Omega_b$ is the density parameter of cold dark matter), and Ω_Λ are typically used.

The parameters of the base model. In Fig. 8.52 we show the constraints on the model parameters as determined by PL + lensing (colored points and black curves), PL + WP (red contours and curves), and by the WMAP 9-year data (grey contours and curves). The probability distribution of each parameter is shown in the panels on the diagonal. We first see that the results for τ are the same for WMAP-9 and PL + WP, since the most important information about the scattering optical depth comes from the CMB polarization at large angular scales, which is used in both data sets. With PL + lensing, i.e., without polarization information, the distribution of τ is considerably broader. The probability distributions for the other four parameters are narrower for the Planck data set than for WMAP-9, and they are slightly

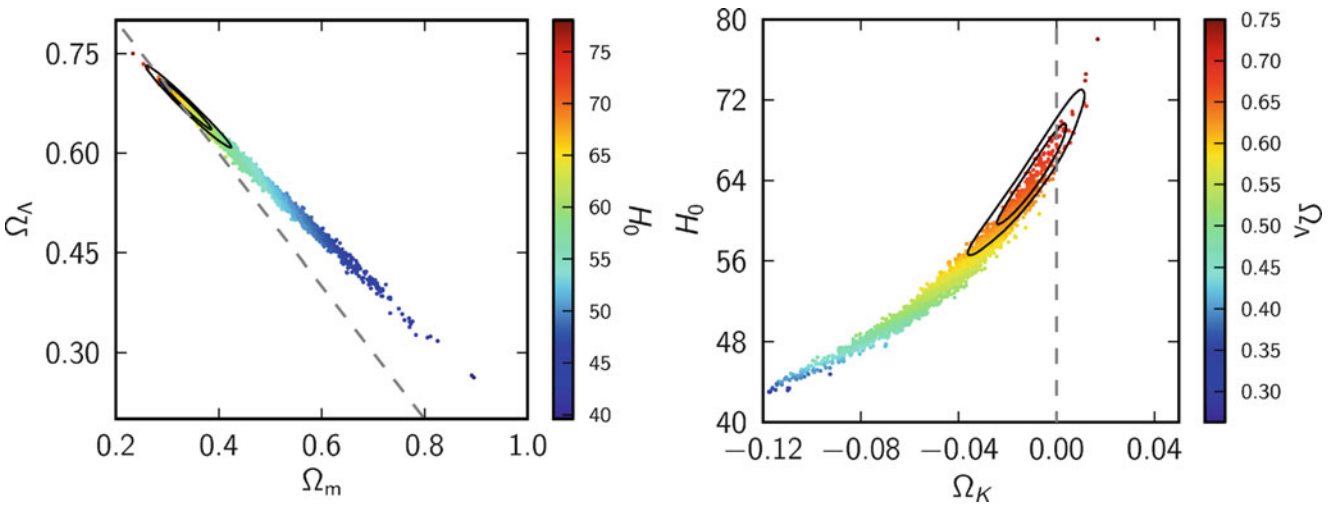


Fig. 8.51 Illustration of the geometric degeneracies in cosmological parameters from CMB temperature fluctuations. *Left panel*: The confidence region in the Ω_m–Ω_Λ-plane as obtained by combining the temperature power spectrum obtained by Planck and, for high ℓ, the ACT and SPT experiments, plus the polarization power spectrum for $\ell \leq 23$ as measured by WMAP, shown as colored points, where the color indicates the Hubble constant (color bar at the right). The *black solid* contours show the confidence region after the lensing results from Planck are included. *Right panel*: The same in the H_0–curvature plane, where $\Omega_K = 1 - \Omega_m - \Omega_\Lambda$, and where the points are color-coded by Ω_Λ. Source: Planck Collaboration 2013, *Planck 2013 results. XVII. Gravitational lensing by large-scale structure*, arXiv:1303.5077, Fig. 15; original source: ESA and the Planck Collaboration

shifted. However, in no case is the shift significant compared to the width of the WMAP-9 probability distribution. This can also be seen from the distributions in the two-dimensional parameter planes, where there is significant overlap in the confidence contours from Planck and WMAP. We thus conclude that the results from WMAP and Planck are fully compatible with each other, not only at the level of the C_ℓ (Fig. 8.48), but also concerning the cosmological analysis.

We also see a clear correlation between the parameter estimates, as mentioned before. In particular, the value of H_0 is seen to depend strongly on the values of the other parameters in this base model. An explicit expression for this degeneracy can be obtained be recalling that θ_1 is the most accurately measured quantity from Planck. Since this scale depends only on the geometry of the universe [through $f_k(z)$] and the total and baryonic matter density in the early universe (through the sound horizon), θ_1 is essentially independent of A, n_s and τ. Translating the measured value of θ_1 into a parameter combination, one finds

$$\Omega_m h^{3.2} \left(\Omega_b h^2\right)^{-0.54} = 0.695 \pm 0.002 \,, \qquad (8.40)$$

hence this combination is determined with an accuracy of better than 0.3 %. The combination of parameters in (8.40) explains many of the parameter correlations shown in Fig. 8.52.

In Table 8.1, the values of the cosmological parameters are listed, as determined by the Planck data in combination with other CMB measurements. For each parameter, the 68 % confidence interval is indicated, together with the central value (note that, if the probability distribution for a parameter is not symmetric around its maximum, the location of the maximum probability can be slightly different from the central value, but these shifts are much smaller than the uncertainty range). We note that many of the parameters are determined with a relative accuracy of one or a few percent. Compared to the best-fitting WMAP cosmology, the matter density is slightly larger (and Ω_Λ corresponding slightly smaller), the normalization of the power spectrum as expressed by σ_8 is somewhat larger, and the Hubble constant is somewhat smaller (with a corresponding slight increase in the age of the Universe). As mentioned before, all these changes are within the 1-σ range of the WMAP measurements. The deviation of n_s from unity is measured with a high significance of $\sim 6\sigma$, confirming one of the robust predictions of inflationary models.

The quality of the best-fitting model can be seen in Fig. 8.47, where its corresponding angular power spectrum (using PL + WP + highL) is shown as solid curve. It provides an excellent fit to the data, with the exception for low values of ℓ, in particular around $\ell \sim 30$, where the measured power is smaller than expected from the best-fit model. This discrepancy was noted already by WMAP and is therefore unlikely to originate from instrumental effects. We shall return to this issue below. However, the fit it excellent for all $\ell \geq 50$. Furthermore, the same model fits the lensing data extremely well, seen in Fig. 8.49, which shows in particular the consistency of our understanding of structure formation in the universe through gravitational collapse.

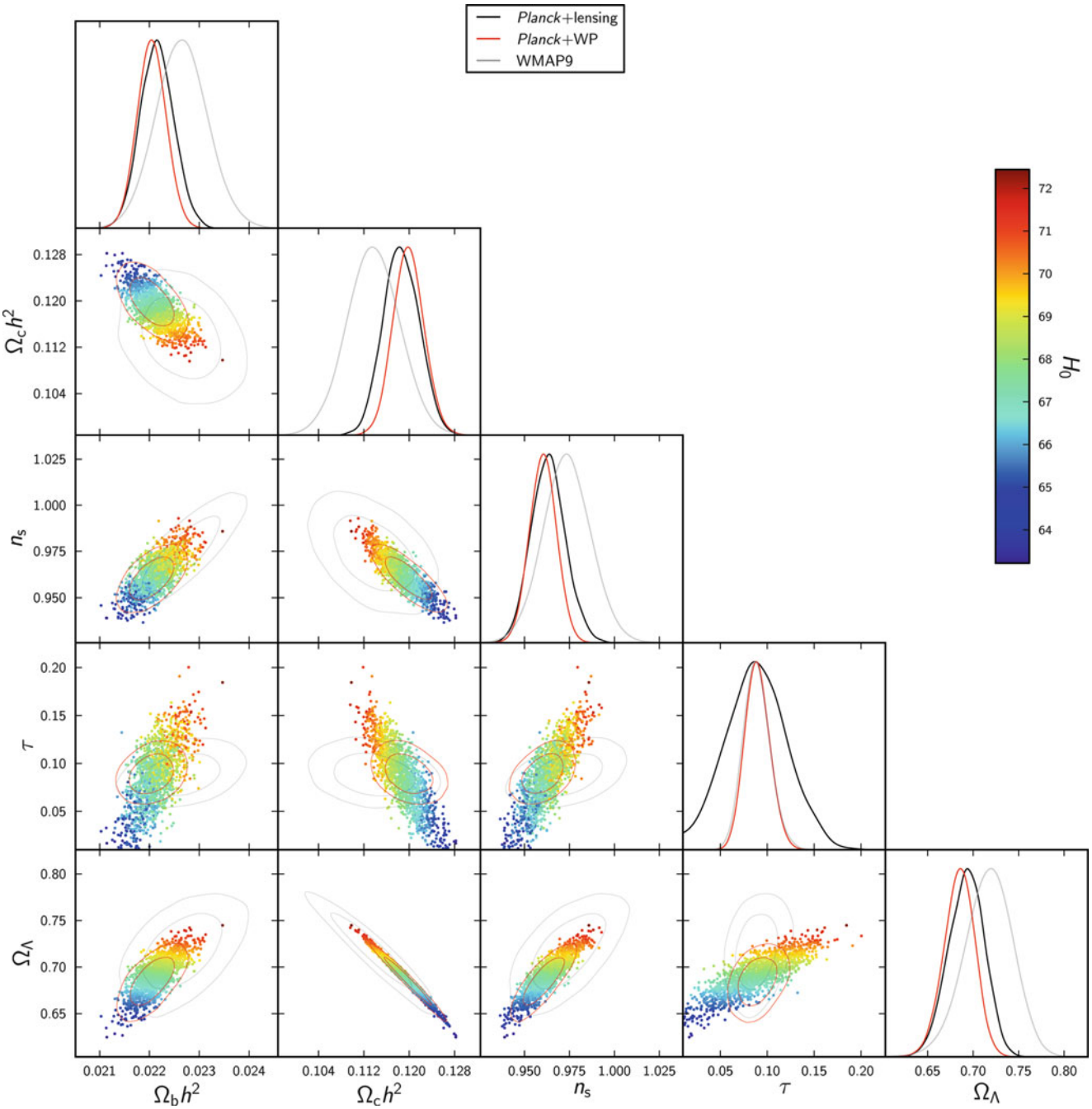


Fig. 8.52 Constraints on the parameters of the cosmological standard model, as obtained (1) by combining Planck temperature anisotropies (PL) with the measurement of the lensing potential, shown by the colored points, where color indicates the corresponding value of H_0, and the *black curves*, (2) by combining PL with WP (*red* contours), and (3) by the WMAP 9-year results (*grey* contours). The diagonal shows the estimated probability for each of the five parameters, whereas the off-diagonal panels show confidence regions in two-parameter planes. Source: Planck Collaboration 2013, *Planck 2013 results. XVI. Cosmological parameters*, arXiv:1303.5076, Fig. 2; original source: ESA and the Planck Collaboration

8.7.2 Consistency and discrepancies with other measurements

Cosmological parameter estimates from the CMB are considered to be more robust than those from most other methods, because predicting the CMB properties requires only well-understood physical processes which, due to the small density fluctuations in the early universe, need to be considered only in the linear regime. The underlying physics of most other methods is far more complicated; for example,

Table 8.1 The cosmological parameters as determined by the Planck 2013 data, in combination with low-ℓ polarization measurements from WMAP (WP), high-ℓ temperature measurements from ACT and SPT (highL), the lensing potential determined from Planck, and results from baryonic acoustic oscillations

Parameter	PL + WP	PL + WP + highL	PL + lensing + WP + highL	PL + WP + highL + BAO
$\Omega_b h^2$	0.02205 ± 0.00028	0.02207 ± 0.00027	0.02218 ± 0.00026	0.02214 ± 0.00024
$\Omega_c h^2$	0.1199 ± 0.0027	0.1198 ± 0.0026	0.1186 ± 0.0022	0.1187 ± 0.0017
Ω_Λ	0.685 ± 0.018	0.685 ± 0.017	0.693 ± 0.013	0.692 ± 0.010
τ	0.089 ± 0.014	0.091 ± 0.014	0.090 ± 0.014	0.092 ± 0.013
n_s	0.9603 ± 0.0073	0.9585 ± 0.0070	0.9614 ± 0.0063	0.9608 ± 0.0054
σ_8	0.829 ± 0.012	0.828 ± 0.012	0.8233 ± 0.0097	0.826 ± 0.012
z_{ion}	11.1 ± 1.1	11.1 ± 1.1	11.1 ± 1.1	11.3 ± 1.1
H_0(km/Mpc/s)	67.3 ± 1.2	67.3 ± 1.2	67.9 ± 1.0	67.80 ± 0.77
Age(Gyr)	13.817 ± 0.048	13.813 ± 0.047	13.794 ± 0.044	13.798 ± 0.037

the use of SNe Ia as standardized candles is essentially purely based on empirical studies, not on a detailed physical understanding of these explosions, and thus may be affected by some yet undiscovered systematic effects.

Therefore, it is of great interest to compare the results from the CMB measurements with those of other cosmological probes. Compatibility strengthens our confidence in the standard model. Possible discrepancies may either mean that one of the methods is burdened by unaccounted systematic effects, or that our standard model needs modifications or even is incorrect. Here we will confront the results from the CMB with other cosmological probes.

Baryonic acoustic oscillations. We discussed the results from BAO studies of the galaxy distribution in Sect. 8.1.4, see in particular Fig. 8.8. The results are perfectly compatible with the cosmological model parameters as determined by WMAP, and they are equally well in agreement with those shown in Table 8.1. Since the physical effects needed to predict the BAO signal are well understood—with the possible caveat that the galaxy biasing even on the acoustic scale may not be fully scale-independent—this perfect agreement is very reassuring. In the final column of Table 8.1, the constraints from BAO measurements were included in the cosmological analysis.

Furthermore, one can also compare the observed power spectrum of galaxies with the predictions from the standard model, using the best-fit parameters. The agreement again is excellent on large scales; on smaller scales, slight discrepancies occur, which most likely are due to non-linear effects in structure formation, and thus also in the clustering of galaxies.

Measurements of the Hubble constant. The Planck analysis yields a small value for the Hubble constant, $H_0 \approx (68 \pm 1)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. Whereas this estimate is in full agreement with the original result from the Hubble Key Project (Sect. 3.9.6), it is considerably smaller than most other recent estimates of H_0. Local measurements of H_0, discussed in Sect. 3.9, give values close to $H_0 \approx (74 \pm 3)\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, giving rise to a discrepancy at the $\sim 2\sigma$ level. These measurements make use of the distance ladder. The dominant source of uncertainty in these local measurements is the first rung in the distance ladder, for example, the distance to the LMC, or the distance to the megamaser galaxy NGC 4258. Indeed, the exact value of H_0 depends on which of these first rungs are chosen, since they calibrate the absolute distance scale used in applying secondary distance indicators.

Independent of the distance ladder, H_0 has been determined from gravitational lens systems, making use of the time delay (see Sect. 3.11.4). Two recent measurements with small quoted uncertainties are $H_0 = 70.6 \pm 3.1\,\mathrm{km\,s^{-1}Mpc^{-1}}$ for B1608 + 656, and $H_0 = 78.7^{+4.3}_{-4.5}\,\mathrm{km\,s^{-1}Mpc^{-1}}$ for RXJ1131–1231. Whereas the first of these values is fully compatible with the CMB determination, the second is significantly off. However, as pointed out in Sect. 3.11.4, the determination of H_0 from lensing can be significantly affected by the mass-sheet degeneracy (see Problem 3.5). The corresponding uncertainties appear not to have been fully included in the allowed range of H_0-values quoted above.

Type Ia supernovae. When the CMB results are compared to the results from SNe Ia, the outcome gives a mixed message: Whereas the distance-redshift relation from the best-fitting CMB model is fully compatible with that obtained from the sample of SNe shown in Fig. 8.24 (the Union 2 compilation), the comparison to a different sample of SNe Ia (the SNLS sample) shows a potential discrepancy. In fact, the SNLS sample was investigated using three different methods to account for the light-curve stretching, with one of them yielding a result compatible with the CMB parameters, the other two being somewhat deviant, in the sense that they prefer a slightly, but significantly smaller value of Ω_m.

Cosmic shear. Lensing be the large-scale structure, as measured from correlated shape distortions of distant galaxies,

was discussed in Sect. 8.4. As mentioned there, the strongest constraints from current cosmic shear surveys concern a combination of the power-spectrum normalization σ_8 and the matter density parameter Ω_m. The combination of these two parameters given in (8.26) is significantly smaller than the prediction from the best-fitting CMB model. A different cosmic shear analysis of the CFHTLenS data set, making explicit use of the multi-band color information (which provide an indication for the source redshifts) found a combination of $\sigma_8(\Omega_\mathrm{m}/0.27)^{0.46} = 0.774 \pm 0.04$, which is smaller than the corresponding value from the CMB, $\sigma_8(\Omega_\mathrm{m}/0.27)^{0.46} = 0.89 \pm 0.03$, by about 2σ. It is currently unclear what the source of this discrepancy is; however, given the challenges that need to be overcome to obtain reliable shape measurements of faint and small galaxies, it is not unlikely that future studies will reveal the origin of this difference.

Abundance of massive galaxy clusters. The CMB results for Ω_m and σ_8 can be compared with the values obtained from studying the abundance of galaxy clusters (Sect. 8.2.1). The combination of these two parameters, as obtained from X-ray selected clusters and given in (8.18), is significantly smaller than that obtained from the CMB. The corresponding result (8.20) from optically-selected clusters presents a better match to the CMB values.

Results on cluster cosmology were derived from the Planck data itself, by studying a complete sub-sample of their SZ-detected galaxy clusters. In order to derive cosmological constraints, the SZ-signal Y (see Sect. 6.4.4) needs to be related to a mass estimate. This has been done by X-ray follow-up studies of about a third of the SZ sample, to calibrate the Y-mass scaling relation. Accounting for a discrepancy between hydrostatic masses M_{500} obtained from X-ray analysis, and the expectation from numerical simulations of clusters, which seem to indicate that there is a relative bias in the mass estimates of $\sim 20\,\%$, a similar discrepancy as for X-ray selected clusters is found. The origin of this discrepancy is yet not clear, but not unlikely to be related to cluster mass calibrations and/or proper modeling of selection effects.

8.7.3 Extensions of the standard model

Whereas the standard model, defined by the six basic parameters mentioned earlier, provide an excellent fit to the CMB data—temperature fluctuation power spectrum out to $\ell = 2500$, low-ℓ polarization fluctuations, and the CMB lensing effects, it is worth to generalize this model by considering extensions in various ways. In particular, one may check whether some of the discrepancies with other data sets can be relaxed if such modifications of the standard model are accounted for. In the analysis of the Planck 2013 data set, several such extensions were considered; as we shall see, they yielded highly interesting results.

Curvature. The standard model assumes spatial flatness. We have seen before (Fig. 8.51) that the measurement of the lensing potential with the CMB data breaks the geometric degeneracy and gives strong constraints on potential deviations from flatness. Without including the lensing effect, the CMB data by itself yield $(\Omega_\mathrm{m} + \Omega_\Lambda - 1) = 4.2^{+4.8}_{-4.3}\%$, whereas if the lensing information is added, this allowed interval shrinks to $1.0^{+1.9}_{-1.8}\%$. If, in addition, the results from the BAO studies are included in the analysis, the limits for a possible deviation from spatial flatness become even more stringent,

$$(\Omega_\mathrm{m} + \Omega_\Lambda - 1) = 0.10^{+0.65}_{-0.62}\,\% \,. \tag{8.41}$$

Hence, there is very little room for generalizing the model away from flatness. The very stringent bounds on a possible curvature thus confirm a robust prediction from inflationary models.

Number of neutrino families and their masses. The standard model assumed that there are three families of neutrinos, and that they are essentially massless. One can relax one of these assumptions, or both together, to see whether the CMB data show a preference for a non-standard picture of the neutrino sector. If the sum of the neutrino masses is significantly larger than the minimum mass required from neutrino oscillations, then they would contribute to the current matter density in the Universe in the form of hot dark matter, and thus affect the shape of the power spectrum. If the number of neutrino families is larger than three, there would be a larger radiation component in the early universe, changing the expansion law and the epoch of matter-radiation equality. All of this affects the CMB fluctuations.

Significant constraints on the neutrino masses and the number of neutrino families can thus be obtained from the CMB results. As shown in Fig. 8.53, the CMB alone yields a 2σ upper limit on the neutrino mass of $\sum m_\nu < 0.60\,\mathrm{eV}$ and limits the effective number of neutrinos—see (4.63)—to $N_\mathrm{eff} = 3.29^{+0.67}_{-0.64}$. These constraints get even tighter when the results from the BAOs are included in the analysis, yielding $\sum m_\nu < 0.28\,\mathrm{eV}$ and $N_\mathrm{eff} = 3.32^{+0.54}_{-0.52}$. The former results presents the strongest bound on the neutrino masses yet available, it is a mere factor of ~ 4 larger than the lower bound derived from neutrino oscillations. The latter result confirms our picture of particle physics, according to which there are three families of leptons, and thus three kinds of neutrinos.

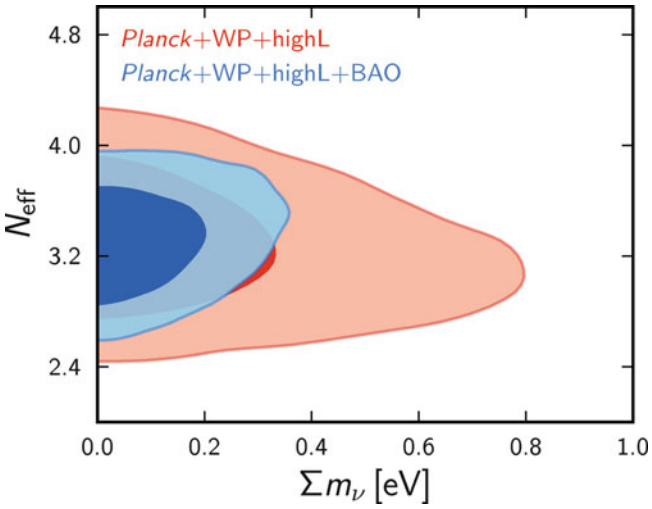


Fig. 8.53 Allowed region in the parameter plane spanned by the sum of neutrino masses and the effective number of neutrino families $N_{\rm eff}$ [see (4.63)]. Constraints are given from CMB data alone (*red*) and in combination with the BAO results (*blue*). Source: Planck Collaboration 2013, *Planck 2013 results. XVI. Cosmological parameters*, arXiv:1303.5076, Fig. 28; original source: ESA and the Planck Collaboration

Big Bang nucleosynthesis. The standard model assumes that the helium abundance following the first few minutes after the Big Bang is given by the theory of nucleosynthesis, and is therefore a function of $\omega_{\rm b} \equiv \Omega_{\rm b} h^2$. This helium abundance then determines the evolution of the number density of free electrons before recombination, since helium recombines earlier than hydrogen because of its higher ionization potential. The value of $\omega_{\rm b}$, as determined from CMB anisotropies, is in excellent agreement with the results obtained from the observed abundance of helium and deuterium in the Universe (see the left panel of Fig. 8.54).

One can now generalize the model by no longer requiring the helium abundance to be fixed by BBN, but leaving is as a free parameter. This parameter can be estimated from the CMB data themselves, as mentioned before: the helium abundance before hydrogen recombination has a significant impact on the CMB spectrum. The right panel of Fig. 8.54 shows the combined constraints on the helium abundance and $\omega_{\rm b}$, the former being in excellent agreement with the value obtained from BBN. This measurement therefore yields independent confirmation of the theory of Big Bang nucleosynthesis and very strongly rules out any exotic model that wants to explain the current helium contents in the Universe solely by nuclear fusion in stars.

Density fluctuations from inflation. The simplest models of inflation predict a power-law primordial density fluctuation spectrum. However, more complex inflationary models allow for a (slight) curvature of the primordial spectrum, called 'running spectral index'. If this curvature is included as a free parameter in the model, one can check whether a 'running' is preferred by the data.

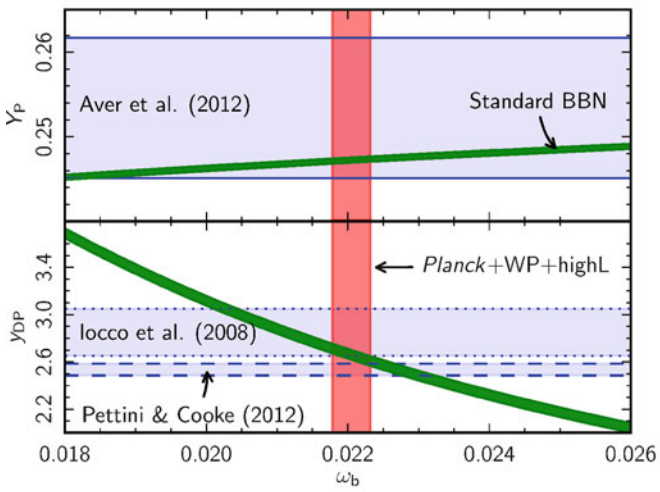
Marginal evidence for a running spectral index was reported from early releases of the WMAP data, and confirmed by WMAP-9. Also the SPT experiment concluded evidence for a significant deviation from a simple power law, when combined with the WMAP data, whereas the ACT data give no support for running.

With the Planck data added, evidence for a curvature in the primordial spectrum is seen at about the 1.5σ level. The sign of the curvature is such that there is relatively less power on the largest scales, compared to a pure power law. The reason for this finding can be traced back to the fluctuation spectrum seen in Fig. 8.47, where it is seen that the measured power for small ℓ lies somewhat below the prediction of the best-fitting model, whereas the model is perfectly compatible with the data for all $\ell \geq 50$. Given that the primary fluctuations for $\ell \lesssim 100$ come from regions on the last scattering surface that had never been in causal contact before that epoch, any feature imprinted there must come from inflation, and not from physical processes in structure evolution.

Inflation also predicts the generation of large-scale gravitational waves, which are perturbations of the geometry of spacetime. These primordial gravitational waves are called tensor fluctuations. Different models of inflation predict different amplitudes of the tensor fluctuations, relative to those of the density fluctuations; therefore, if one can determine the amplitude of these gravitational waves, different models of inflation could be ruled out. Tensor fluctuations generate weak features in the polarization pattern of the CMB, but they will be very difficult to detect, unless the ratio of tensor-to-density fluctuations is unexpectedly large. More indirect signs of tensor fluctuations are imprinted on the higher acoustic peaks, which allows one to derive upper bounds on the tensor-to-density ratio.

Varying physical 'constants'? If we recall the basic assumptions of our cosmological model, namely the cosmological principle which postulates that on large scales, the Universe is homogeneous and isotropic, and that the physical laws as we have determined them here and now are valid at all times and everywhere, it is a question of fundamental importance whether one can see variations in the physical laws. In particular, one may investigate whether the physical constants in our physical laws are indeed constant or change with cosmic epoch. Among the quantities that cosmologists have studied are the electron-to-proton mass ratio, the value of the gravitational constant, and the fine-structure constant $\alpha = e^2/\hbar c$, where e is the elementary charge, and $\hbar$ the reduced Planck constant.

There are reports in the literature about a variation of α with redshift. These results are based on quasar absorption-



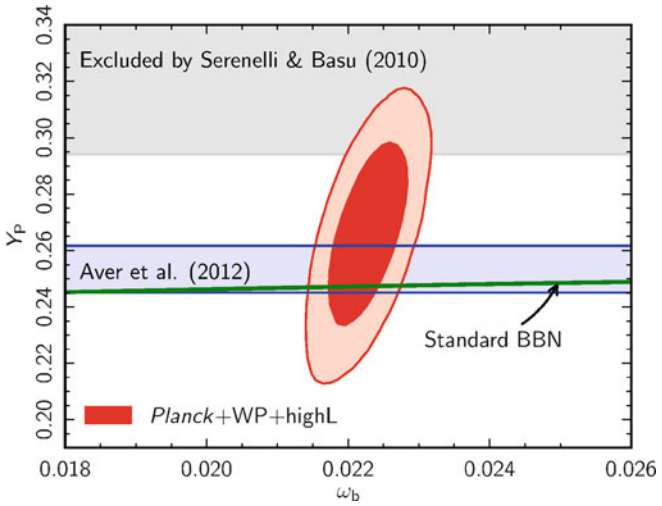


Fig. 8.54 *Left panel*: The primordial abundance Y_P of ^{4}He and 10^5 times the abundance of deuterium y_{DP}, as a function of $\omega_b \equiv \Omega_b h^2$. The *blue horizontal bands* show recent determinations of the primordial abundances of these two elements. The *green bands* display the dependence of the element abundances on ω_b, using the standard BBN model. The *orange band* is the allowed interval of ω_b from the CMB results. *Right panel*: In generalizing the standard model, the helium abundance is left unconstrained by BBN, but treated as a free parameter of the model. The CMB fluctuations yield stringent constraints on the combination of ω_b and the helium abundance, shown by the *red ellipse*. For comparison, the same primordial helium abundance as in the *left panel* is shown, as well as a very conservative upper bound obtained from the helium abundance of the Sun. Source: Planck Collaboration 2013, *Planck 2013 results. XVI. Cosmological parameters*, arXiv:1303.5076, Figs. 29, 30; original source: ESA and the Planck Collaboration

line studies, with the value of α determining the ratios of wavelengths in fine-structure lines. However, these claims have not been confirmed by independent observations from other groups and at other telescopes.

The CMB offers a handle on the variation of the fine-structure constant between the epoch of recombination and today. A change of α modifies the redshift of recombination, as it determines the energy levels in an atom and it affects the cross section for Compton scattering. As a consequence, the positions of the acoustic peaks are shifted, and the Silk damping is affected. Hence, the CMB fluctuations offer a window to constraining α at $z \sim 1100$. As a result, the constraint

$$\frac{\alpha(z_{\rm rec})}{\alpha(0)} = 0.9989 \pm 0.0037 \qquad (8.42)$$

was obtained. Hence, any variation of α with redshift must be minute, and this result provides very strong support for the view that physical constants are indeed constant in time.

Discussion. The standard model turned out to be robust against possible extensions. None of the extensions mentioned here are preferred, except the running of the spectral index, driven by the relatively low temperature fluctuations at low ℓ. It seems that the spectrum for small ℓ is the only peculiarity in the CMB power spectrum, and it has been seen by several instruments independently.

None of the extensions of the standard model removes the tensions between the CMB data and some other data sets described in Sect. 8.7.2. We must point out that these apparent discrepancies are significant only because the error bars of the parameter estimates decreased substantially in recent years, due to the enormous progress in observational cosmology. It remains to be seen whether future results lead to even higher significances of these differences, or whether the impact of systematics were slightly underestimated. The final data release of Planck will not only contain more accurate data, due to the longer observing time, but most likely will include a precise measurement of the CMB polarization properties. Since the information from polarization is partly independent of the temperature fluctuations, one may expect significant changes of the results—either related to the size of the confidence regions, or in terms of best-fitting parameters.

Apart from the low amplitude of some low-ℓ fluctuations, other peculiarities in the CMB data on large scales have been pointed out. The low amplitude quadrupole, the fact that the directions of the $\ell = 2$ and the $\ell = 3$ fluctuations are very much aligned, and the occurrence of a 'cold spot' in the CMB temperature map with properties unexpected from a Gaussian temperature field have all gained considerable attention in the literature. The question was investigated of what is the probability that the CMB sky, assuming the standard model, has such a low quadrupole, such a strong alignment of the $\ell = 2$ and $\ell = 3$ modes, and such a peculiar cold spot. Well, the probability is exceedingly small! However, this is not necessarily a reason to worry, because one asks these question *after* having seen the data. This is called *a posteriori* statistics and it is always dangerous to apply. The human

Table 8.2 Summary of the cosmological parameters as determined by the Planck 2013 data, in combination with low-ℓ polarization measurements from WMAP (WP)

Ω_Λ	0.685 ± 0.018
Ω_m	0.315 ± 0.018
$\Omega_\mathrm{b}h^2$	0.02205 ± 0.00028
H_0(km s^{-1} Mpc^{-1})	67.3 ± 1.2
z_ion	11.1 ± 1.1
τ	0.089 ± 0.014
σ_8	0.829 ± 0.012
Y_P	0.24770 ± 0.00012
t_0(Gry)	13.817 ± 0.048
z_rec	1090 ± 0.54
z_eq	3391 ± 60

brain is enormously efficient in finding peculiarities. For example, if one considers randomly distributed points on a plane, then a short view will allow us to identify peculiarities in the point distribution—a large region in which no point is found, with the shape of a heart, an overdensity of points with the shape of an 'S', etc. One can then calculate the probability for a random distribution of points to show such a heart-shaped void or an S-shaped overdensity, and these probabilities are small—but there are many, many more peculiarities the point distribution could have, but doesn't (like, instead of heart-shaped, the void could have the shape of a banana, a bottle, an eight, ... the overdensity could be A-shaped, B-shaped ...). For that reason, defining criteria after looking at the data and then calculating their probability yields misleading conclusions.[18]

8.7.4 Cosmic harmony

The detailed studies of a large variety of cosmological probes have converged to a standard model of the Universe, which is characterized by a flat ΛCDM model, in which baryons contribute about 4.5 % of the cosmic energy density, cold dark matter makes some 25 %, and the rest, about 70 %, is made of dark energy with properties very similar to that of a cosmological constant; the best current values of the cosmological parameters are summarized in Table 8.2. The structure in the current Universe has evolved from tiny primordial density fluctuations whose spectrum is described by a power law, presumably an outcome of an early inflationary phase of cosmic expansion that occurred some 13.8 Gyr ago. Structure growth occurred through gravitational instability, and in combination with baryonic processes, leads to the structure of the Universe as we observe it today, with its galaxies, AGNs, galaxy clusters, and their large-scale distribution.

While keeping in mind that the values of some cosmological parameters as determined by different probes are apparently deviant at the $\sim 2\sigma$ level, we note again that these tensions are due to the very small statistical error bars enabled by modern cosmological surveys. It must be pointed out that these discrepancies—though formally statistically significant—are small, at a level of 10 % or less (say, in σ_8 or H_0). Legitimate or not, if we assume for a moment that all error bars are underestimated by a factor ~ 1.5 due to as yet undetected systematic effects, these tensions would disappear. Therefore, we will ignore them for now.

Instead, we want to present here a broader picture of the current status of cosmology, focusing on the internal consistency. The density fluctuations in the Universe are observed with a broad range of methods, starting from the largest scales at $z \sim 1100$ seen in the CMB anisotropies, the clustering of galaxies in the more local Universe, the density fluctuations giving rise to the weak gravitational lensing effect measurable by cosmic shear and by the lensing of the CMB, the abundance of galaxy clusters with depends sensitively on the density fluctuations on scales of a few Mpc, and at the smallest scales, fluctuations in the Lyman-α forest. Some of them are purely gravitational probes (like the cosmic shear), others depend on electrodynamics and relativistic hydrodynamics just 400 000 years after the Big Bang or on the relation between galaxies and the underlying mass distribution. Accounting for redshift evolution of the amplitude of density fluctuations, all these probes can be put onto a common scale, as is shown in Fig. 8.55. It is seen that the corresponding results smoothly join together, despite the fact that they come from vastly different probes, techniques, and redshifts. Moreover, these fluctuations follow very closely that of a CDM spectrum.

In addition, we are in a situation where the basic cosmological parameters are not only known with an accuracy that had been unimaginable only a few years ago, but also each of these individual values was measured by more than one independent method, confirming the self-consistency of the model in an impressive manner.

- **Hubble constant.** H_0 was determined with the Hubble Key Project, by means of the distance ladder, particularly using Cepheids, and later refinements and extensions of

[18] You have probably made an experience similar to that: Being far from home, traveling in a different part of the world, e.g., sitting in a cafe, when you suddenly see a person you know—say the former mathematics teacher of your brother. How unlikely that is! Indeed, meeting this person now in this town, in this cafe is extremely unlikely. But you could have met her 2 h ago in the park, or 3 h later during dinner, or the day before, or the day after, or a week before in a different town, or last year on a different trip, and it doesn't have to be the math teacher of your brother, but maybe his biology teacher, the history teacher, or one of your former teachers, or nor teacher at all, but someone else from your school, someone living your neighborhood, or someone in your sports club ... there are incredibly many possibilities for such an incredibly unlikely event so that one of them *will* occur one of these days.

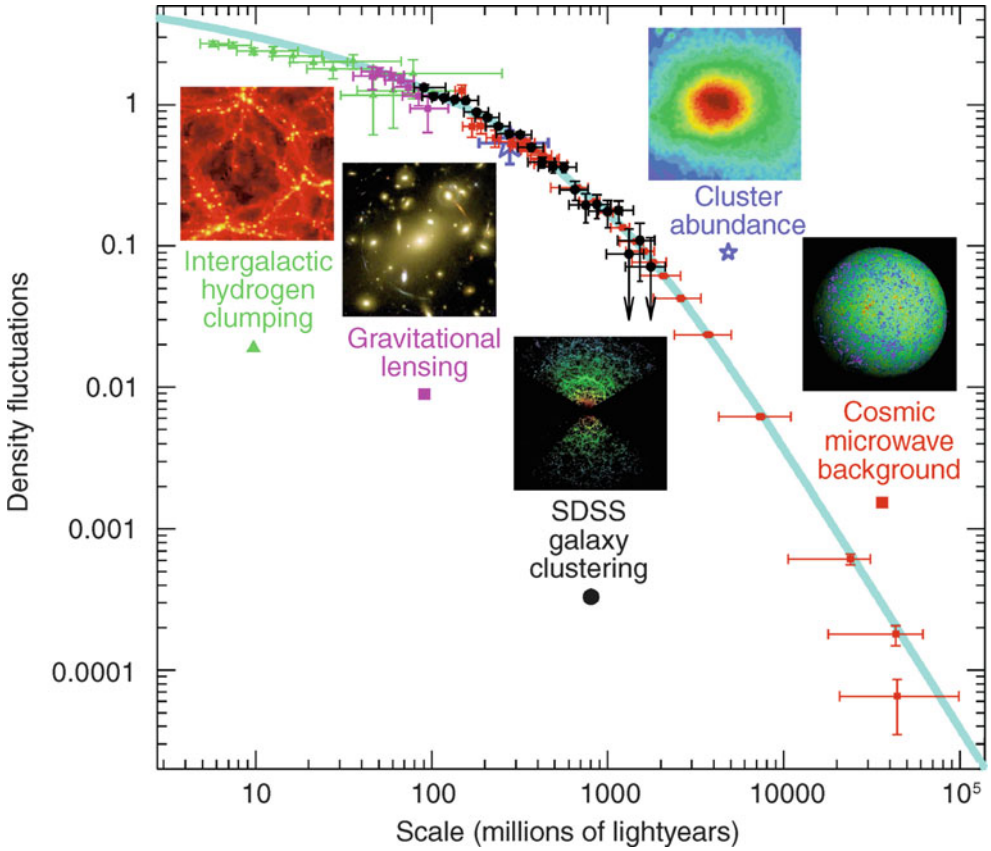


Fig. 8.55 The power spectrum of density fluctuations in the Universe, as determined by different methods. Here, $\Delta^2(k) \propto k^3 P(k)$ is plotted. Note that small length-scales (or large k, respectively) are towards the left in the plot. Going from large to small scales, the results presented here are obtained from CMB temperature fluctuations, from the abundance of galaxy clusters, from the large-scale distribution of galaxies, from cosmic shear, and from the statistical properties of the Lyα forest. One can see that the power spectrum of a ΛCDM model is able to describe all these data over many orders of magnitude in scale. Credit: Max Tegmark

method. These local measurements directly probe the current expansion rate of the Universe. The Hubble constant is also determined with time delay gravitational lenses, which yields an estimate of the expansion rate, averaged between the redshift of the source and redshift zero. Thus, the value of H_0 from lensing makes use of the functional behavior of the expansion rate on redshift, i.e., the Friedmann equation. Third, the Hubble constant is determined from the CMB measurements, this time by comparing the physical density parameters ω_b and ω_c that determine the fluctuations on the last-scattering surface, with the redshift-distance relation which depends on Ω_m and Ω_Λ. Thus, to translate these measurements into a value of H_0, quite a number of properties of the early phase of the Universe are employed. All of these methods yield concordant results, in that the Hubble constant is within 5 % of 71 km s^{-1}Mpc^{-1}, or $h \approx 1/\sqrt{2}$. This concordance provides a powerful test of the standard model.

- **Contribution of baryons to the total matter density.** The ratio $\Omega_\mathrm{b}/\Omega_\mathrm{m}$ is determined from the baryon fraction in clusters of galaxies (Fig. 8.22), from redshift surveys of galaxies (Fig. 8.6), and from the CMB fluctuations, all yielding $\Omega_\mathrm{b}/\Omega_\mathrm{m} \approx 0.15$.
- **Baryon density.** The value for $\Omega_\mathrm{b}h^2$ determined from primordial nucleosynthesis combined with measurements of the deuterium abundance in Ly-α systems was impressively confirmed by the CMB results. Moreover, the confirmation of the BBN value of the helium abundance by the measurements of CMB fluctuations provides a stringent test of our picture of the early thermal history of the Universe.
- **Matter density.** Assuming the value of H_0 to be known, Ω_m is determined from the distribution of galaxies in redshift surveys (8.5), from the CMB, and from the evolution of the number density of galaxy clusters [see (8.19)]. Furthermore, Ω_m is also constrained by combining the value of Ω_b as determined from BBN with the ratio $\Omega_\mathrm{b}/\Omega_\mathrm{m}$ measured from galaxy clusters [see (8.23)].
- **Curvature.** The location of the peaks of the CMB power spectrum yields very tight constraints on any deviation from flatness, if combined either with estimates of the Hubble constant, or with the measurements of baryonic acoustic oscillations in the galaxy distribution.
- **Vacuum energy.** The very tight limits on the curvature of the Universe obtained from the CMB measurements, and the implied tight limits on the deviation of $\Omega_\mathrm{m} + \Omega_\Lambda$ from unity, allows us to determine Ω_Λ from the measurement of Ω_m and the integrated Sachs–Wolfe effect. These values are in excellent agreement with the SN Ia measurements, as shown in Fig. 8.56, as well as with BAOs (Fig. 8.8).
- **Normalization of the power spectrum.** The CMB fluctuations probe the power spectrum of matter inhomo-

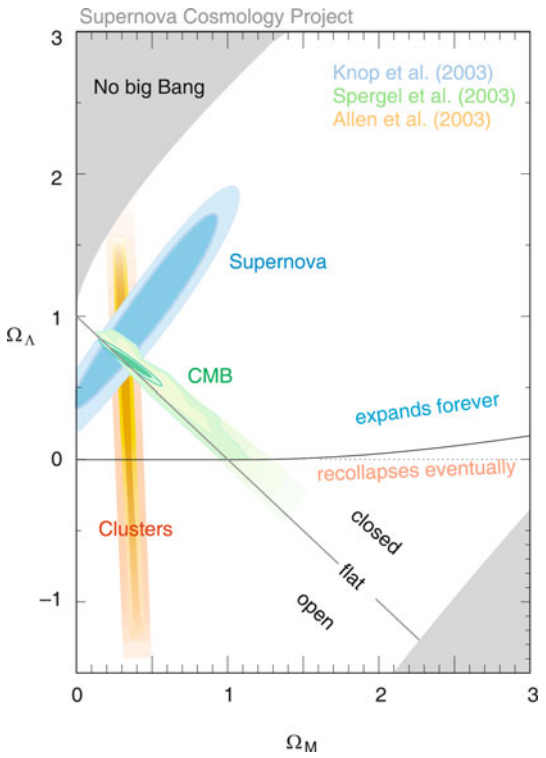


Fig. 8.56 This figure shows the allowed regions of the parameter pair Ω_m and Ω_Λ, as derived from the CMB anisotropy, SN Ia measurements, and the z-evolution of the abundance of galaxy clusters. Since the individual confidence regions have substantially different orientations in this parameter plane, their combination provides much better constraints on these parameters than each method by itself. The smallness of the individual confidence regions and the fact that they are overlapping is an impressive demonstration of the self-consistency of our cosmological model. Credit: Supernova Cosmology Project, adapted from: R.A. Knop et al. 2003, *New Constraints on Ω_m, Ω_Λ, and w from an Independent Set of 11 High-Redshift Supernovae Observed with the Hubble Space Telescope*, ApJ 598, 102

geneities at very large length-scales, and thus an estimate of σ_8 from CMB measurements involves the shape of the matter power spectrum, which in turn depends on the CDM transfer function, the shape parameter Γ, and the primordial slope n_s, to relate these large-scale fluctuations to those on a scale of $8h^{-1}$ Mpc. In contrast to that, cluster abundance and cosmic shear probe much more directly these scales. Despite these very different methods, the results remarkably agree and yield $\sigma_8 \approx 0.8$, with an uncertainty of probably less than 5 %.

- **Age of the Universe.** The age of the Universe derived from the CMB data, $t_0 \approx 13.8 \times 10^9$ yr, is compatible with the age of globular clusters and of the oldest white dwarfs in our Galaxy, as well as with the age of stellar populations in high-redshift elliptical galaxies (see Fig. 6.67).

The observational results described in this chapter opened an era of precision cosmology. On the one hand, the accuracy of the individual cosmological parameters will most likely be improved in the coming years by new observational results; on the other hand, the interest of cosmology will increasingly shift towards observations of the early Universe. Studies of the evolution of cosmic structure, of the formation of galaxies and clusters, and of the history of the reionization of the Universe will increasingly become the focus of cosmological research.

The search for the constituents of dark matter will keep physicists busy in the coming years. Experiments at particle accelerators (e.g., the LHC at CERN) and the direct search, in underground laboratories, for particles which may be candidates for dark matter, are promising. In any case, dark matter (if it indeed consists of elementary particles) will open up a new field in particle physics. For these reasons, the interests of cosmology and particle physics are increasingly converging—in particular since the Universe is the largest and cheapest laboratory for particle physics.

Another central objective of future cosmological research will remain the investigation of dark energy. For the foreseeable future, it will be accessible only through astronomical observations. Due to the enormous importance of a non-vanishing dark energy density for fundamental physics, studying its properties will be at the center of interest of more than just astrophysicists. It is expected that a successful theory describing dark energy will necessitate a significant breakthrough in our general understanding of fundamental physics, as will be discussed next

8.8 Dark energy: Cosmological constant, or something else?

As mentioned several times before, the origin of the accelerated expansion of the Universe is arguably the least understood aspect of fundamental physics. 'Gravity sucks', and does not push. The introduction of a cosmological constant by Einstein for (as we learned after Hubble's discovery of the expanding Universe) wrong reasons has served as an 'explanation' for this cosmic acceleration.

The cosmological constant and Einstein's field equation. Let us schematically write Einstein's field equation of General Relativity in a symbolic form,

$$G = T \,, \tag{8.43}$$

where G (called the Einstein tensor) describes the curvature of spacetime and thus the effects of gravity, whereas T (the so-called energy-momentum tensor) contains information about the matter and energy density. Hence, (8.43) generalizes the Poisson equation in Newtonian gravity. As it stands, (8.43) does not allow a static cosmos, and so Einstein modified his equation to include the cosmological constant,

which then reads

$$G - \Lambda = T \ . \tag{8.44}$$

With this modification, one can construct a static model; however, this model is unstable against small perturbations and, moreover, irrelevant since it is not at all similar to the Universe we live in.

A modern interpretation of the cosmological constant is obtained by slightly rewriting (8.44) as

$$G = T + \Lambda \ , \tag{8.45}$$

where now the Λ-term is seen as a contribution to the source of the gravitational field; it has the same structure as one would get from a constant, uniform vacuum energy density. The difference in the interpretation of (8.44) and (8.45) is then, that (8.44) is a modification of the laws of gravity, whereas (8.45) adds a new energy component as a source of gravity.

Of course, one cannot possibly distinguish between these two equations: if the dark energy has indeed the properties of the Λ-term, it is only a matter of interpretation whether gravity or the energy content is modified. It is almost a semantic issue. But the difference is that the cosmological constant in (8.44) would be another constant of nature, characterizing the law of gravity, whereas in the case of (8.45), the vacuum energy density Λ may in principle be calculated from the properties of the quantum fields corresponding to the elementary particles.

Why so small, and why now? We mentioned before that simple estimates of the value of Λ from quantum field theory differ by some 120 orders of magnitude from the value observed in cosmology—indeed the worst estimate in physics. There is as yet no plausible explanation for the smallness of Λ—why is it so incredibly small, and yet non-zero? The other issue related to Λ is the 'why now' question: why does the transition between matter domination and Λ-domination of the expansion rate happens at a scale factor a within a factor of 2 of the current epoch (see also Fig. 4.13)? If one selects a random epoch on a logarithmic axis of a, the most likely situation is one where the universe is either fully radiation dominated, or fully matter dominated, or where the cosmological constant is the only relevant term. We happen to exist at one of the rather rare moments in the cosmic evolution where two components have almost the same density. A pure coincidence?

Time-varying dark energy. There are ideas that the value of the vacuum energy density is actually not a constant in time, but that it may vary. In that case, the dark energy density would be a function of cosmic epoch, $\rho_{\rm DE}(a)$. For example, there exist models in which the dark energy component somehow interacts with dark matter, and that the two densities trace each other to some degree. This, however, cannot be done arbitrarily; the cosmological constraints we have, derived from observations at low, intermediate, and high redshifts, give strong constraints on the behavior of $\rho_{\rm DE}(a)$ in the past. For example, the SN Ia observations and BAOs tell us that the Universe was decelerating at redshifts larger than about unity, hence at these epochs, dark energy cannot have been the dominant component.

Nevertheless, the question of whether dark energy is compatible with a cosmological constant, or has more complicated properties, is a very interesting one: if it were not constant in time, than it must have a more dynamical origin, which would clearly argue against it being another fundamental constant of nature. One therefore considers as a possible variant of the cosmological constant an equation-of-state of dark energy of the form

$$P_{\rm DE} = w \rho_{\rm DE} c^2 \ , \tag{8.46}$$

where $w = -1$ corresponds to the cosmological constant. In order for this component to potentially lead to an accelerated expansion, the second Friedmann equation (4.22) requires $w < -1/3$. If we insert (8.46) into the first law of thermodynamics (4.17), we obtain $a^3\rho' + 3(1+w)\rho a^2 = 0$, where a prime denotes a derivative with respect to a. Making the power-law ansatz $\rho(a) \propto a^\beta$ and inserting this into the foregoing equation, one obtains

$$\rho_{\rm DE}(a) = \rho_{\rm DE,0}\, a^{-3(1+w)} \ . \tag{8.47}$$

For $w = -1$, we recover the behavior that $\rho_{\rm DE}$ is constant in time.

In fact, the equation-of-state parameter w does not need to be a constant, it can as well vary with cosmic epoch. Since we have no clue of what its behavior in time should be, we parametrize our ignorance by making some ansatz for a possible functional form, the most common one being

$$w(a) = w_0 + w_a(1-a) \ . \tag{8.48}$$

Note that there is no physical basis for this ansatz, it is just a linear (in a) expansion of w around its current value w_0.

Can one observationally distinguish between the case $w = -1$ for a cosmological constant (or a vacuum energy density that is indistinguishable from a cosmological constant) and the more interesting, dynamical case $w \neq -1$?

The first major impact of $w \neq -1$ on cosmology would be a change of the expansion rate $H(a)$ of the universe, which becomes (for constant w and a flat universe)

$$H^2(a) = H_0^2 \left(\frac{\Omega_{\rm r}}{a^4} + \frac{\Omega_{\rm m}}{a^3} + \frac{\Omega_{\rm DE}}{a^{3(1+w)}} \right) \ , \tag{8.49}$$

where $\Omega_{\rm DE} = \rho_{\rm DE,0}/\rho_{\rm cr}$ is the density parameter of dark energy at the current epoch. Second, the growth rate of structure would be affected for $w \neq -1$. The growth factor is the growing solution of (7.15), but for $w \neq -1$, the Hubble function is no longer a solution of the growth equation, and thus D_+ is no longer given by the explicit expression (7.17). Amazingly enough, one finds that the logarithmic derivative of $D_+(a)$, i.e., the function f [see (7.21)], is still very closely approximated by (7.22), with $\gamma \approx 0.55$, even for a varying w.

Outlook. Whereas before 2000, the main emphasis of determining the bulk cosmological parameters were focussed on H_0, $\Omega_{\rm m}$ and Ω_Λ, these parameters are now determined with previously unimaginable precision. The focus has since then switched to the possibilities for distinguishing $w \equiv -1$ from dynamical dark energy, also because this question lies at the heart of fundamental physics. A Dark Energy Task Force was established by the National Science Foundation in the USA to study the most sensitive probes for the equation-of-state of dark energy; a similar exercise was carried out by a joint Working Group of ESA and ESO. There are four cosmological probes which appear to be most promising for shedding light on w: SN Ia and BAOs are purely geometrical probes that are sensitive to the distance-redshift relation, and thus to the impact of w on the expansion rate. Recent constraints on w obtained by these two methods, in combination with the CMB fluctuations, are shown in Fig. 8.57.[19] The other two probes, cosmic shear and galaxy clusters, are in addition sensitive to the power spectrum of density fluctuations and the growth of structure. All four of these probes have their individual issues on systematics which we discussed before; in none of these cases is there a fundamental reason why these systematics cannot be overcome. However, considerable work needs to be invested to beat down systematics, such as the potential redshift-dependence of SN Ia luminosity, the potentially different clustering behavior of galaxies and matter on the baryonic acoustic scale, the shape measurements of faint galaxies and potential intrinsic alignment effects for cosmic shear, and the calibration of the mass-observable relations for clusters. For each of these probes, major surveys are currently planned to exploit their potential to its full extent. These include the German-Russian eROSITA space mission which will conduct an all-sky survey in X-rays and is expected to detect some 10^5 galaxy clusters up to, and above, redshift unity, with a projected launch date of 2015, and the ESA mission Euclid which will observe half the sky to study cosmic shear and BAOs up to redshift ~ 2.

[19] The 2013 Planck data analysis including Planck lensing measurements and results from BAOs yield the constraint $w = -1.08^{+0.11}_{-0.09}$.

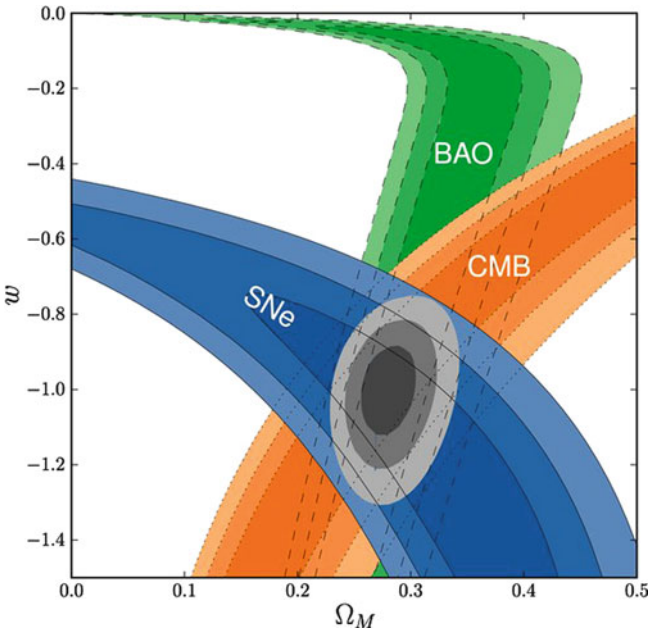


Fig. 8.57 Confidence regions in the parameter space spanned by $\Omega_{\rm m}$ and the dark energy equation-of-state parameter w. Three different techniques are combined in this figure: Type Ia supernovae (*blue*), the angular-diameter distance determination from baryonic acoustic oscillations in the galaxy distribution (*green*), and the CMB anisotropies from WMAP (*orange*). A flat cosmological model was assumed here. Any of the methods yields highly degenerate constraints on these two parameters; however, since their confidence regions have quite different 'orientations', the combined confidence region is quite localized, shown in *grey*. Source: R. Amanullah et al. 2010, *Spectra and Hubble Space Telescope Light Curves of Six Type Ia Supernovae at* $0.511 < z < 1.12$ *and the Union2 Compilation*, ApJ 716, 712, p. 731, Fig. 11. ©AAS. Reproduced with permission

Alternatives. But what if we are fully on the wrong track, fooled by our lack of knowledge on gravity? General Relativity, our current model of gravity, has passed all its tests with flying colors. Deviations of the law of gravity from the predictions of General Relativity have been constrained by precision observations within the Solar System and in the strong gravitational fields around neutron stars, possible through the exquisite timing precision of rotating neutron stars in the form of pulsars. However, General Relativity has not been tested in the very weak-field regime, or on very large scales. Is it feasible that our conclusions about dark energy is just a consequence of assuming the wrong law of gravity? Maybe the same is true for dark matter?

Starting with the latter issue, the answer is almost certainly 'no'. Whereas ad-hoc modifications of Newton's law of gravity can account for the rotation curves of spiral galaxies without the need for invoking dark matter, the CMB anisotropies cannot be explained without the presence of a dark matter component. We hope that the cosmic harmony presented in this chapter is seen as ample evidence that the ΛCDM model yields a coherent picture of our Universe, accounting for all relevant cosmological observations.

The same does not hold for the dark energy issue. Indeed, the introduction of Λ in (8.44) [but not in (8.45)] *is* a modification of the law of gravity! But maybe, there are other classes of gravity models which compete for the 'correct one'? Yes, there are, and they will be judged by their ability to provide a similarly coherent framework of cosmology (and other tests of gravity) as General Relativity does in the form of the ΛCDM model. Given that the logarithmic derivative f (7.21) of the growth factor shows an almost universal behavior in terms of $\Omega_{\rm m}(a)$ for General Relativity + dark energy models, as mentioned above, cosmological probes which are sensitive to f can find evidence for, or against, the validity of General Relativity. Of particular sensitivity are combined studies of large-scale structure and redshift-space distortions. In fact, several alternative gravity models have already been excluded by such cosmological tests. The upcoming dark energy surveys will have a largely increased sensitivity to such probes of the law of gravity.

8.9 Problems

8.1. Limber equation. Derive the Limber equation (8.17).

1. Let $n_3(\boldsymbol{x})$ be the comoving number density of galaxies of a certain kind. The number density on the sky is denoted by $n(\boldsymbol{\theta})$, and is given by the integral over n_3 along the line-of-sight, chosen to be in the x_3-direction,

$$n(\boldsymbol{\theta}) = \int \mathrm{d}x_3 \; \nu(x_3)\, n_3(f_k(x_3)\boldsymbol{\theta}, x_3) \;, \qquad (8.50)$$

where $\nu(x_3)$ is a selection function which contains the comoving volume element as a function of comoving distance x_3, and accounts for the fact that the observed galaxies are flux limited, and hence only the most luminous galaxies at large distances will make it into our survey. Show that the probability distribution of the observed galaxies in distance x_3, $p_x(x_3)$, is related to the selection function $\nu(x_3)$ by

$$p_x(x_3) = \nu(x_3)\, \bar{n}_3(x_3)/\bar{n} \;, \qquad (8.51)$$

where $\bar{n}_3(x_3)$ is the mean number density of galaxies at the cosmic epoch corresponding to the comoving distance x_3, and $\bar{n}$ the mean number density of observed galaxies on the sky.

2. Calculate $n(\boldsymbol{\theta})$ in terms of the distance probability distribution $p_x(x_3)$ and the comoving number density $n_3(\boldsymbol{x})$, and write the latter in terms of the galaxy number density contrast $\delta_{\rm g}(\boldsymbol{x})$.
3. Now derive the Limber equation (8.17), making use of (7.27). For this, you need to make some approximations; state them clearly.

8.2. Scaling of the observed baryon mass fraction in clusters. Assume to have X-ray data that cover an angular radius θ around the center of a cluster at given redshift $z_{\rm cl}$.

1. Relate the observed X-ray flux S_X to the X-ray luminosity L_X; furthermore, derive the relation between X-ray luminosity, electron density and θ, assuming that the emission region is spherical. By combining these two relations, show that the estimated electron density scales as $n_{\rm e} \propto D_{\rm A}^{-1/2}$, where $D_{\rm A}$ is the angular diameter distance to $z_{\rm cl}$. Hint: Recall the relation between angular-diameter- and luminosity distance in cosmology.
2. With the same assumptions, obtain an estimate for the gas mass inside θ and show that it scales like $M_{\rm gas} \propto D_{\rm A}^{5/2}$.
3. Finally, employing the assumption of hydrostatic equilibrium and using (6.37), show that the estimated total mass scales like $M \propto D_{\rm A}$. With the previous result, this now shows that the estimated gas-mass fraction scales as $f_{\rm gas} \propto D_{\rm A}^{1.5}$.

8.3. Flatness problem. If the dark energy equation-of-state is w, calculate the total density parameter $\Omega_0(z)$ and show that the flatness problem (see Sect. 4.5.2) still remains.

8.4. Slope of correlation functions. Show that a spatial correlation function of the form $\xi(r) \propto r^{-\gamma}$ yields an angular correlation function of $w(\theta) \propto \theta^{-(\gamma-1)}$. What is the corresponding behavior of the projected correlation function $w_{\rm p}(r_{\rm p})$?

The Universe at high redshift

9

In the previous chapter we explained by what means the cosmological parameters may be determined, and what progress has been achieved in recent years. This might have given the impression that, with the determination of the values for $\Omega_{\rm m}$, Ω_Λ etc., cosmology is nearing its conclusion. As a matter of fact, for several decades cosmologists have considered the determination of the density parameter and the expansion rate of the Universe as their prime task, and now this goal has largely been achieved. However, from this point on, the future evolution of the field of cosmology will probably proceed in two directions. First, we will try to uncover the nature of dark energy and to gain new insights into fundamental physics along the way. Second, astrophysical cosmology is much more than the mere determination of a few parameters. We want to understand how the Universe evolved from a very primitive initial state, as seen in the almost isotropic CMB radiation, into what we are observing around us today—galaxies of different morphologies, luminosities and spectral properties, the large-scale structure of their distribution, groups and clusters of galaxies, active galaxies, and the intergalactic medium. We seek to study the formation of stars and of metals, the cosmic history of star formation, and also the processes that reionized and enriched the intergalactic medium with heavy elements.

The boundary conditions for studying these processes are now very well defined. Until about the year 2000, the cosmological parameters in models of galaxy evolution, for instance, could be chosen from within a large range, because they had not been determined sufficiently well at that time. That allowed these models more freedom to adjust the model outcomes such that they best fit with observations. Today however, a successful model needs to make predictions compatible with observations, but using the parameters of the standard model. In terms of the cosmological parameters, there is little freedom left in designing such models. In other words, the stage on which the formation and evolution of objects and structure takes place is prepared, and now the cosmic play can begin.

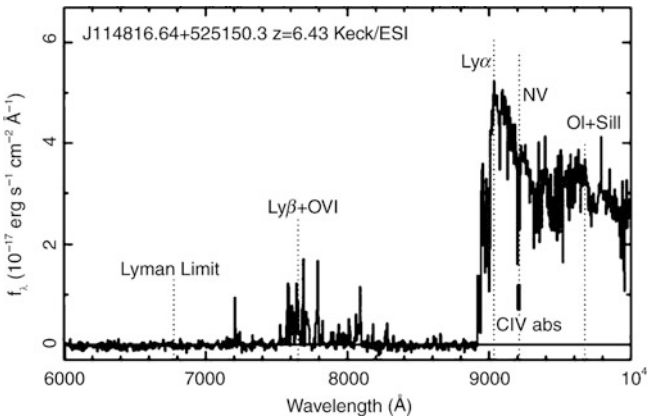


Fig. 9.1 Spectrum of a QSO at the high redshift of $z = 6.419$. Like many other QSOs at very high redshift, this source was discovered with the Sloan Digital Sky Survey. The spectrum was obtained with the Keck telescope. The redshifted Lyα line is clearly visible, its blue side 'eaten' away by intergalactic absorption. Almost all radiation bluewards of the central wavelength of the Lyα line is absorbed; however, a low level of this radiation is getting through, as is most clearly seen from the Lyβ line. For $\lambda \leq 7200\,$Å the spectral flux is consistent with zero; intergalactic absorption is too strong here. Source: X. Fan et al. 2003, *A Survey of $z > 5.7$ Quasars in the Sloan Digital Sky Survey. II. Discovery of Three Additional Quasars at $z > 6$*, AJ 125, 1649, p. 16554, Fig. 6. ©AAS. Reproduced with permission

Progress in recent years, with developments in instrumentation having played a vital role, has allowed us to examine the Universe at very high redshift. An obvious indication of this progress is the increasingly high maximum redshift of sources that can be observed; as an example, Fig. 9.1 presents the spectrum of a QSO at redshift $z = 6.419$ whose precise redshift was measured from molecular CO lines. Today, we know quite a few galaxies at redshift $z > 6$, i.e., we observe these objects at a time when the Universe had less than 10 % of its current age and when the density of the neutral hydrogen in the intergalactic medium was apparently considerably higher than at later epochs, as concluded from the very strong absorption blueward of the Lyα emission line (see Fig. 9.1). As we shall see, the detection of galaxies

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_9,

at even higher redshifts has been claimed. Besides larger telescopes, which enabled these deep images of the Universe, gaining access to new wavelength domains is of particular importance for our studies of the distant Universe. This can be seen, for example, from the fact that the optical radiation of a source at redshift $z \sim 1$ is shifted into the NIR. Because of this, near-infrared astronomy is about as important for galaxies at $z \gtrsim 1$ as optical astronomy is for the local Universe. Furthermore, the development of sub-millimeter astronomy has provided us with a view of sources that are nearly completely hidden to the optical eye because of strong dust absorption.

In this chapter, we will attempt to provide an impression of astronomy of the distant Universe, and shed light on some interesting aspects that are of particular importance for our understanding of the evolution of the Universe, whereas in Chap. 10, we will try to provide an impression of our theoretical understanding of the evolution of galaxies throughout the Universe. Both, observational as well as theoretical and numerical studies, are currently very rapidly developing fields of research, so we will simply address some of the main topics in this field today. We begin in Sect. 9.1 with a discussion of methods to specifically search for high-redshift galaxies, and we will then focus on a method by which galaxy redshifts can be determined solely from photometric information in several bands (thus, from the color of these objects). This method can be applied to deep multi-band sky images, and we will present some of the results from deep HST surveys, described in Sect. 9.2. We will also emphasize the importance of gravitational lenses as 'natural telescopes', which provide us with a deeper view into the Universe due to their magnification effect.

Gaining access to new wavelength domains paves the way for the discovery of new kinds of sources; in Sect. 9.3 we will present high-redshift galaxy populations, some of which have been identified by sub-millimeter and NIR observations. Some key properties of the high-redshift galaxy population will be described in Sect. 9.4, including their luminosity function; as will be shown there, the properties of galaxies in the early phases of our Universe are quite different from the present galaxies. In Sect. 9.5 we will show that, besides the CMB, background radiation also exists at other wavelengths, but whose nature is considerably different from that of the CMB; recent progress has allowed us to identify the nature of these cosmic backgrounds. Then, in Sect. 9.6, we will focus on the history of cosmic star formation, and show that at redshift $z \gtrsim 1$ the Universe was much more active than it is today—in fact, most of the stars that are observed in the Universe today were already formed in the first half of cosmic history. This empirical discovery is one of the aspects that one attempts to explain in the framework of models of galaxy formation and evolution. Finally, in Sect. 9.7 we will discuss the sources of gamma-ray bursts. These are explosive events which, for a very short time, appear brighter than all other sources of gamma rays on the sky put together. For about 25 years the nature of these sources was totally unknown; even their distance estimates were spread over at least seven orders of magnitude. Only since 1997 has it been known that these sources are of extragalactic origin.

9.1 Galaxies at high redshift

In this section we will first consider how distant galaxies can be found, and how to identify them as such. The properties of these high-redshift galaxies can then be compared with those of galaxies in the local Universe, which were described in Chap. 3. The question then arises as to whether galaxies at high z, and thus in the early Universe, look like local galaxies, or whether their properties are completely different. One might, for instance, expect that the mass and luminosity of galaxies are evolving with redshift since, as we have seen in Sect. 7.5.2, the mass function of dark matter halos changes during cosmic evolution. Examining the galaxy population as a function of redshift, one can trace the history of global cosmic star formation and analyze when most of the stars visible today have formed, and how the density of galaxies changes as a function of redshift. We will investigate some of these questions in this and the following sections.

How to find high-redshift galaxies? Until about 1995 only a few galaxies with $z > 1$ had been known; most of them were radio galaxies discovered by optical identification of radio sources. The most distant normal galaxy with $z > 2$ then was the source of the giant luminous arc in the galaxy cluster Cl 2244−02 (see Fig. 6.49). Very distant galaxies are expected to be faint, and so the question arises of how galaxies at high z can be detected at all.

The most obvious answer to this question may perhaps be by spectroscopy of a sample of faint galaxies. This method is not feasible though, since galaxies with $R \lesssim 22$ have redshifts $z \lesssim 0.5$, and spectra of galaxies with $R > 22$ are observable only with 4-m telescopes and with a very large investment of observing time.[1] Also, the problem of finding a needle in a haystack arises: most galaxies with $R \lesssim 24.5$ have redshifts $z \lesssim 2$ (a fact that was not established before 1995), so how can we detect the small fraction of galaxies with larger redshifts?

Narrow-band photometry. A more systematic method that has been applied is narrow-band photometry. Since hydrogen

[1] Readers not familiar with the optical/near-IR filter system may find it useful to consult Sect. A.4.2 in the Appendix at this point. We will also follow the usual practice and write $R = 22$ instead of $R = 22\,\mathrm{mag}$ in the following.

is the most abundant element in the Universe, one expects that some fraction of galaxies feature a Lyα emission line (as do all QSOs). By comparing two sky images, one taken with a narrow-band filter centered on a wavelength λ, the other with a broader filter also centered roughly on λ, this line emission can be searched for specifically. If a galaxy at $z \approx \lambda/(1216\,\text{Å}) - 1$ has a strong Lyα emission line, it should be particularly bright in the narrow-band image in comparison to the broad-band image, relative to other sources. This search for Lyα emission line galaxies had been almost without success until the mid-1990s. Among other reasons, one did not know what to expect, e.g., how faint galaxies at $z \gtrsim 3$ are and how strong their Lyα line would be. Another reason, which was found only later, was the leakage of the narrow-band filters for radiation at shorter and longer wavelength—the transmission of these filters was not close enough to zero for wavelengths outside the considered narrow range. We will see later that more recent narrow-band photometric surveys have indeed uncovered a population of high-redshift galaxies.

9.1.1 Lyman-break galaxies (LBGs)

The method. The breakthrough was obtained with a method that became known as the *Lyman-break method*. Since hydrogen is so abundant and its ionization cross section so large, one can expect that photons with $\lambda < 912$ Å are very heavily absorbed by neutral hydrogen in its ground state. Therefore, photons with $\lambda < 912$ Å have a low probability of escaping from a galaxy without being absorbed.

Intergalactic absorption also contributes. In Sect. 5.7 we saw that each QSO spectrum features a Lyα forest and often also Lyman-limit absorption. The intergalactic gas absorbs a large fraction of photons emitted by a high-redshift source at $\lambda < 1216$ Å, and virtually all photons with a rest-frame wavelength $\lambda \lesssim 912$ Å. As also discussed in Sect. 8.5.2, the strength of this absorption increases with increasing redshift. Combining these facts, we conclude that spectra of high-redshift galaxies should display a distinct feature—a 'break'—at $\lambda = 912$ Å for redshifts $z \lesssim 4$, and for higher redshifts, the break shifts more towards $\lambda = 1216$ Å. Furthermore, radiation with $\lambda \lesssim 912$ Å should be strongly suppressed by intergalactic absorption, as well as by absorption in the interstellar medium of the galaxies themselves, so that only a very small fraction of these ionizing photons will reach us.

From this, a strategy for the detection of galaxies at $z \gtrsim 3$ emerges. We consider three broad-band filters with central wavelengths $\lambda_1 < \lambda_2 < \lambda_3$, where their spectral ranges are chosen to not (or only marginally) overlap. If $\lambda_1 \lesssim (1+z)912\,\text{Å} \lesssim \lambda_2$, a galaxy containing young stars should appear relatively blue as measured with the filters λ_2 and λ_3, and be virtually invisible in the λ_1-filter: because of the absorption, it will drop out of the λ_1-filter (see Fig. 9.2). For this reason, galaxies that have been detected in this way are called *Lyman-break galaxies (LBGs)* or *drop-outs*. An example of this is displayed in Fig. 9.3.

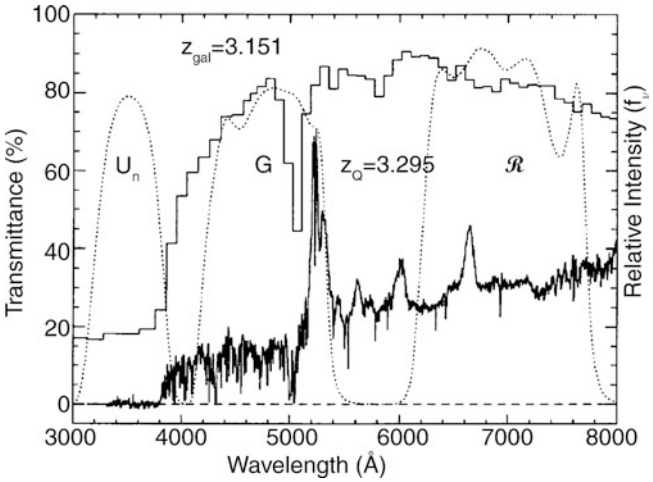


Fig. 9.2 Principle of the Lyman-break method. The histogram shows the synthetic spectrum of a galaxy at $z = 3.15$, generated by models of population synthesis; the spectrum belongs to a QSO at slightly higher redshift. Clearly, the decline of the spectrum at $\lambda \leq 912(1+z)$Å is noticeable. Furthermore, we see that the flux for $\lambda \leq 1216(1+z)$Å is reduced relative to the radiation on the red side of the Lyman-α emission line due to the integrated absorption of the intergalactic Lyman-α forest. For higher redshift sources, this latter effect becomes stronger, so that for them the break occurs already at a rest wavelength of $\lambda = 1216$ Å. The three *dotted curves* are the transmission curves of three broad-band filters, chosen such that one of them (U_n) blocks all photons with wavelengths above the Lyman-break. The color of this galaxy would then be blue in $G - \mathcal{R}$, and very red in $U_n - G$. Source: C.C. Steidel et al. 1995, *Lyman Imaging of High-Redshift Galaxies. III. New Observations of Four QSO Fields*, AJ 110, 2519, p. 2520, Fig. 1. ©AAS. Reproduced with permission

Large samples of LBGs. The method was first applied systematically in 1996, using the filters specified in Fig. 9.2. As can be read from Fig. 9.4, the expected location of a galaxy at $z \sim 3$ in a color-color diagram with this set of filters is nearly independent of the type and star formation history of the galaxy. Hence, sources in the relevant region of the color-color diagram are very good candidates for being galaxies at $z \sim 3$. The redshift needs to be verified spectroscopically, but the crucial point is that the color selection of candidates yields a very high success rate per observed spectrum, and thus spectroscopic observing time at the telescope is spent very efficiently in confirming the redshift of distant galaxies. With the commissioning of the Keck telescope (and later also of other telescopes of the 10-m class), spectroscopy of galaxies with $B \lesssim 25$ became possible (see Fig. 9.5). Employing this method, thousands of galaxies with $2.5 \lesssim$

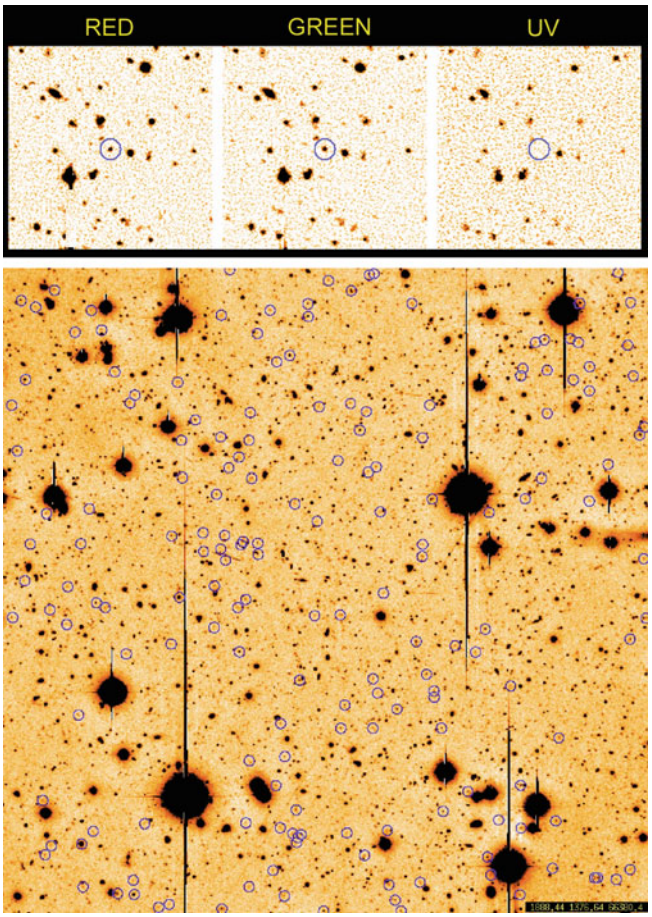


Fig. 9.3 *Top panel:* A U-band drop-out galaxy. It is clearly detected in the two redder filters, but vanishes almost completely in the U-filter. *Bottom panel:* In a single CCD frame, a large number of candidate Lyman-break galaxies (~ 150) are found. They are marked with *circles* here; their density is about 1 per square arcminute. Credit: C.C. Steidel

$z \lesssim 3.5$ have been detected and spectroscopically verified to date.

From the spectra shown in Fig. 9.5, it also becomes apparent that not all galaxies that fulfill the selection criteria also show a Lyα emission line, which provides one of the explanations for the lack of success in earlier searches for high-redshift galaxies using narrow-band filters. The spectra of the high-redshift galaxies which were found by this method are very similar to those of starburst galaxies at low redshift. It should come as no surprise that the galaxies selected by the drop-out technique feature active star formation, since it was required that the spectrum on the red side of the break—i.e., at (rest-frame) wavelengths above 1216 Å—shows a blue spectrum. Such a blue spectrum in the rest-frame UV is produced only by a stellar population which features active star formation. Furthermore, the luminosity of galaxies in the rest-frame UV and blue range strongly depends on the star-formation rate, so that preferentially galaxies with the highest (unobscured—see below) star-formation rate are selected.

This is a prominent example of the effect that the physical properties of objects selected depend on the selection criteria. One must always bear in mind that, when comparing galaxy populations detected by different methods, the properties can differ substantially. One of the challenges of studies of (high-redshift) galaxies is to get a coherent picture of the galaxy population from samples with a vast variety of selection methods.

The correlation function and halo masses of LBGs. For a large variety of objects, and over a broad range of separations, the spatial correlation function of objects can be described by the power law (7.28), with a slope of typically $\gamma \sim 1.7$. However, the amplitude of this correlation function varies between different classes of objects. For example, we saw in Sect. 8.2.4 that the amplitude of the power spectrum of galaxy clusters is larger by about a factor 7 than that of galaxies (see Fig. 8.23); the same ratio holds of course for the corresponding correlation functions. As we argued there, the strength of the correlation depends on the mass of objects; in the simple picture of biasing shown in Fig. 7.22, the correlation of objects is larger the rarer they are. High-mass peaks exceeding the density threshold needed for gravitational collapse have a lower mean number density than low-mass peaks, so they are therefore expected to be more biased (see Sect. 8.1.3) and thus more strongly correlated.

If we now assume that each galaxy lives in a dark matter halo, we can estimate the dark halo mass from the observed correlation function of these galaxies. As we discussed in Sect. 7.6.3, dark matter halos have clustering properties which differ from the clustering of the underlying matter density field, and we described that in terms of the halo biasing $b_{\rm h}(M,z)$, which is a function of halo mass and redshift. The dark matter correlation function can be determined quite accurately from numerical simulations. The ratio of the observed correlation function to the dark matter correlation function then yields the square of the halo bias parameter (7.68), and comparing that to the numerically-determined function $b_{\rm h}(M,z)$, the corresponding halo mass can be obtained.

Considering the spatial distribution of LBGs, we find a large correlation amplitude. The (comoving) correlation length of LBGs at redshifts $1.5 \lesssim z \lesssim 3.5$ is $r_0 \sim 4.2h^{-1}$ Mpc, i.e., not very different from the correlation length of L_*-galaxies in the present Universe. Since the bias factor of present-day galaxies is about unity, implying that they are clustered in a similar way as the dark matter distribution, this result then implies that the bias of LBGs at high redshift must be considerably larger than unity. This conclusion is based on the fact that the dark matter correlation at high redshifts (on large scales, i.e., in the linear regime) was smaller than today by the factor $D_+^2(z)$,

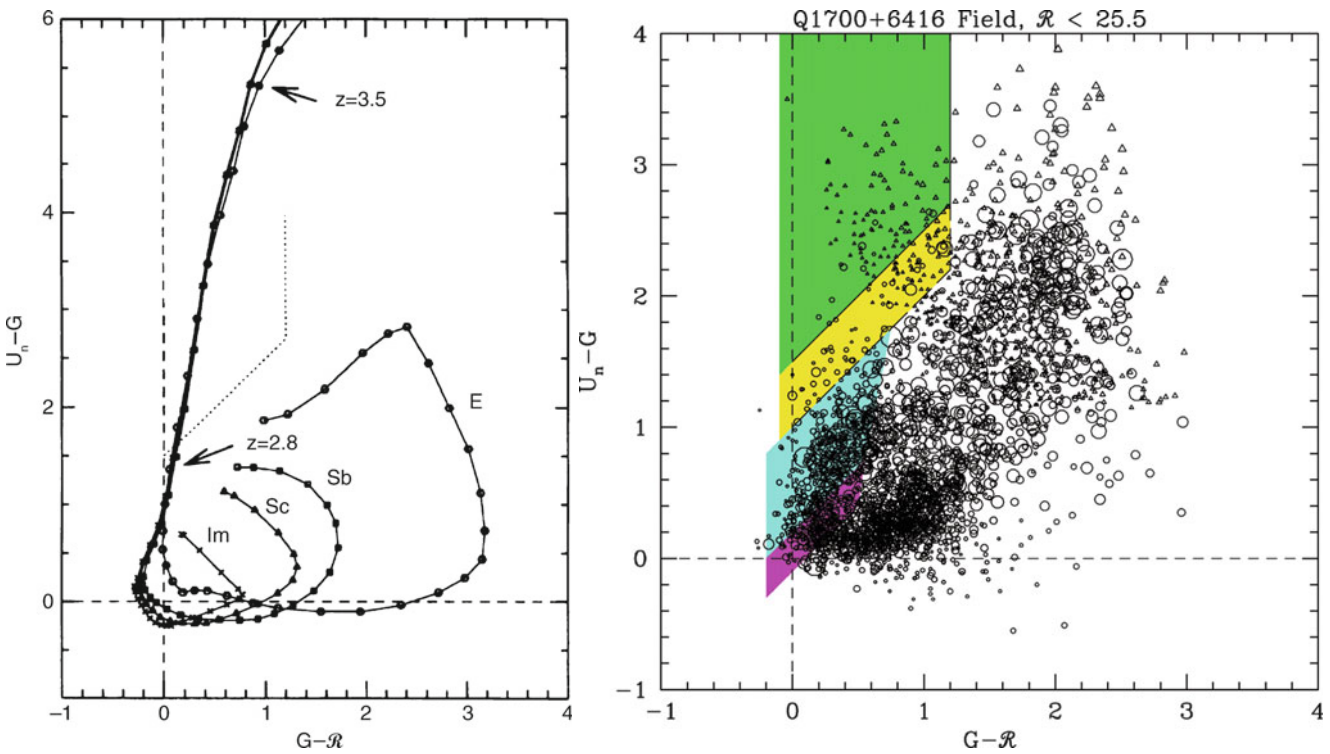


Fig. 9.4 *Left panel*: Evolutionary tracks of galaxies in the (G − $\mathcal{R}$) – (U_n−G) color-color diagram, for different types of galaxies, as obtained from population synthesis models. All evolutionary tracks start at $z = 0$, and the *symbols* along the curves mark intervals of $\Delta z = 0.1$. The colors of the various galaxy types are very different at lower redshift, but for $z \geq 2.7$, the evolutionary tracks for the different types nearly coincide—a consequence of the Lyα absorption in the intergalactic medium. Hence, a color selection of galaxies in the region between the *dotted* and *dashed curves* should select galaxies with $z \geq 3$. Indeed, this selection of candidates has proven to be very successful; thousands of galaxies with $z \sim 3$ have been spectroscopically verified. *Right panel*: The same color-color diagram, with objects selected from one survey field. The *green* and *yellow shaded* regions show the selection criteria for $z \sim 3$ Lyman-break galaxies, the *cyan* and *magenta* regions indicate the selection windows for galaxies with $z \sim 2.2$ and $z \sim 1.7$, respectively. The symbols are coded according to the brightness of the sources, and *triangles* denote sources for which only lower limits in the U_n − G color were obtained. Source: *Left*: C.C. Steidel et al. 1995, *Lyman Imaging of High-Redshift Galaxies. III. New Observations of Four QSO Fields*, AJ 110, 2519, p. 2522, Fig. 2. ©AAS. Reproduced with permission. *Right*: C.C. Steidel et al. 2004, *A Survey of Star-forming Galaxies in the* $1.4 \lesssim z \lesssim 2.5$ *Redshift Desert: Overview*, ApJ 604, 534, p. 537, Fig. 1. ©AAS. Reproduced with permission

where D_+ is the growth factor of linear perturbations introduced in Sect. 7.2.2. Thus we conclude that LBGs are rare objects and thus correspond to high-mass dark matter halos. Comparing the observed correlation length r_0 with numerical simulations, the characteristic halo mass of LBGs can be determined, yielding $\sim 3 \times 10^{11} M_\odot$ at redshifts $z \sim 3$, and $\sim 10^{12} M_\odot$ at $z \sim 2$. Furthermore, the correlation length is observed to increase with the luminosity of the LBG, indicating that more luminous galaxies are hosted by more massive halos, which are more strongly biased than less massive ones. If these results are combined with the observed correlation functions of galaxies in the local Universe and at $z \sim 1$, and with the help of numerical simulations, then this indicates that a typical high-redshift LBG will evolve into a massive elliptical galaxy by today.

Proto-clusters. Furthermore, the clustering of LBGs shows that the large-scale galaxy distribution was already in place at high redshifts. In some fields the observed overdensity in angular position and galaxy redshift is so large that one presumably observes galaxies which will later assemble into a galaxy cluster—hence, we observe some kind of proto-cluster. We have already shown such a proto-cluster in Fig. 6.71. Galaxies in such a proto-cluster environment seem to have about twice the stellar mass of those LBGs outside such structures, and the age of their stellar population appears older by a factor of two. This result indicates that the stellar evolution of galaxies in dense environments proceeds faster than in low-density regions, in accordance with expectations from structure formation. It also reveals a dependence of galaxy properties on the environment, which we have seen before manifested in the morphology-density relation (see Sect. 6.7.2). Proto-clusters of galaxies have also been detected at higher redshifts up to $z \sim 6$, using narrow-band imaging searches for Lyman-alpha emission galaxies (see below).

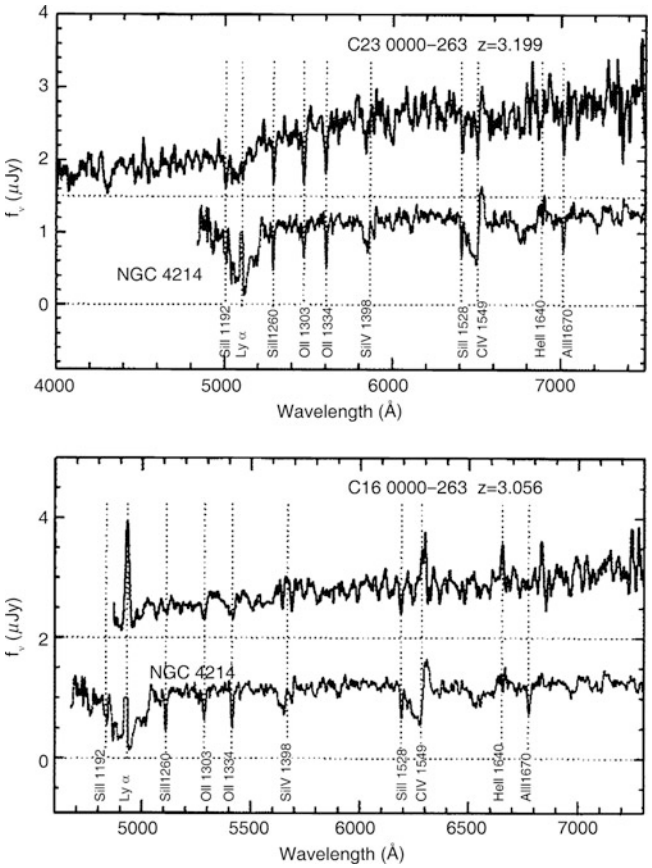


Fig. 9.5 Spectra of two galaxies at $z \sim 3$, detected by means of the U-drop-out technique. Below each spectrum, the spectrum of a nearby starburst galaxy (NGC 4214)—shifted to the corresponding redshift—is plotted; it becomes apparent that the spectra of galaxies at $z \sim 3$ are very similar to those of present-day star-forming galaxies. One of the two U-drop-out galaxies features a strong Lyα emission line, the other shows absorption at the respective wavelength. Source: C.C. Steidel et al. 1996, *Spectroscopic Confirmation of a Population of Normal Star-forming Galaxies at Redshifts* $z > 3$, ApJ 462, L17, PLATE L3, Fig. 1. ©AAS. Reproduced with permission

Satellite galaxies at high redshifts. Whereas the clustering of LBGs is well described by the power law (7.28) over a large range of scales, the correlation function exhibits a significant deviation from this power law at very small scales: the angular correlation function exceeds the extrapolation of the power law from larger angles at $\Delta\theta \lesssim 7''$, corresponding to comoving length-scales of $\sim$ 200 kpc. It thus seems that this scale marks a transition in the distribution of galaxies. To get an idea of the physical nature of this transition, we note that this length-scale is about the virial radius of a dark matter halo with $M \sim 3 \times 10^{11} M_\odot$, i.e., the mass of halos which host the LBGs. On scales below this virial radius, the correlation function thus no longer describes the correlation between two distinct dark matter halos. An interpretation of this fact is provided in terms of merging: when two galaxies and their dark matter halos merge, the resulting dark matter halo hosts both galaxies, with the more massive one close to the center and the other one as 'satellite galaxy'. The correlation function on scales below the virial radius thus indicates the clustering of galaxies within the same halo, whereas on larger scales, where it follows the power-law behavior, it indicates the correlation between different halos. Note that this effect is also well described in the halo model which we discussed in Sect. 7.7.3. On large scales, the correlation function is dominated by the two-halo term, whereas on smaller scales, the one-halo term takes over. The transition between these two regimes, which at low redshifts occurs on scales of several hundred kiloparsecs (see Fig. 7.27), is at smaller scales for high-redshift galaxies, since the high-mass population of galaxy clusters has not formed yet at these early epochs.

Winds of star-forming galaxies. The inferred high star-formation rates of LBGs implies an accordingly high rate of supernova explosions. These release part of their energy in the form of kinetic energy to the interstellar medium in these galaxies. This process will have two consequences. First, the ISM in these galaxies will be heated locally, which slows down (or prevents) further star formation in these regions. This thus provides a feedback effect for star formation which prevents all the gas in a galaxy from turning into stars on a very short time-scale, and is essential for understanding the formation and evolution of galaxies, as we shall see in Sect. 10.4.4. Second, if the amount of energy transferred from the SNe to the ISM is large enough, a galactic wind may be launched which drives part of the ISM out of the galaxy into its halo. Evidence for such galactic winds has been found in nearby galaxies, for example from neutral hydrogen observations of edge-on spirals which show an extended gas distribution outside the disk. Furthermore, the X-ray corona of spirals (see Fig. 3.26) is most likely linked to a galactic wind in these systems.

Indeed, there is now clear evidence for the presence of massive winds from LBGs. The spectra of LBGs often show strong absorption lines, e.g., of CIV, which are blueshifted relative to the velocity of the emission lines in the galaxy. An example of this effect can be seen in the spectra of Fig. 9.5, where in the upper panel the emission line of CIV is accompanied by an absorption to the short-wavelength side of the emission line. Such absorption can be produced by a wind moving out from the star-forming regions of the galaxy, so that its redshift is smaller than that of the emission regions. Characteristic velocities are $\sim$ 200 km/s. In one case where the spectral investigation has been performed in most detail (the LBG cB58; see Fig. 9.17), the outflow velocity is $\sim$

255 km/s, and the outflowing mass rate exceeds the star-formation rate. Whereas these observations clearly show the presence of outflowing gas, it remains undetermined whether this is a fairly local phenomenon, restricted to the star-formation sites, or whether it affects the ISM of the whole galaxy.

Connection to QSO absorption lines. A slightly more indirect argument for the presence of strong winds from LBGs comes from correlating the absorption lines in background QSO spectra with the position of LBGs. These studies have shown that whenever the sight-line of a QSO passes within ~ 40 kpc of an LBG, very strong CIV absorption lines (with column density exceeding $10^{14}\,\mathrm{cm}^{-2}$) are produced, and that the corresponding absorbing material spans a velocity range of $\Delta v \gtrsim 250$ km/s; for about half of the cases with impact parameters within 80 kpc, strong CIV absorption is produced. This frequency of occurrence implies that about 1/3 of all CIV metal absorption lines with $N \gtrsim 10^{14}\,\mathrm{cm}^{-2}$ in QSO spectra are due to gas within ~ 80 kpc from those LBGs which are bright enough to be included in current surveys. It is plausible that many of the remaining 2/3 are due to fainter LBGs.

The association of CIV absorption line systems with LBGs by itself does not prove the existence of winds in such galaxies; in fact, the absorbing material may be gas orbiting in the halo in which the corresponding LBG is embedded. In this case, no outflow phenomenon would be implied. However, in that case one might wonder where the large amount of metals implied by the QSO absorption lines is coming from. They could have been produced by an earlier epoch of star formation, but in that case the enriched material must have been expelled from its production site in order to be located in the outer part of $z \sim 3$ halos. It appears more likely that the production of metals in QSO absorption systems is directly related to the ongoing star formation in the LBGs. We shall see in Sect. 9.3.5 that clear evidence for superwinds has been discovered in one massive star-forming galaxy at $z \sim 3$.

Finally, we mention another piece of evidence for the presence of superwinds in star-forming galaxies. There are indications that the density of absorption lines in the Lyα forest is reduced when the sight-line to the QSO passes near a foreground LBG. This may well be explained by a wind driven out from the LBG, pushing neutral gas away and thus leaving a gap in the Lyα forest. The characteristic size of the corresponding 'bubbles' is estimated to be ~ 0.5 Mpc for luminous LBGs.

Lyman-break galaxies at low redshifts. One might ask whether galaxies similar to the LBGs at $z \sim 3$ exist in the current Universe. Until recently this question was difficult to investigate, since it requires imaging of lower redshift galaxies at ultraviolet wavelengths. With the launch of GALEX an appropriate observatory became available with which to observe galaxies with restframe UV luminosities similar to those of LBGs. UV-selected galaxies show a strong inverse correlation between the stellar mass and the surface brightness in the UV. Lower-mass galaxies are more compact than those of higher stellar mass. On the basis of this correlation we can consider the population of large and compact UV-selected galaxies separately. The larger ones show a star-formation rate of a few $M_\odot$/yr; at this rate, their stellar mass content can be built up on a time-scale comparable to the Hubble time, i.e., the age of the Universe. These galaxies are typically late-type spiral galaxies, and they show a metallicity similar to our Galaxy.

In contrast, the compact galaxies have a lower stellar mass and about the same star-formation rate, which allows them to generate their stellar population much faster, in about 1 Gyr. Compared to normal low-redshift galaxies, their metallicity is smaller by about a factor of 2 for a given stellar mass. In addition, they show similar extinction and outflow properties as the LBG at $z \sim 3$. Hence, the properties of the compact UV-selected galaxies, which are sometimes called Lyman-break analogs, are quite similar to those of the LBGs seen at higher redshifts, and they may indeed be closely related to the LBG population.

Lyman-break galaxies at high redshift. By variation of the filter set, drop-outs can also be discovered at larger wavelengths, thus at accordingly higher redshifts. The object selection at higher z implies an increasingly dominant role of the Lyα forest whose density is a strongly increasing function of redshift (see Sect. 8.5.2). This method has been routinely applied with ground-based observations up to $z \sim 4.5$, yielding so-called B-drop-outs. Galaxies at considerably higher redshifts are difficult to access from the ground with this method. One reason for this is that galaxies become increasingly faint with redshift, rendering observations substantially more difficult. Furthermore, one needs to use increasingly redder filter sets. At such large wavelengths the night sky gets significantly brighter, which further hampers the detection of very faint objects. For detecting a galaxy at redshift, say, $z = 5.5$ with this method, the Lyα line, now at $\lambda \approx 7900$ Å, is located right in the I-band, so that for an efficient application of the drop-out technique only the I- and z-band filters or NIR-filters are viable, and with those filters the brightness of the night sky is very problematic (see Fig. 9.6 for an example of a drop-out galaxy at very high redshift). Furthermore, candidate very high-redshift galaxies detected as drop-outs are very difficult to verify spectroscopically

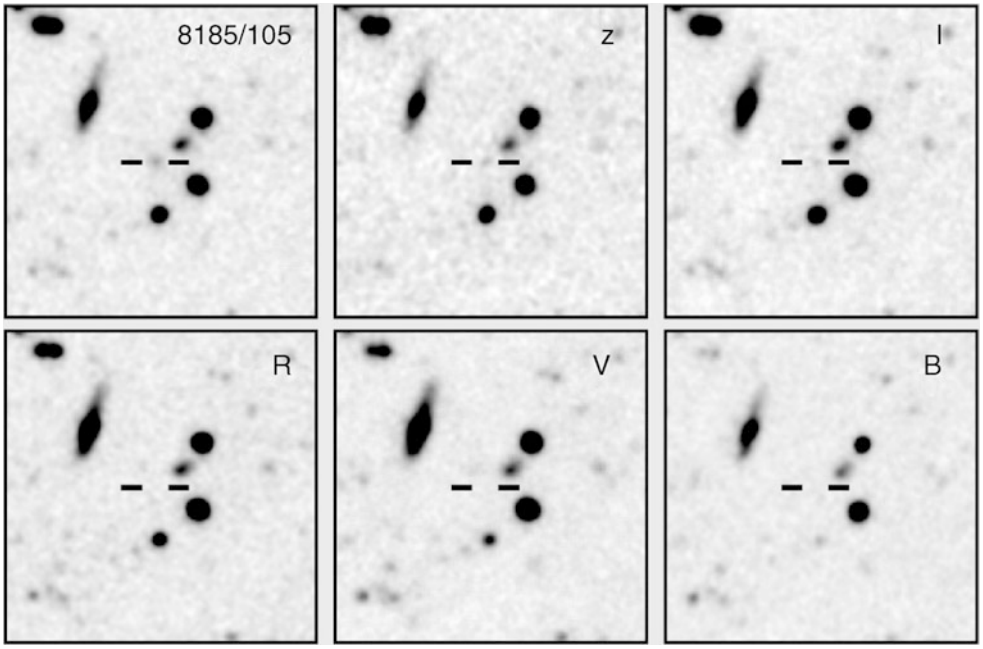


Fig. 9.6 A galaxy at $z = 5.74$, which is visible in the narrow-band filter (*upper left panel*) and in the I- and z-band (located between the two horizontal dashes), but which does not show any flux in the three filters at shorter wavelength (*lower panels*). Source: Hu et al. 1999, *An Extremely Luminous Galaxy at* $z = 5.74$, ApJ 522, L9, p. L10, Fig. 1. ©AAS. Reproduced with permission

due to their very faint flux and the fact that most of the diagnostic spectral features are shifted to the near-IR. In spite of this, we will see later that the drop-out method has achieved spectacular results even at redshifts considerably higher than $z \sim 4$, where the HST played a central role.

9.1.2 Photometric redshift

Spectral breaks. The Lyman-break technique is a special case of a method for estimating the redshift of galaxies (and QSOs) by multi-color photometry. This technique can be employed due to the spectral breaks at $\lambda = 912\,\text{Å}$ and $\lambda = 1216\,\text{Å}$, respectively. Spectra of galaxies also show other characteristic features. As was discussed in detail in Sect. 3.5, the broad-band energy distribution is basically a superposition of stellar radiation (if we ignore for a moment the presence of dust, which can yield a substantial infrared emission from galaxies). A stellar population of age $\gtrsim 10^8\,\text{yr}$ features a 4000 Å-break because, due to a sudden change in the opacity at this wavelength, the spectra of most stars show such a break at about 4000 Å (see Fig. 3.33). Hence, the radiation from a stellar population at $\lambda < 4000\,\text{Å}$ is less intense than at $\lambda > 4000\,\text{Å}$; this is the case particularly for early-type galaxies, as can be seen in Fig. 3.36, due to their old stellar population.

Principle of the method. If we assume that the star-formation histories of galaxies are not too diversified, the spectral energy distributions of these galaxies are expected to follow certain patterns. For example, if all galaxies had a single episode of star formation, starting at redshift z_f and lasting for a time τ, then the spectra of these galaxies, for a given initial mass function, would be characterized by these two parameters, as well as the total stellar mass formed (see Sect. 3.5); this latter quantity then yields the amplitude of the spectrum, but does not affect the spectral shape. When measuring the magnitude of these galaxies in n broad-band filters, we can form $n - 1$ independent colors. Next suppose we form a multi-dimensional color-color diagram, in which every galaxy is represented by a point in this $(n-1)$-dimensional color space. Considering only galaxies at the present epoch, all these points will lie on a two-dimensional surface in this multi-dimensional space, instead of being more or less randomly distributed.

Next, instead of plotting $z = 0$ galaxies, we consider the distribution of galaxies at some higher redshift $z < z_\text{f}$. This distribution of points will be different, mainly due to two different effects. First, a given photometric filter corresponds to a different rest-frame spectral range of the galaxy, due to redshift. Second, the ages of the stellar populations are younger at an earlier cosmic epoch, and thus the spectral energy distributions are different. Both of these effects will cause these redshift z galaxies to occupy a different two-dimensional surface in multi-color space.

Generalizing this argument further, we see that in this idealized consideration, galaxies will occupy a three-dimensional subspace in $(n-1)$-dimensional color space, parametrized by formation redshift z_f, time-scale τ and the galaxy's redshift z. Hence, from the measurement of the broad-band energy distribution of a galaxy, we might expect to be able to determine, or at least estimate, its redshift, as well as other properties such as the age of its stellar population; this is the principle of the method of *photometric redshifts*.

Of course, the situation is considerably more complicated in reality. Galaxies most likely have a more complicated star-formation history than assumed here, and hence they will not be confined to a two-dimensional surface at fixed redshift, but instead will be spread around this surface. The spectrum of a stellar population also depends on its metallicity, as well as absorption, either by gas and dust in the interstellar medium or hydrogen in intergalactic space (of which the Lyman-break method makes proper use). On the other hand, we have seen in Sect. 3.6 that the colors of current-day galaxies are remarkably similar, best indicated by the red sequence. Therefore, the method of photometric redshifts may be expected to work, even if more complications are accounted for than in the idealized example considered above.

The method is strongly aided if the galaxies have characteristic spectral features, which shift in wavelength as the redshift is changed. If, for example, the spectrum of a galaxy was a power law in wavelength, then the redshifted spectrum would as well be a power law, with the same spectral slope—if we ignore the different age of the stellar population. Therefore, for such a spectral energy distribution is would be impossible to estimate a redshift. However, if the spectrum shows a clear spectral break, then the location of this break in wavelength depends directly on the redshift, thus yielding a particularly clean diagnostic. In this context the 4000 Å-break and the Lyα-break play a central role, as is illustrated in Fig. 9.7.

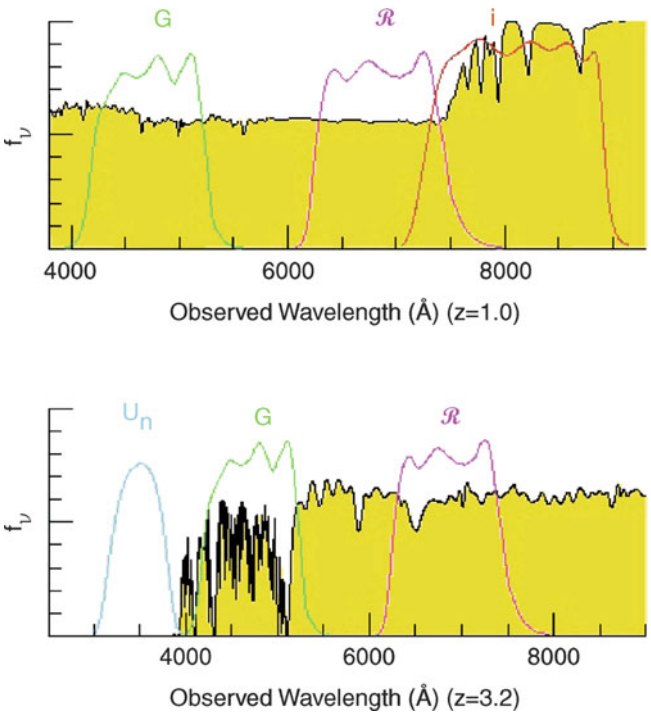


Fig. 9.7 The *bottom panel* illustrates again the principle of the drop-out method, for a galaxy at $z \sim 3.2$. Whereas the Lyman-α forest absorbs part of the spectral flux between (rest-frame wavelength) 912 and 1216 Å, the flux below 912 Å vanishes almost completely. By using different combinations of filters (*top panel*), an efficient selection of galaxies at other redshifts is also possible. The example shows a galaxy at $z = 1$ whose 4000 Å-break is located between the two redder filters. The 4000 Å-break occurs in stellar populations after several 10^7 yr (see Fig. 3.33) and is one of the most important features for the method of photometric redshift. Source: K.L. Adelberger 1999, *Star Formation and Structure Formation at Redshifts* $1 < z < 4$, astro-ph/9912153, Fig. 1

Calibration. In order to apply this method, one needs to find the characteristic domains in color space where (most of) the galaxies are situated. This can be done either empirically, using observed energy distributions of galaxies, or by employing population synthesis model. More precisely, a number of standard spectra of galaxies (so-called templates) are used, which are either selected from observed galaxies or computed by population synthesis models. Each of these template spectra can then be redshifted in wavelength. For each template spectrum and any redshift, the expected galaxy colors are determined by integrating the spectral energy distribution, multiplied by the transmission functions of the applied filters, over wavelength [see (A.25)]. This set of colors can then be compared with the observed colors of galaxies, and the set best resembling the observation is taken as an estimate for not only the redshift but also the galaxy type.

Practical considerations. The advantage of this method is that multi-color photometry is much less time-consuming than spectroscopy of galaxies. Whereas some modern spectrographs allow one to take spectra of $\sim$ 1000 objects simultaneously, images taken with wide-field cameras of $\sim 1\,\mathrm{deg}^2$ on 4-m class telescopes record the fluxes of $\sim 10^5$ galaxies in a one hour exposure. In addition, this method can be extended to much fainter magnitudes than are achievable for spectroscopic redshifts. The disadvantage of the method becomes obvious when an insufficient number of photometric bands are available, since then the photometric redshift estimates can yield a completely wrong z; these are often

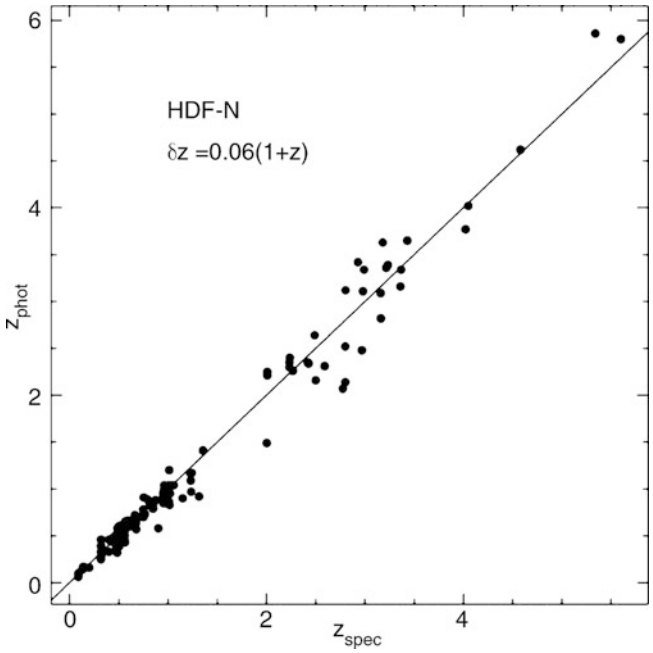


Fig. 9.8 Photometric redshift versus the spectroscopic redshift for galaxies in the HDF-North. Photometric data in four optical and two NIR bands have been used here. We see how accurate photometric redshifts can be—their quality depends on the photometric accuracy in the individual filters, the number of filters used, the redshift and the type of the galaxy, and also on details of the applied analysis method. Source: N. Benítez 2000, *Bayesian Photometric Redshift Estimation*, ApJ 536, 571, p. 579, Fig. 5. ©AAS. Reproduced with permission

called *catastrophic outliers*. One example for the occurrence of extremely wrong redshift estimates is provided by a break in the spectral energy distribution. Depending of whether this break is identified as the Lyman-break or the 4000 Å-break, the resulting redshift estimates will be very different. To break the corresponding degeneracy, a sufficiently large number of filters spread over a broad spectral range must be available to probe the spectral energy distribution over a wide range in wavelengths. As a general rule, the more photometric bands are available and the smaller the uncertainties in the measured magnitudes, the more accurate the estimated redshift. Normally, data from four or five photometric bands are required to obtain useful redshift estimates. In particular, the reliability of the photometric redshift benefits from data over a large wavelength range, so that a combination of several optical and NIR filters is desirable.

The successful application of this method also depends on the type of the galaxies. As we have seen in Sect. 6.8, early-type galaxies form a relatively well-defined color-magnitude sequence at any redshift, due to their old stellar populations (manifested in clusters of galaxies in form of the red cluster sequence), so that the redshift of this type of galaxy can be estimated very accurately from multi-color information. However, this is only the case if the 4000 Å-break is located in between two of the applied filters. For $z \gtrsim 1$ this is no longer the case if only optical filters are used. Other types of galaxies show larger variations in their spectral energy distribution, depending, e.g., on the star formation history, as mentioned before.

Photometric redshifts are particularly useful for statistical purposes, for instance in situations in which the exact redshift of each individual galaxy in a sample is of little relevance. However, by using a sufficient number of optical and NIR filters, quite accurate redshift estimates for individual galaxies are achievable. For example, with eight optical and NIR bands and accurate photometry, a redshift accuracy of $\Delta z \sim 0.03(1+z)$ was obtained, as demonstrated in Fig. 9.8 by a comparison of photometric redshifts with redshifts determined spectroscopically for galaxies in the field of the HDF-North. If data in additional photometric bands are available, e.g., using filters of smaller transmission curves ('intermediate width filters'), the redshift accuracy can be increased even more, e.g., $\Delta z \sim 0.01(1+z)$ was obtained using a total of 30 photometric bands.

9.1.3 Other few-band selection techniques

The Lyman-break technique is a special case of the photometric redshift method; it relies on only three photometric bands to select galaxies in a given redshift range, whereas in general, more bands are needed to obtain reliable redshift estimates. There are other cases where a few bands are sufficient for a fairly reliable selection of particular kinds of galaxies, or particular redshift regimes, some of which should be mentioned here.

Selection of $1.5 \lesssim z \lesssim 2.5$ galaxies. The success of the Lyman-break method is rooted in the fact that the observed colors of star-forming galaxies in a carefully selected triplet of filters is essentially independent on details of the star-formation history, metallicity and other effects, due to a very strong spectral break. This is illustrated in Fig. 9.4. The same figure also shows that the colors of galaxies with somewhat lower redshift are also very similar; for example, one sees that galaxies with $z \sim 1.8$ all have $U_n - G \sim 0$ and $G - \mathcal{R} \sim 0$. At that redshift, the Lyα-line is shortward of the U_n-band filter, and thus a star-forming galaxy has a rather flat spectrum across all three filters. As the redshift increases above $z \sim 2$, the Lyα-line moves into the U_n-band filter and thus increases the flux there; however, as we have seen, a large fraction of LBGs have rather low Lyα-flux, thus affecting the color only marginally. For redshifts higher than ~ 2.5, the break moves into the U_n-band, and the objects redden in $U_n - G$ and move onto the same sequence where LBGs are selected. Thus, with a single set of three filters (and thus the same optical images), one can select galaxies over the broad range of $1.5 \lesssim z \lesssim 3.5$.

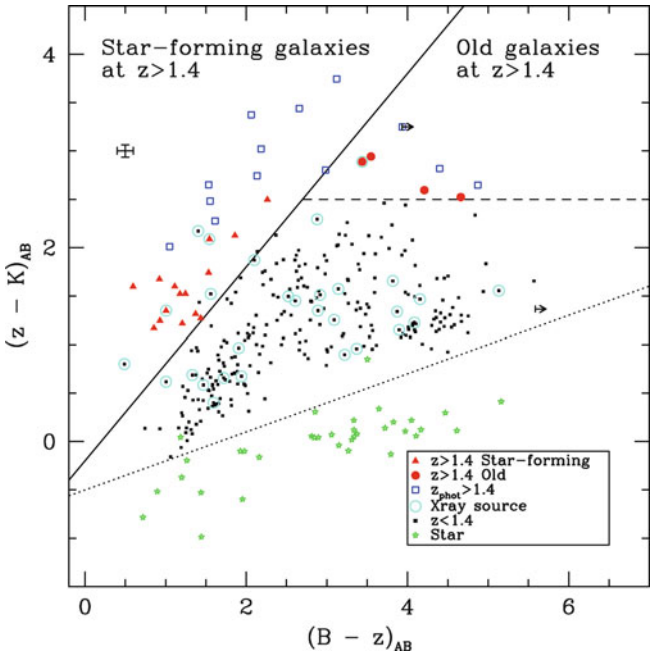


Fig. 9.9 Two color diagram $(B-z)$ vs $(z-K)$ for K-band selected galaxies of the K20 survey in the GOODS field. *Red solid triangles* and *circles* denote star-forming and passive galaxies, respectively, at $z \geq 1.4$, and *blue open squares* correspond to additional $z \geq 1.4$ objects as determined from their photometric redshifts. *Black solid squares* are galaxies with redshift below 1.4, and the *green asterisks* are stars. *Encircles symbols* are galaxies detected in X-rays. The various lines delineate regions of photometric selection of $z > 1.4$ galaxies—see text. Source: E. Daddi et al. 2004, *A New Photometric Technique for the Joint Selection of Star-forming and Passive Galaxies at* $1.4 \lesssim z \lesssim 2.5$, ApJ 617, 746, p. 749, Fig. 3. ©AAS. Reproduced with permission

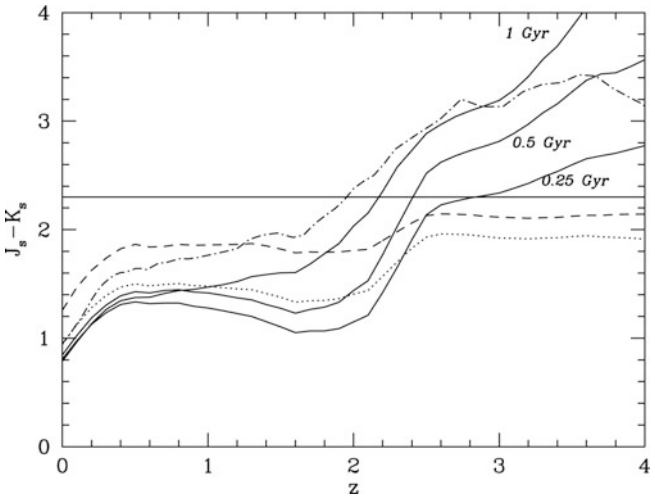


Fig. 9.10 The evolution of $(J-K)$ color as a function of redshift. *Solid curves* show the color for different ages of the stellar population. *Dashed* and *dotted curves* correspond to stellar populations with continuous star formation, for different ages and reddening. The *dash-dotted curve* corresponds to a single age population formed at $z = 5$. The color is redder than $(J-K) = 2.3$ for the single-age stellar populations at $z > 2.5$, and for the one formed at $z = 5$, this color criterion is satisfied for all $z > 2$. Source: M. Franx et al. 2003, *A Significant Population of Red, Near-Infrared-selected High-Redshift Galaxies*, ApJ 587, L79, p. L80, Fig. 1. ©AAS. Reproduced with permission

BzK selection. While the filter combination used for the Lyman-break galaxies selects star-forming galaxies at high redshift, it misses galaxies with a passive stellar population. One has therefore investigated whether another combination of filters, and thus different colors, may be able to identify high-redshift passive galaxies. Indeed, such a filter set was found; the combination of the B-, z- and K-band filters provides a successful tool to search for galaxies with $1.4 \lesssim z \lesssim 2.5$, as illustrated in Fig. 9.9. K-band selected galaxies with $1.4 \lesssim z \lesssim 2.5$ occupy specific regions in a $B-z$ versus $z-K$ color-color diagram.[2] In this redshift range, the 4000 Å-break is located redward of the z-band, thus such galaxies display a fairly red $z-K$ color if they are not forming stars at a high rate. The lack of active star formation also causes the $B-z$ color to be rather red, since the B-band probes to the rest-frame UV-region of the spectrum. Such galaxies are located in the upper right corner of the diagram in Fig. 9.9. In case the galaxies in this redshift range are actively forming stars, the 4000 Å-break is weaker, but instead the $B-z$ color is rather blue, so that these galaxies are located in the upper left corner of the diagram. As Fig. 9.9 shows, this selection of high-redshift galaxies is very efficient.

The BzK-selected galaxies with active star formation have redder colors in the rest-frame UV than the Lyman-break galaxies which are selected based on their UV flux, although there is a significant overlap between the two populations in the sense that a substantial fraction of galaxies are found by both methods. However, the most actively star-forming galaxies are missed with the BzK-method since those show little-to-no 4000 Å-break, thus no longer have a sufficiently red $z-K$ color, and would lie below the solid line in Fig. 9.9.

Distant red galaxies. Another method to select high-redshift passive galaxies is based on their rest-frame optical colors. From local galaxies we know that the 4000 Å-break is the most prominent feature in the spectral energy distribution of stellar populations with no or little star formation. At redshifts $2 \lesssim z \lesssim 4$, this break is located between the observed J- and K-band filters; hence we expect that passive galaxies are red in their $J-K$ color. As Fig. 9.10 shows, $J-K \gtrsim 2.3$ as soon as the redshift increases beyond $z \gtrsim 2$. Perhaps surprisingly, this is true even if the stellar population is as young as 0.25 Gyr, for which the redshift of the transition to $J-K > 2.3$ occurs at only slighter larger redshift. Furthermore, this color selection is able to find also galaxies with ongoing star formation, provided they also have an old stellar population; this is due to the fact that

[2] Whereas the symbols for redshift and the z-band magnitudes are identical, we trust that no confusion will arise by that, as the meaning will always be clear by the context.

much of the star formation is accompanied by substantial dust obscuration. At redshift $z = 2$, the J-band corresponds to the rest-frame B-band, which is substantially affected by extinction, leading to a red $J - K$ color. High-redshift galaxies selected according to $J - K > 2.3$ are called *distant red galaxies (DRGs)*. The fact that there is very little overlap in the galaxy population selected according to their UV-properties and the DRG population immediately shows the necessity to apply several very different selection criteria for high-redshift galaxies to obtain a complete census of their population.

Narrow-band selection. We mentioned the method of narrow-band selection before. If a source has a strong emission line, and if the observed wavelength of the emission line matches the spectral response of a narrow-band filter, then the ratio of fluxes obtained in this narrow-band image compared to a broad-band image would be much larger than for other sources without a strong emission line at the corresponding wavelength.

After a substantial population of high-redshift galaxies were found with the Lyman Break technique, it became known that about 60 % of these galaxies show very strong Lyα emission lines. It was then possible to design narrow-band filters that were particularly tuned to detect objects with strong Lyα emission lines at a particular redshift. Several thousand Lyα emitters (LAEs) were detected with this method, extending up to redshift $z \sim 7$. These galaxies are on average considerably fainter than LBGs, and therefore allow one to probe the fainter end of the luminosity function of star-forming galaxies. Their faintness, on the other hand, make more detailed spectroscopic studies very challenging, and thus the relation of these Lyα emitters to the other galaxy populations at similar redshifts is not easy to determine.

Furthermore, candidate objects detected in narrow-band images require spectroscopic follow-up, since there are many possible contaminants that may enter the selection. Galaxies, and in particular AGNs, at lower redshifts can display strong emission lines of other atomic transitions and need to be ruled out with a spectrum. Due to the cumulative effect of the Lyα forest, a high-redshift ($z \gtrsim 4$) Lyα emitter should show essentially no flux at shorter wavelengths, and so some of the Lyα emission-line candidates can be rejected if continuum flux bluewards of the narrow band is detected.

9.2 Deep views of the Universe

Very distant objects in the Universe are expected to be exceedingly faint. Therefore, in order to find the most distant, or earliest, objects in the Universe, very deep images of the sky are needed to have a chance to detect them.

In order to get further out into the Universe, astronomers use their most sensitive instruments to obtain extremely deep sky images. The Hubble Deep Field, already discussed briefly in Sect. 1.3.3, is perhaps the best-known example for this. As will be discussed below, further instrumental developments have led to even deeper observations with the HST. Deep fields are taken also with ground-based optical and near-IR telescopes. Although the sensitivity limit from the ground is affected by the atmosphere, in particular at longer wavelengths, this drawback is partly compensated by the larger field-of-view that many ground-based instruments offer, compared to the relatively small field-of-view of the HST. Dep field observations are conducted also at other wavelengths, preferentially in the same sky areas as the deep optical fields, to enable cross-identification and thus provide additional information on the detected sources. As we shall see, the availability of such deep fields has allowed us to take a first look at the first 10 % of the Universe's life.

9.2.1 Hubble Deep Fields

The Hubble Deep Field North. In 1995, an unprecedented observing program was conducted with the HST. A deep image in four filters (U_{300}, B_{450}, V_{606}, and I_{814}) was observed with the Wide Field/Planetary Camera 2 (WFPC2) on-board HST, covering a field of ~ 5.3 arcmin^2, with a total exposure time of about 10 days. This resulted in the deepest sky image of that time, displayed in Fig. 1.37. The observed field was carefully selected such that it did not contain any bright sources. Furthermore, the position of the field was chosen such that the HST was able to continually point into this direction, a criterion excluding all but two relatively small regions on the sky, due to the low HST orbit around the Earth. Another special feature of this program was that the data became public right after reduction, less than a month after the final exposures had been taken. Astronomers worldwide immediately had the opportunity to scientifically exploit these data and to compare them with data at other frequency ranges or to perform their own follow-up observations. Such a rapid and wide release was uncommon at that time, but is now seen more frequently. Rarely has a single data set inspired and motivated a large community of astronomers as much as the *Hubble Deep Field* (HDF) did (after another HDF was observed in the Southern sky—see below—the original HDF was called HDF North, or HDFN).

Follow-up observations of the HDF were made in nearly all accessible wavelength ranges, so that it became the best-observed region of the extragalactic sky. Within a few years, more than $\sim$ 3000 galaxies, 6 X-ray sources, 16 radio sources, and fewer than 20 Galactic stars were detected in the HDF, and redshifts were determined spectroscopically for more than 150 galaxies in this field, with about 30 at

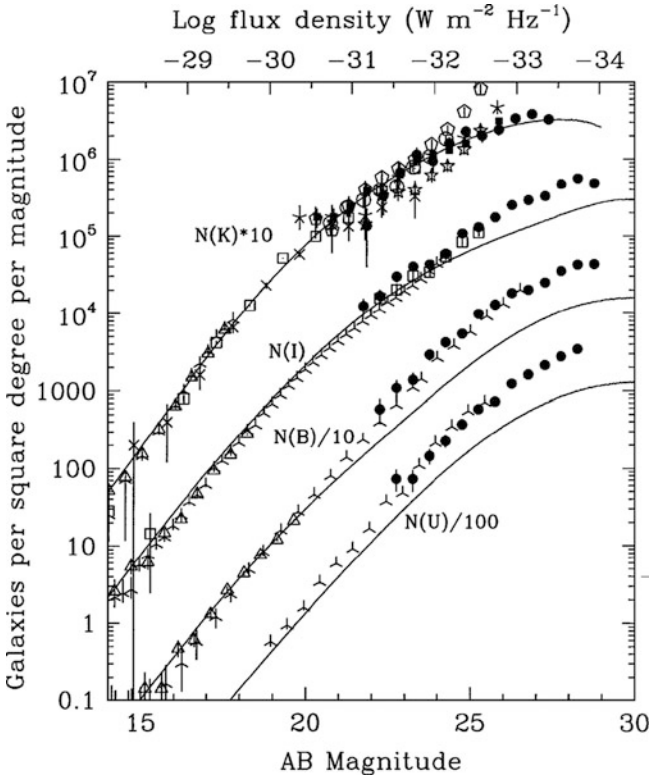


Fig. 9.11 Galaxy counts from the HDF and other surveys in four photometric bands: U, B, I, and K. *Solid symbols* are from the HDF, *open symbols* from various ground-based observations. The *curves* represent predictions from models in which the spectral energy distribution of the galaxies does not evolve—the counts lie significantly above these so-called non-evolution models: clearly, the galaxy population must be evolving. Note that the counts in the different color filters are shifted by a factor 10 each, simply for display purposes. Source: H. Ferguson et al. 2000, *The Hubble Deep Fields*, ARA&A 38, 667, Fig. 4. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 38 ©2000 by Annual Reviews www.annualreviews.org

$z > 2$. Never before could galaxy counts be conducted to magnitudes as faint as it became possible in the HDF (see Fig. 9.11); several hundred galaxies per square arcminute could be photometrically analyzed in this field.

Detailed spectroscopic follow-up observations were conducted by several groups, through which the HDF became, among other things, a calibration field for photometric redshifts (see, for instance, Fig. 9.8). Most galaxies in the HDF are far too faint to be analyzed spectroscopically, so that one often has to rely on photometric redshifts.

HDFS and the Hubble Ultra Deep Field. Later, in 1998, a second HDF was observed, this time in the southern sky. In contrast to the HDFN, which had been chosen to be as empty as possible, the HDFS contains a QSO. Its absorption line spectrum can be compared with the galaxies found in the HDFS, by which one hopes to obtain information on the relation between QSO absorption lines and galaxies. In addition to the WFPC2 camera, the HDFS was simultaneously observed with the cameras STIS ($51'' \times 51''$ field-of-view, where the CLEAR 'filter' was used, which has a very broad spectral sensitivity; in total, STIS is considerably more sensitive than WFPC2) and NICMOS (a NIR camera with a maximum field-of-view of $51'' \times 51''$) which had both been installed in the meantime. Nevertheless, the overall impact of the HDFS was smaller than that of the HDFN; one reason for this may be that the requirement of the presence of a QSO, combined with the need for a field in the continuous viewing zone of HST, led to a field close to several very bright Galactic stars. This circumstance makes photometric observations from the ground very difficult, e.g., due to stray-light.

One of the immediate results from the HDF was the finding that the morphology of faint galaxies is quite different from those in the nearby Universe. Locally, most luminous galaxies fit into the morphological Hubble sequence of galaxies. This ceases to be the case for high-redshift galaxies. In fact, galaxies at $z \sim 2$ are much more compact than local luminous galaxies, they show irregular light distributions and do not resemble any of the Hubble sequence morphologies. By redshifts $z \sim 1$, the Hubble sequence seems to have been partly established.

In 2002, an additional camera was installed on-board HST. The *Advanced Camera for Surveys* (ACS) has, with its side length of 3′.4, a field-of-view about twice as large as WFPC2, and with half the pixel size (0″.05) it better matches the diffraction-limited angular resolution of HST. Therefore, ACS is a substantially more powerful camera than WFPC2 and is, in particular, best suited for surveys. With the *Hubble Ultra Deep Field* (HUDF), the deepest image of the sky was observed and published in 2004 (see Fig. 9.12). The HUDF is, in all four filters, deeper by about one magnitude than the HDF, reaching a magnitude limit of $m_{\rm AB} \approx 29$. The depth of the ACS images in combination with the relatively red filters that are available provides us with an opportunity to identify drop-out candidates out to redshifts $z \sim 6$; quite a number of such candidates have already been verified spectroscopically.

Lyman-break galaxies at $z \sim 6$ seem to have stellar populations with masses and lifetimes comparable to those at $z \sim 3$. This implies that at a time when the Universe was 1 Gyr old, a stellar population with mass $\sim 3 \times 10^{10} M_\odot$ and age of a few hundred million years (as indicated by the observed 4000 Å break) was already in place. This, together with the apparently high metallicity of these sources, is thus an indication of how quickly the early Universe evolved. The $z \sim 6$ galaxies are very compact, with half-light radii of $\sim$ 1 kpc, and thus differ substantially from the galaxy population known in the lower-redshift Universe.

Half of the HUDF was also imaged by the near-IR NICMOS camera onboard HST, but only after the installment

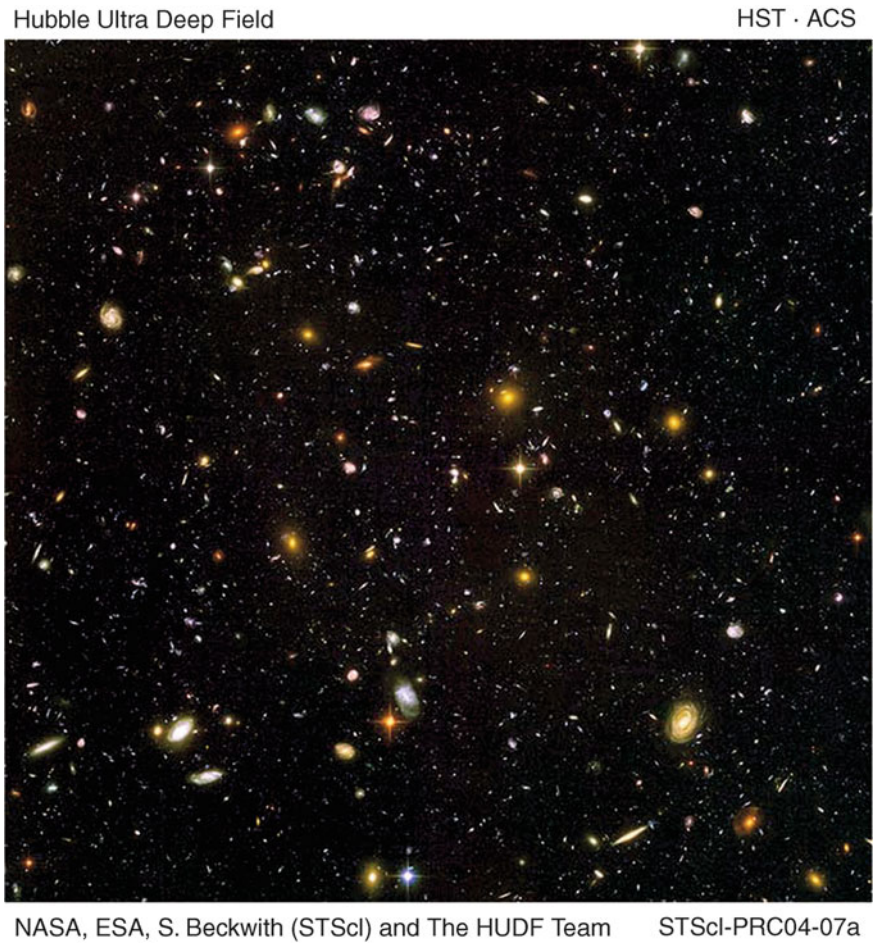


Fig. 9.12 The Hubble Ultra Deep Field, a field of $\sim 3'.4 \times 3'.4$ observed by the ACS camera. The limiting magnitude up to which sources are detected in this image is about one magnitude fainter than in the HDF. More than 10 000 galaxies are visible in the image, many of them at redshifts $z \geq 5$. Credit: NASA, ESA, S. Beckwith/Space Telescope Science Institute, and the HUDF Team

of the *Wide-Field Camera 3* (WFC3) on HST, with a near-IR channel and a much larger field-of-view and much better sensitivity than NICMOS, could the HUDF be imaged to comparable sensitivity levels in the NIR as in the optical. WFC3 mapped the HUDF in three NIR bands, Y, J and H, down to a limiting magnitude of $m_{\rm AB} \approx 28.5$. These long wavelengths allowed the systematic search for Lyman-break galaxies at redshifts beyond 6, as will be discussed below.

In September 2012, the deepest view of the Universe ever taken was released: the *eXtreme Deep Field* (XDF), shown in Fig. 9.13. It covers about 4.7 arcmin^2 in nine optical and NIR filters, reaching a limiting magnitude of $m_{\rm AB} \approx 30$.

Further deep-field projects with HST: GOODS, GEMS, COSMOS. The great scientific harvest from the deep HST images, particularly in combination with data from other telescopes and the readiness to make such data available to the scientific community for multi-frequency analysis, provided the motivation for additional HST surveys. The GOODS (Great Observatories Origins Deep Surveys) project is a joint observational campaign of several observatories, centering on two fields of $\sim 16' \times 10'$ size each that have been observed by the ACS camera at several epochs between 2003 and 2005. One of these two regions (GOODS-North) contains the HDFN, the other a field that became known as the Chandra Deep Field South (CDFS), also containing the HUDF. The Chandra satellite observed both GOODS fields with a total exposure time of $\sim 2 \times 10^6$ s and $\sim 4 \times 10^6$ s, corresponding to about 24 and 48 days, respectively. Also, the Spitzer observatory took long exposures of these two fields, and several ground-based observatories are involved in this survey, for instance by contributing an ultra-deep wide-field image ($\sim 30' \times 30'$) centered on the Chandra Deep Field South and NIR images in the K-band. The data themselves and the data products (like object catalogs, color information, etc.) are all publicly available and have led to a large number of scientific results.

The multi-wavelength approach by GOODS yields an unprecedented view of the high-redshift Universe. Although these studies and scientific analysis are ongoing (at the time of writing), quite a large number of very high-redshift ($z \gtrsim 5$–6) galaxies were discovered and studied: a sample of more than 500 I-band drop-out candidates was obtained from deep ACS/HST images.

Even larger surveys were conducted with the HST, including the Galaxy Evolution from Morphology and SEDs

Fig. 9.13 The Hubble eXtreme Deep Field (XDF) covers an area of $2'.3 \times 2'$, centered on the HUDF, and was composed with HST observations spread over many years. The total exposure time amounts to about 2 million seconds, or 22 days. This color composite was made from data in eight different optical and NIR bands, taken with the ACS and WFC3/IR instruments. Credit: NASA, ESA, G. Illingworth, D. Magee, and P. Oesch (University of California, Santa Cruz), R. Bouwens (Leiden University), and the HUDF09 Team

(GEMS) survey, covering a field of $30' \times 30'$ centered on the CDFS mapped in two filters. For this field, full coverage with a 17-band (5 broad bands, and 12 intermediate width bands) optical imaging survey (COMBO17) is available. The largest contiguous field imaged with the angular resolution of HST is the $\sim 2\,\mathrm{deg}^2$ COSMOS survey. This sky area was also imaged with other space-based (Spitzer, GALEX, XMM-Newton, Chandra) and ground-based (Subaru, VLA, VLT, UKIRT and others) observatories. Its large field enables the study of the large-scale galaxy and AGN distribution at high redshifts. The broad wavelength coverage from the radio to the X-ray regime, renders the COSMOS field a treasure for observational cosmology for years to come. As discussed in Sect. 8.4, the COSMOS field was used for a detailed cosmic shear analysis, where the broad wavelength coverage helped enormously to determine accurate photometric redshifts of the source galaxies.

9.2.2 Deep fields in other wavebands

Deep fields have been observed in many other frequency ranges, and as mentioned before, they are often taken on the same sky areas to allow for multi-frequency studies of the sources. We mention here the *Chandra Deep Fields*, one in the North (CDFN) in the direction of the HDFN, the other in the South (CDFS, see Fig. 9.14) in the direction of the HUDF, with a total exposure time of two and four million seconds, respectively.

With the extremely faint limiting flux of the CDFS, the source density on the sky reaches more than $10^4\,\mathrm{deg}^{-2}$, which is comparable to the source density seen in rather shallow optical imaging, like the SDSS images. However, the mean redshift of the X-ray sources is considerably higher than that of the optical sources, due to the large fraction of AGNs. As we can see from Fig. 9.15, the X-ray population of compact sources in ultra-deep X-images is composed of three components. By far dominating is the population of AGNs, whose number counts exhibit the shape of a broken power law—steep at the bright end, flatter at the low-flux end. For high fluxes, they follow approximately the Euclidean slope, $\mathrm{d}N/\mathrm{d}S \propto S^{-2.5}$, whereas at fluxes below the breakpoint at $S_\mathrm{b} \sim 6 \times 10^{-15}\,\mathrm{erg\,cm^{-2}\,s^{-1}}$, one finds $\mathrm{d}N/\mathrm{d}S \propto S^{-1.5}$ for the soft band, and even slightly flatter in the hard band. Going to fainter X-ray fluxes, the fraction of obscured AGNs at high redshifts increases.

Approaching the fainter flux levels, the population of normal galaxies becomes increasingly important. At the

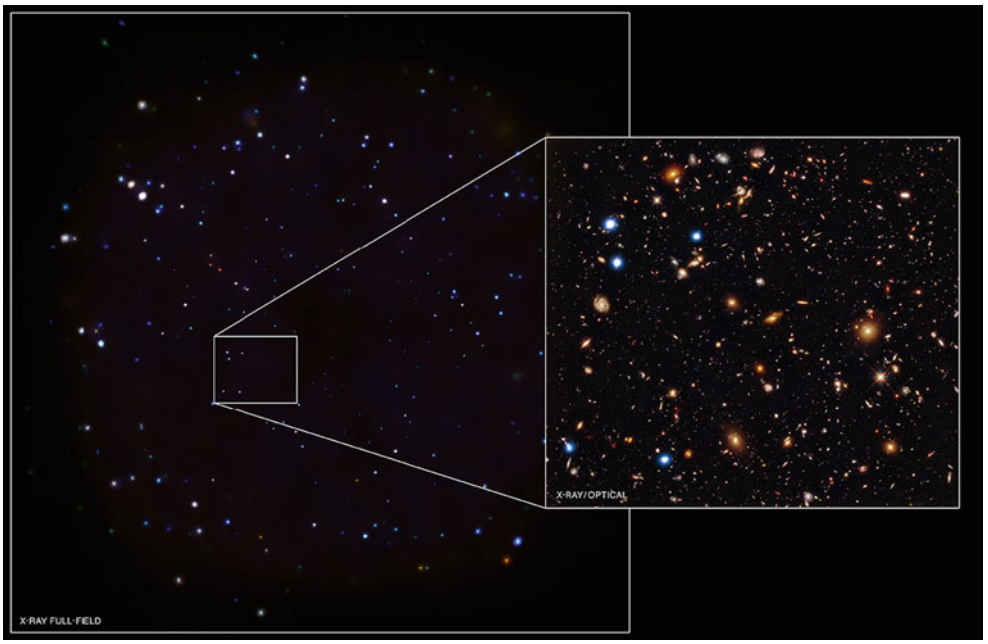


Fig. 9.14 The Chandra Deep Field South (CDFS), an X-ray image of a $\sim$ 450 arcmin^2 field with a total exposure time of 4×10^6 s. This is the deepest image ever taken of the X-ray sky. The *right panel* shows a color composite image taken with HST. Credit: X-ray: NASA/CXC/U.Hawaii/E.Treister et al.; optical: NASA/STScI/S.Beckwith et al.)

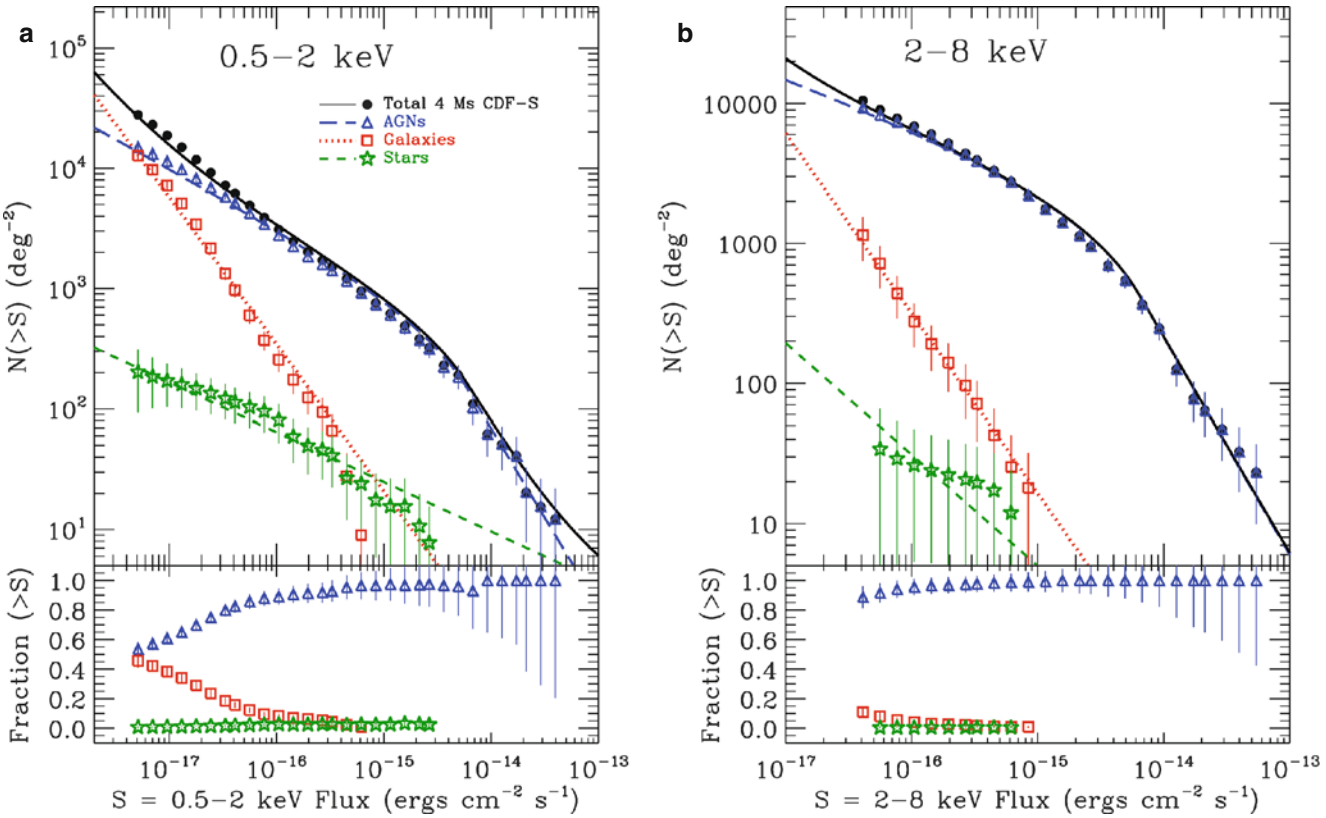


Fig. 9.15 Cumulative compact source counts $N(>S)$ as obtained from the CDFS (see Fig. 9.14), in the soft (**a**) and hard (**b**) energy bands, split according to source populations: AGNs (*blue*), galaxies (*red*) and Galactic stars (*green*). Source: B.D. Lehmer et al. 2012, *The 4 Ms Chandra Deep Field-South Number Counts Apportioned by Source Class: Pervasive Active Galactic Nuclei and the Ascent of Normal Galaxies*, ApJ 752:46, p. 7, Fig. 5. ©AAS. Reproduced with permission

Fig. 9.16 HST images of the lensing cluster Abell 68 magnifying many faint background galaxies. Several $z > 5$ candidates are found in this field, at least one of them spectroscopically verified. Credit: NASA, ESA, and the Hubble Heritage/ESA-Hubble Collaboration, Acknowledgment: N. Rose

limiting flux of the CDFS, they constitute almost 50 % of the source population. They are mainly late-type star-forming galaxies, with the X-ray emission mostly due to X-ray binaries. Early-type galaxies contribute just a small fraction of $\sim$ 10 % to the galaxy counts, where their X-ray emission is a due to a combination of X-ray binaries and hot gas.

9.2.3 Natural telescopes

Galaxies at high redshift are faint and therefore difficult to observe spectroscopically. For this reason, the brightest galaxies are preferentially selected (for detailed examination), i.e., basically those which are the most luminous ones at a particular z—resulting in undesired, but hardly avoidable selection effects. For example, those Lyman-break galaxies at $z \sim 3$ for which the redshift is verified spectroscopically are typically located at the bright end of the LBG luminosity function. The sensitivity of our telescopes is insufficient in most cases to spectroscopically analyze a rather more typical galaxy at $z \sim 3$.

The magnification by gravitational lenses can substantially boost the apparent magnitude of sources; gravitational lenses can thus act as natural (and inexpensive!) telescopes. The most prominent examples are the arcs in clusters of galaxies: many of them have a very high redshift, are magnified by a factor $\gtrsim$ 5, and hence are brighter by about $\gtrsim$1.5 mag than they would be without the lens effect (see Fig. 9.16).[3] In addition to boosting the observable fluxes, the gravitational lens effect yields a spatial magnification of the sources: the sources appear larger than they really are, thus increasing the effective angular resolution with which they can be observed.

Lyman-break galaxies at $z \sim 3$. An extreme first example of this effect is represented by the galaxy cB58 at $z = 2.72$, which is displayed in Fig. 9.17. It was discovered in the background of a galaxy cluster and is magnified by a factor $\sim$ 30. Hence, it appears brighter than a typical Lyman-break galaxy by more than three magnitudes. For this reason, one of the most detailed spectra of all galaxies at $z \sim 3$ was taken of this particular source. Several more such examples were found subsequently, including the so-called Cosmic Eye, a

[3]Note that a factor of 5 in magnification corresponds to a factor 25 in the exposure time required for spectroscopy. This factor of 25 makes the difference between an observation that is feasible and one that is not. Whereas the proposal for a spectroscopic observation of 3 h exposure time at an 8-m telescope may be successful, a similar proposal of 75 h would be hopelessly doomed to failure.

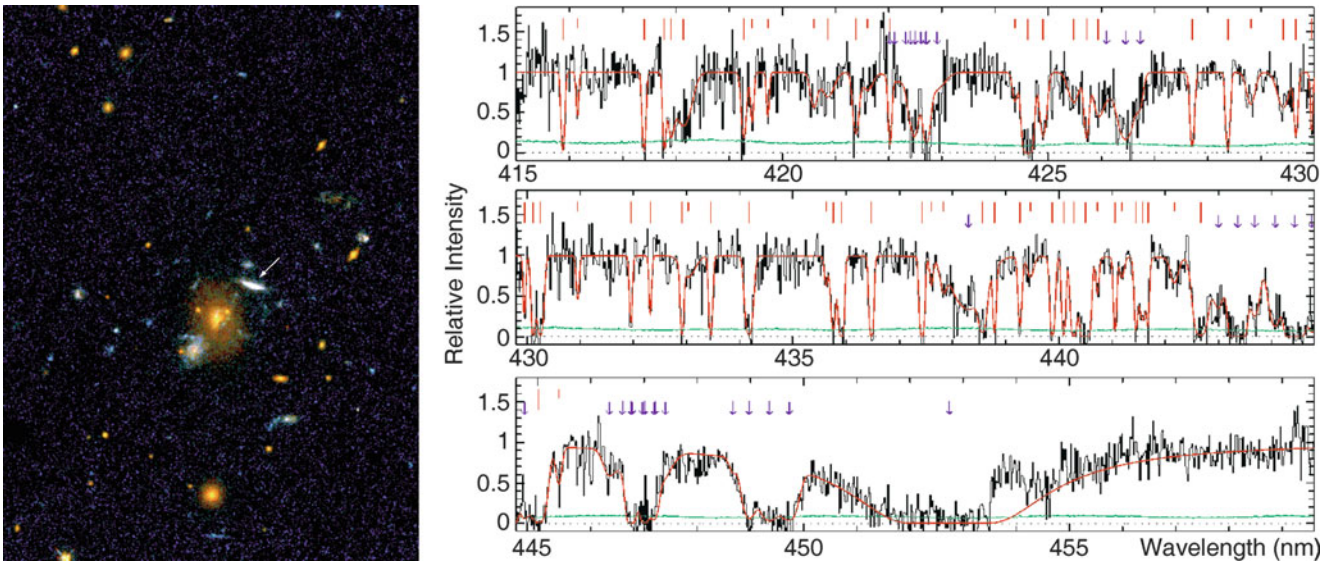


Fig. 9.17 The image *on the left* was taken by the Hubble Space Telescope. It shows the cluster of galaxies MS 1512+36, which has a redshift of $z = 0.37$. To the right, and slightly above the central cluster galaxy, an extended and apparently very blue object is seen, marked by an *arrow*. This source is not physically associated with the cluster but is a background galaxy at a redshift of $z = 2.72$. With this HST image it was proved that this galaxy is strongly lensed by the cluster and, by means of this, magnified by a factor of $\sim$ 30. Due to the magnification, this Lyman-break galaxy is the brightest normal galaxy at redshift $z \sim 3$, a fact that can be profitably used for a detailed spectroscopic analysis. *On the right*, a small section from a high-resolution VLT spectrum of this galaxy is shown. The Lyα transition of the galaxy is located at $\lambda = 4530$ Å, visible as a broad absorption line. Absorption lines at shorter wavelengths originate from the Lyα-forest along the line-of-sight (indicated by *short vertical lines*) or by metal lines from the galaxy itself (indicated by *arrows*). Credit: *Left*: S. Seitz, HST; corresponding research article: S. Seitz et al. 1998, *The $z = 2.72$ galaxy cB58: a gravitational fold arc lensed by the cluster MS1512+36*, MNRAS 298, 945. *Right*: European Southern Observatory

Lyman-break galaxy at $z = 3.07$, shown in Fig. 9.18. The typical magnification of these lensed LBGs is $\mu \sim 30$.

Further examples. Furthermore, at least two highly magnified Lyα emitters (LAEs) at $z \approx 5$ were found. Whereas the study of LAEs is usually hampered by the faint continuum, the strong magnification ($\mu > 10$ in these two systems) enables an investigation of the underlying stellar population from broad-band photometry. One of these two systems has an estimated stellar mass of $\lesssim 10^8 M_\odot$, one of the lowest mass systems found so far at high redshifts. Thanks to the increased effective angular resolution provided by gravitational lensing magnification, a spatially resolved view of the star-forming regions in a $z = 2.33$ sub-millimeter galaxy was obtained. The observations are compatible with the picture that star formation in this object occurs in the cores of giant molecular clouds, as in the local Universe, but these regions are $\sim$ 100 times larger than local star-forming sites.

Apparently extreme sources. One can argue that there is a high probability that the flux of the apparently most luminous sources from a particular source population is magnified by lensing. The apparently most luminous IRAS galaxy, F10214+47, is magnified by a factor $\sim$ 50 by the

Fig. 9.18 The Cosmic Eye is a Lyman-break galaxy at $z = 3.07$, gravitationally lensed by an early-type galaxy at $z = 0.73$, which itself is located in the background of a massive galaxy cluster at $z = 0.33$. The angular extent of the arc systems is $\sim 3''$; this large image splitting is due to the combined lensing effect of the main lens galaxy and the foreground cluster, which magnify the source galaxy by a factor $\mu \sim 30$. Hence, this image is almost 4 magnitudes brighter than the unlensed source. Credit: D.P. Stark, HST

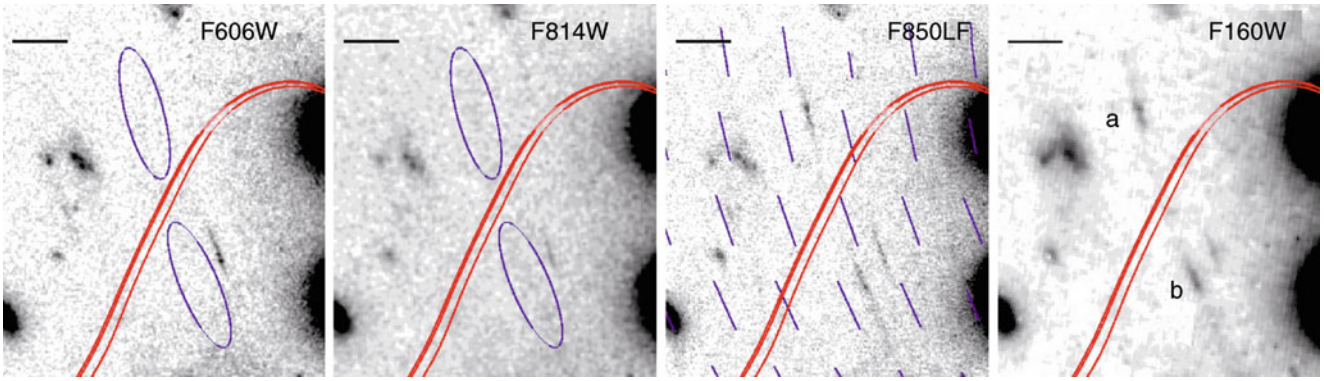


Fig. 9.19 A section of the galaxy cluster Abell 2218 ($z = 0.175$), observed with the HST in four different filters. This region was selected because the magnification by the gravitational lens effect for sources at high redshift is expected to be very large here. This fact has been established by a detailed mass model of this cluster which could be constructed from the geometrical constraints provided by the numerous arcs and multiple images (Fig. 6.51). The *red lines* denote the critical curves of this lens for source redshifts of $z = 5$, 6.5, and 7. A double image of an extended source is clearly visible in the NIR image (*on the right*); this double image was not detected at shorter wavelengths—the expected position is marked by two ellipses in the two images *on the left*. The direction of the local shear, i.e., of the expected image distortion, is plotted in the *second image from the right*; the observed elongation of the two images a and b is compatible with the shear field from the lens model. Together with the photometry of these two images, a redshift between $z = 6.8$ and $z = 7$ was derived for the source of this double image. Source: J.-P. Kneib et al. 2004, *A Probable $z \sim 7$ Galaxy Strongly Lensed by the Rich Cluster A2218: Exploring the Dark Ages*, ApJ 607, 697, p. 698, Fig. 1. ©AAS. Reproduced with permission

lens effect of a foreground galaxy, where the exact value of the magnification depends on the wavelength, since the intrinsic structure and size of the source is wavelength-dependent—hence the magnification is differential. Other examples are the QSOs B1422+231 and APM 08279+5255, which are among the brightest QSOs despite their high redshifts; hence, they belong to the apparently most luminous sources in the Universe. In both cases, multiple images of the QSOs were discovered, verifying the action of the lens effect. Their magnification, and therefore their brightness, renders these sources preferred objects for QSO absorption line spectroscopy (see Fig. 5.55). The Lyman-break galaxy cB58 mentioned previously is another example, and we will see below that the most luminous sub-millimeter sources are gravitationally lensed.

Employing natural telescopes. Most of the examples of highly magnified sources mentioned before were found serendipitously. However, the magnification effect can also be utilized deliberately, by searching for high-redshift sources in regions which are known to produce strong magnification effects, i.e., fields around clusters of galaxies: for a massive cluster, one knows that distant sources located behind the cluster center are substantially magnified. It is therefore not surprising that some of the most distant galaxies known have been detected in systematic searches for drop-out galaxies near the centers of massive clusters. One example of this is shown in Fig. 9.19, where a galaxy at $z \sim 7$ is doubly imaged by the cluster Abell 2218 (see Fig. 6.51), and by means of this it is magnified by a factor ~ 25.

The CLASH and HLS surveys. In order to fully exploit the power of natural telescopes, a large treasury program was carried out with HST, imaging 25 massive clusters with 16 filters of the ACS and WFC3/IR instruments. This CLASH (Cluster Lensing And Supernova survey with Hubble) survey is designed to obtain accurate spectral energy distributions of very faint galaxies in the cluster fields. Furthermore, the data will allow high-quality strong and weak lensing analysis of these clusters, and many results are already available at the time of writing. The observing strategy with multiple visits per cluster field also enables the detection of supernovae, either of cluster galaxies, or gravitationally lensed background supernovae.

In a similar spirit, the Herschel Lensing Survey (HLS) targeted 44 massive clusters, in order to find magnified far-IR sources; some of the results from this survey will be discussed in Sect. 9.3.3 below.

9.2.4 Towards the dark ages

We have seen that in order to apply the Lyman-break technique, one needs images in three filters, one on the blue side of the Lyman break, and two on the red side. These two redder filters are required to demonstrate that these galaxies have a rather flat spectrum for wavelengths longer than the break, i.e., that they are indeed actively star forming, to minimize the possibility that the source is simply a very red one and drops out of the shorter wavelength filter just for this reason. Therefore, the application of the Lyman-break method on deep optical images is restricted to redshifts of

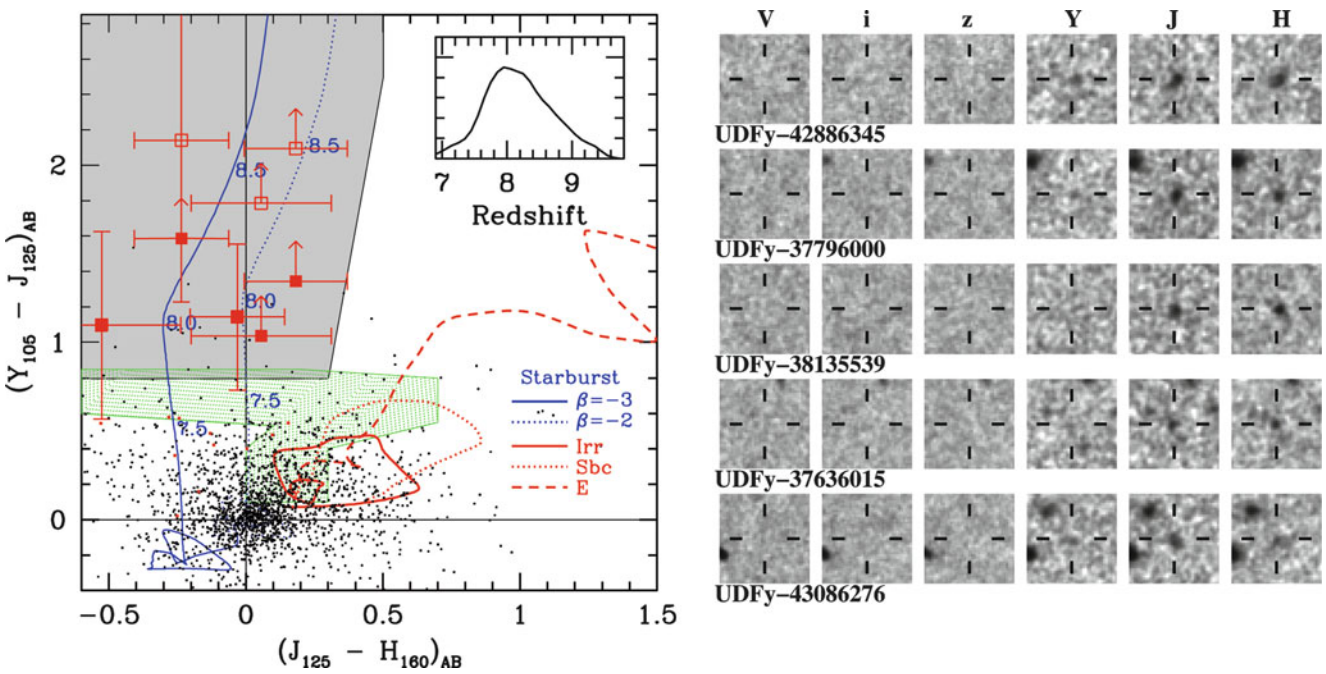


Fig. 9.20 *Left panel*: Similar to Fig. 9.4, evolutionary tracks of stellar populations in a near-IR color-color diagram are shown, for different galaxy types. The two *blue tracks* correspond to starburst galaxies, and they differ in the assumed slope β of the continuum UV spectrum. The corresponding redshifts are written near the tracks. *Black points* show the near-IR colors of galaxies detected at optical wavelengths in the HUDF. The *red squares* with error bars correspond to five galaxies which are clearly detected in at least two near-IR bands, but with no detection in any optical filter—see the *right panel*, where cut-outs around these sources in the HUDF optical and WFC3/IR images are shown. These were selected by the color criteria indicated by the *grey* region. Source: R.J. Bouwens et al. 2010, *Discovery of* $z \sim 8$ *Galaxies in the Hubble Ultra Deep Field from Ultra-Deep WFC3/IR Observations*, ApJ 709, L133, p. L135, Figs. 1, 2, 3. ©AAS. Reproduced with permission

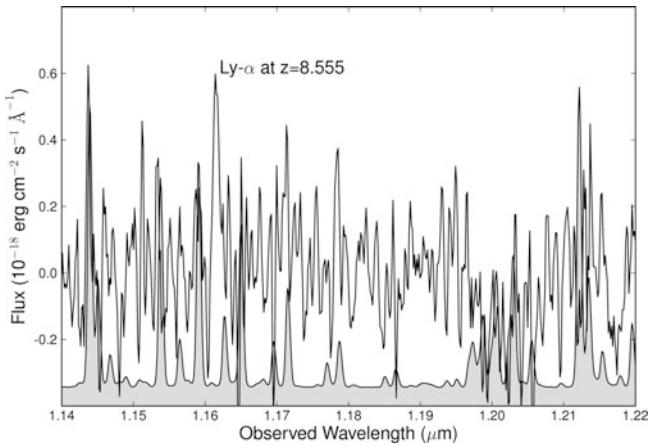


Fig. 9.21 Spectrum of the galaxy UDFy-38135539, for which multi-band images are shown in Fig. 9.20, obtained with integral field spectrograph SINFONI at the VLT. The *grey band* at the bottom shows the shape of the night-sky spectrum. The total integration time for this spectrum was 15 h. The *Ly*α line marked at $\lambda = 11\,616$ Å is broader than the instrumental resolution, and is detected with $\sim 6\sigma$ significance. Source: M.D. Lehnert et al. 2010, *Spectroscopic confirmation of a galaxy at redshift z=8.6*, Nature 467, 940, Fig. 1. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2010

$z \lesssim 5$. In order to move towards higher redshifts, images in the near-IR are required.

Combining the images of the HUDF with deep NIR imaging from NICMOS and with ground-based telescopes, the redshift $z \sim 6$ barrier could be overcome, by searching for i-band dropouts. Many of these were found, and several of the brighter ones were spectroscopically confirmed. Furthermore, ultra-deep optical and NIR imaging from the ground, covering larger sky areas than possible with the HST, have brought large harvest in selecting $z \sim 6$ galaxies using the Lyman-break technique. Currently, more than 30 galaxies are known with spectroscopic redshifts between 6 and 6.5.

As we saw before, QSOs at these redshifts have essentially no flux shortward of the Lyα emission line, i.e., the intergalactic medium at $z \sim 6$ is sufficiently neutral to absorb almost all UV-photons. This could mean that we are approaching the redshift of reionization of the Universe, and it is not immediately obvious that one may expect a substantial population of higher-redshift objects. On the other hand, the photons needed for reionizing the Universe must come from the first galaxies formed, and these must occur at even higher redshifts. In addition, since the first results from WMAP, we have known that the redshift of reionization is substantially higher than $z = 6$, with the Planck CMB results estimating this redshift to be at $z \sim 10$. Thus, searching for higher-redshift sources is a highly valuable scientific aim,

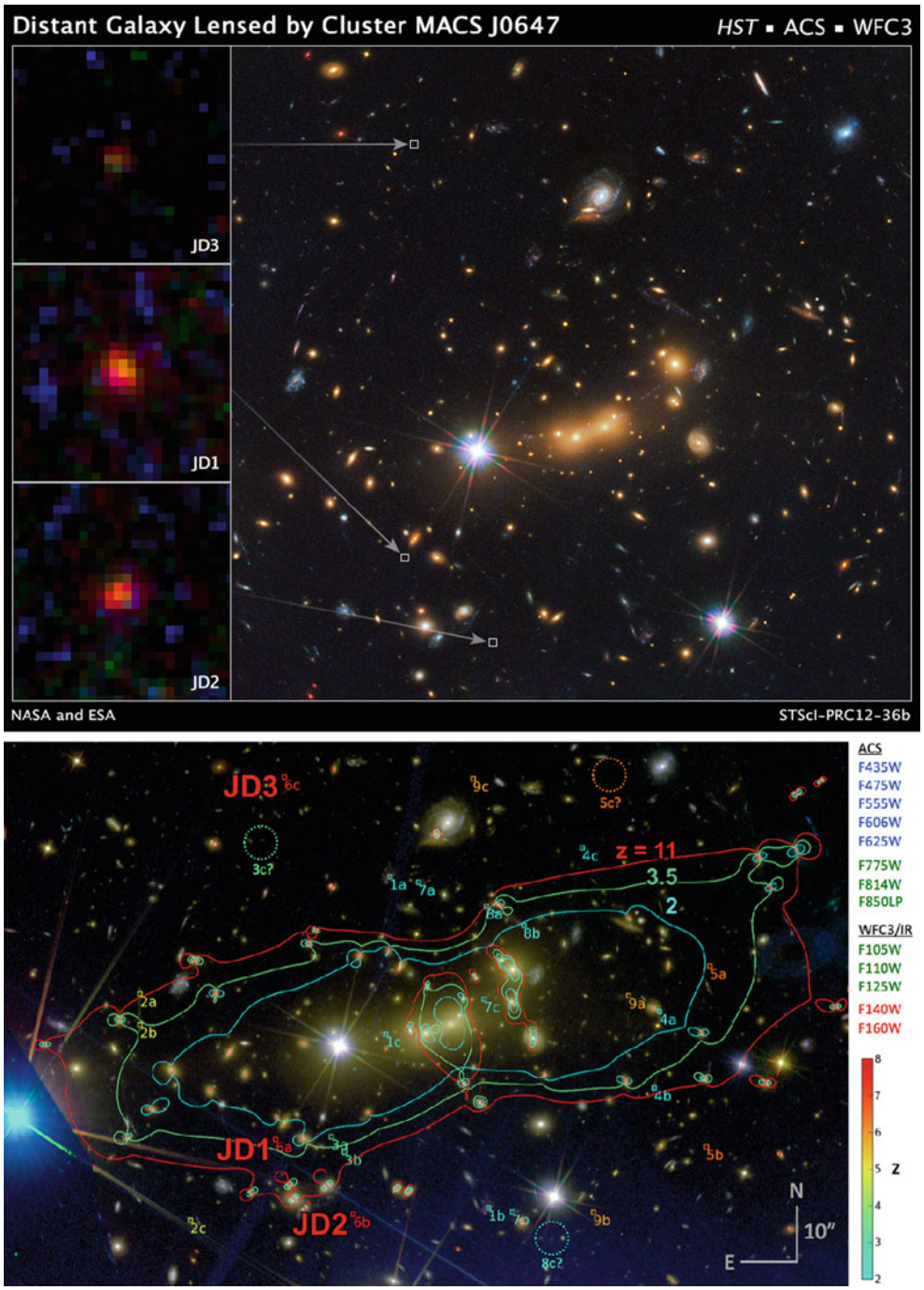


Fig. 9.22 The galaxy cluster MACS J0647+7015 ($z = 0.591$) as observed by the HST in the framework of the CLASH survey. The image is a multi-band composite, based on images in 13 different bands. The three *small panels on the left of the top image* are zooms of a multiply imaged source. The *bottom panel* shows the critical curves of the lens model for this cluster, for sources at different redshifts of $z = 2.0$ (*cyan*), $z = 3.5$ (*green*) and $z = 11.0$ (*red*). Several strongly lensed images are identified and labeled, where the number of the label identify the same source, and the lower-case letter are assigned to the multiple images. From the lens model of the cluster, which yields an estimate of the magnification $\mu \approx 7$ for the brightest image, and the multi-band photometry, a likely redshift of $z \sim 10.7$ is estimated for the source. *Top*: Credit: NASA, ESA, M. Postman and D. Coe (STScI), and the CLASH Team. *Bottom*: Source: D. Coe et al. 2013, *CLASH: Three Strongly Lensed Images of a Candidate $z \approx 11$ Galaxy*, ApJ 762, 32, p 3, Fig. 1. ©AAS. Reproduced with permission

in order to understand the early stages of the evolution of galaxies.

Once at $z \sim 7$, the Lyα line is at $\lambda \sim 1\,\mu$m, and hence one needs at least two NIR images, plus a very deep optical image in the z-band, in order to find LBG candidates at these redshifts. As discussed above, the latest camera onboard HST, the WFC3, has a sensitive near-IR channel, and its mapping of the HUDF provided an excellent data set for this purpose. Together with less deep, but larger-area imaging by WFC3 over the GOODS field, some 100 candidates at $z \sim 7$, and more than 50 at $z \sim 8$ have been found (see Fig. 9.20 for several $z \sim 8$ candidates). Spectroscopic verification of these sources, however, is very challenging, since now very sensitive near-IR spectroscopy is needed.

For one of the high-z candidates in Fig. 9.20, the detection of a spectral line in a deep NIR spectrum was claimed—see Fig. 9.21—yielding a redshift of $z = 8.55$. Whereas the redshift is based on a single emission line, the identification

of this line as Lyα is considered to be secure, since the HST photometry shown in Fig. 9.20 rules out all other redshift, except a low-probability alternative at $z \sim 2$. In this case, the emission line would be [OII], which is a doublet. This doublet would have been clearly resolved in the spectrum and can safely be ruled out. Thus, if the emission line in Fig. 9.21 is real, this galaxy is the highest-redshift spectroscopically confirmed galaxy yet found. However, reobservation of the same source with two other NIR spectrographs failed to reproduce the emission line. Whether or not this galaxy is at $z = 8.55$ thus remains an open issue, but this example impressively illustrates the difficulty of securing redshifts for $z \gtrsim 7$ galaxies. We also need to keep in mind that many LBGs at lower redshift do not show a Lyα emission line; no spectroscopic redshift of analogs of such sources at high redshifts can be obtained with current telescopes.

Employing natural telescopes, the hunt for even higher-redshift sources becomes promising. The cluster MACS J0647+7015 shown in Fig. 9.22 is a target of the CLASH survey, and thus has been imaged with HST in many different filters. One of the multiply-imaged sources in this cluster is a J-band drop-out; its spectral energy distribution puts it at a redshift $z \sim 11$. Such a high redshift is also supported by the lensing geometry and the location of the three images. Whereas no spectroscopic confirmation of this high redshift is available up to now, the next generation space telescope JWST (see Chap. 11) will probably be able to confirm (or not) this redshift estimate.

On the other hand, the narrow-band selection of high-z sources is more promising in this respect. Although there are many contaminating sources—emission line objects at lower redshifts—a clear detection of a source through the narrow-band technique at least yields a good indication that the source has an emission line at that wavelength, and thus spectroscopy is promising.

The wavelengths of the narrow-band filters are best chosen as to minimize the sky brightness. This then defines preferred spectral windows, which in turn define the redshifts of the Lyα emitting galaxies that can be detected with this technique. By far the most productive telescope in this respect is Subaru, due to its unique combination of aperture and field-of-view of its Suprime-Cam camera. Specifically designed narrow-band filters for the highest redshifts target LAEs at $z \sim 5.7$, 6.6, 7.0 and 7.3. Many hundreds of LAEs were found with this techniques, including several dozens spectroscopically confirmed at redshift $z \gtrsim 6.6$. The redshift record holder are galaxies at $z = 6.96$ and $z = 7.213$.

In many cases photometric redshifts may be the only method for obtaining redshift information, until more powerful telescopes come into operation. As mentioned before, photometric redshifts are the more reliable the more bands are available, and the better the photometric accuracy is. Deep fields like the GOODS field are therefore best suited for photo-z studies of high-redshift galaxies, owing to the broad wavelength coverage and their depth. In Fig. 9.23, two galaxies selected by the Lyman-break technique are shown, with optical, NIR and MIR observations available from HST and Spitzer. In both cases, the best-fitting high-redshift and low-redshift spectral energy distributions are shown, and in both cases, the low-redshift fit is very much worse than the high-z one, so that a low redshift can be ruled out with very high confidence in both cases. This photo-z technique has

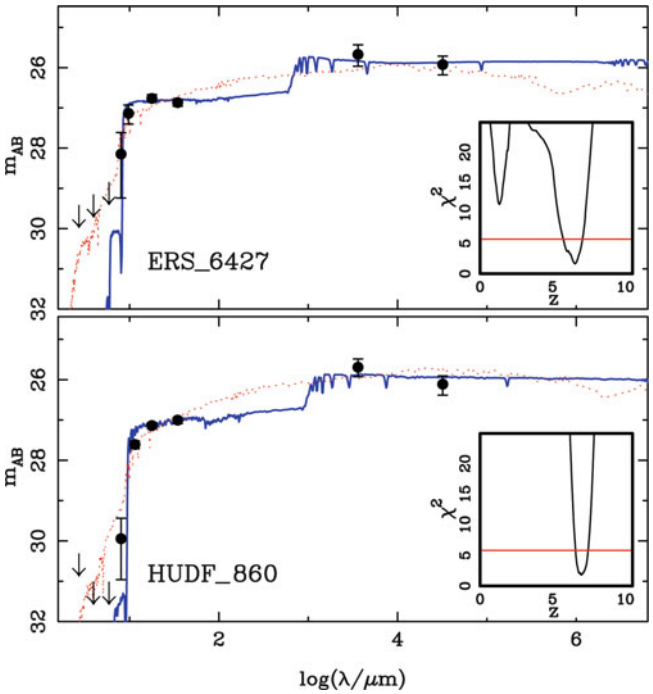


Fig. 9.23 Two high-redshift galaxy candidates, selected by the Lyman-break technique from deep HST fields. The data points with error bars are flux measurements from the HST data, complemented by MIR data from Spitzer. The *small arrows* indicated upper limits to the flux in these bands. The *inserts* show the figure-of-merit function χ^2 as a function of redshift. The *thick blue line* in each panel shows the best fitting spectral energy distribution for a high-redshift galaxy, yielding $z \sim 6.65$ and $z = 6.96$ for the upper and lower galaxy, respectively, corresponding to the minimum of the χ^2 function. A second spectral fit can be found when assuming a lower redshift, shown by the *dotted red curves*. In the *upper panel*, this fit is indeed almost acceptable, though the high-z fit is considerably better—the corresponding χ^2-values as a function of redshift is shown in the *inset*, and rules out the low-redshift solution with very high confidence. The lower-redshift fit shown in the *bottom panel* can be rejected fully; this $z \sim 7$ estimate seems to be very robust. Source: R.J. McLure et al. 2011, *A robust sample of galaxies at redshifts* $6.0 < z < 8.7$*: stellar populations, star formation rates and stellar masses*, MNRAS 418, 2074, Fig. 2. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

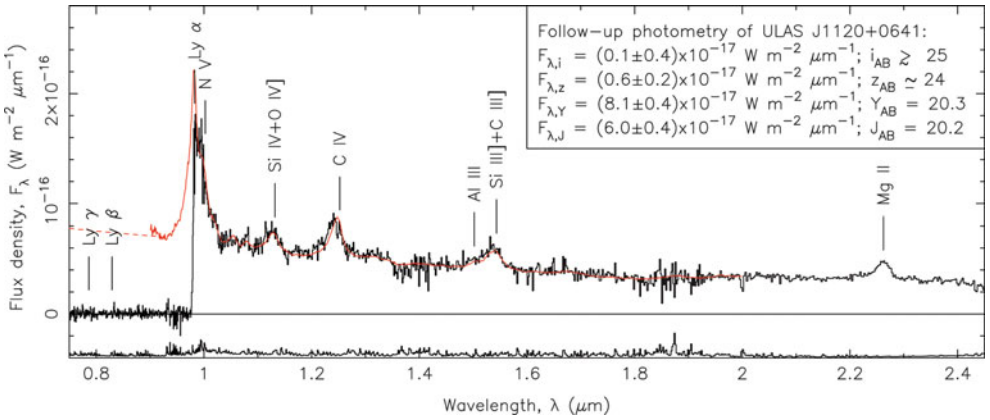


Fig. 9.24 The spectrum of the QSO ULAS J1120+0641 at $z = 7.085$ is shown in *black*, superposed on a composite spectrum of lower-redshift QSOs. Several emission lines redwards of the Lyman-α line are marked. The two spectra are in very good agreement, except of course for rest wavelengths smaller than the Lyman-α line, due to exceedingly strong absorption by the Lyα forest. However, the CIV line of ULAS J1120+0641 seems to be significantly blueshifted, relative to the composite spectrum. Source: D.J. Mortlock et al. 2011, *A luminous quasar at a redshift of z = 7.085*, Nature 474, 616, Fig. 1. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2011

yielded substantial samples of $z \lesssim 8$ galaxies, which form a rather robust base for statistical studies of the high-z galaxy population.

The highest redshift QSO. Whereas over most of the past $\sim$ 50 years QSOs were the redshift record holders, this is currently no longer the case: independent of whether the galaxy shown in Fig. 9.21 is indeed at $z = 8.55$, the photo-z of several galaxies are sufficiently robust to place them at $z > 7$. Concerning QSOs, the SDSS has discovered several $z \sim 6$ objects, with the highest redshift one at $z = 6.42$ (see Fig. 9.1). This was for many years the record holder; only recently, a QSO with $z > 7$ was found, whose spectrum is shown in Fig. 9.24. It was found with a color selection, based on optical and near-IR photometry. The spectrum clearly confirms its high redshift of $z = 7.085$, based on several emission lines. Its high luminosity of $\sim 6 \times 10^{13} L_\odot$ implies a very massive black hole with mass $M \sim 2 \times 10^9 M_\odot$, as estimated from the line width of the MgII emission line in combination with the luminosity. This large SMBH mass had to be assembled within the first 800 million years of the Universe. This strengthens the constraints on rapid black hole formation, relative to the previously discovered QSOs with redshifts $z \lesssim 6.4$.[4]

[4] At the time of writing, $\sim$ 90 QSOs with $z > 5.7$ and known, of which $\sim$ 40 have $z > 6.0$; those were found from several wide-field imaging surveys, including SDSS, the CFHT quasar survey, the UKIRT Infrared Deep Sky Survey (UKIDSS), and Pan-STARRS.

9.3 New types of galaxies

The Lyman-break galaxies discussed above are not the only galaxies that are expected to exist at high redshifts. We have argued that LBGs are galaxies with strong star formation. Moreover, the UV radiation from their newly-born hot stars must be able to escape from the galaxies. From observations in the local Universe we know, however, that a large fraction of star formation is hidden from our direct view when the star formation region is enveloped by dust. The latter is heated by absorbing the UV radiation, and re-emits this energy in the form of thermal radiation in the FIR domain of the spectrum. At high redshifts such galaxies would certainly not be detected by the Lyman-break method.

Instrumental developments opened up new wavelength regimes which yield access to other types of galaxies. Two of these will be described in more detail here: EROs (Extremely Red Objects) and sub-millimeter sources. But before we discuss these objects, we will first investigate starburst galaxies in the relatively local Universe.

9.3.1 Starburst galaxies

One class of galaxies, the so-called *starburst galaxies*, is characterized by a strongly enhanced star-formation rate, compared to normal galaxies. Whereas our Milky Way is forming stars with a rate of $\sim 3 M_\odot$/yr, the star-formation rate in starburst galaxies can be larger by a factor of more than a hundred. Dust heated by hot stars radiates in the FIR, rendering starbursts very strong FIR emitters. Many of them

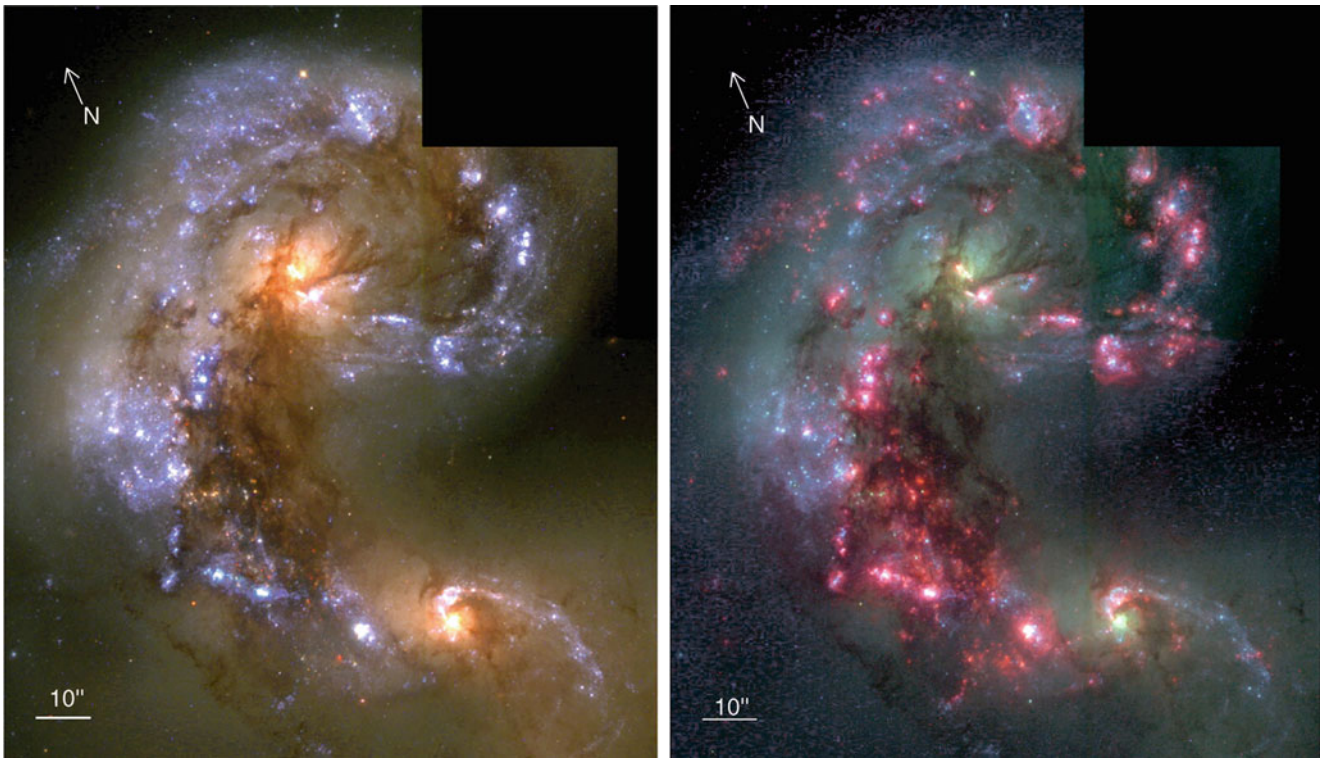


Fig. 9.25 The Antennae galaxies. *On the left*, the 'true' optical colors are shown, whereas in the *right-hand image* the *reddish color* shows Hα emission. This pair of merging galaxies (also see Figs. 1.16 and 3.6 for other examples of merging galaxies) is forming an enormous number of young stars. Both the UV emission (*bluish in the left image*) and the Hα radiation (*reddish in the right image*) are considered indicators of star formation. The individual knots of bright emission are not single stars but star clusters with typically $10^5 M_\odot$; however, it is also possible to resolve individual stars (red and blue supergiants) in these galaxies. Source: B.C. Whitmore et al. 1999, *The Luminosity Function of Young Star Clusters in "the Antennae" Galaxies (NGC 4038-4039)*, AJ 118, 1551, p. 1556, 1557, Figs. 3, 4. ©AAS. Reproduced with permission

were discovered by the IRAS satellite ('IRAS galaxies'); they are also called ULIRGs (Ultra Luminous InfraRed Galaxies).

The reason for this strongly enhanced star formation is presumably the interaction with other galaxies or the result of merger processes, an impressive example of which is the merging galaxy pair known as the 'Antennae' (see Fig. 9.25). In this system, stars and star clusters are currently being produced in very large numbers. The images show a large number of star clusters with a characteristic mass of $10^5 M_\odot$, some of which are spatially resolved by the HST. Furthermore, particularly luminous individual stars (supergiants) are also identified. The ages of the stars and star clusters span a wide range and depend on the position within the galaxies. For instance, the age of the predominant population is about 5–10 Myr, with a tendency for the youngest stars to be located in the vicinity of strong dust absorption. However, stellar populations with an age of 100 and 500 Myr, respectively, have also been discovered; the latter presumably originates from the time of the first encounter of these two galaxies, which then led to the ejection of the tidal tails. This seems to be a common phenomenon; for example, in the starburst galaxy Arp 220 (see Fig. 1.15) one also finds star clusters of a young population with age $\lesssim 10^7$ yr, as well as older ones with age $\sim 3 \times 10^8$ yr. It thus seems that during the merging process several massive bursts of star-cluster formation are triggered.

It was shown by the ISO satellite that the most active regions of star formation are not visible on optical images, since they are completely enshrouded by dust (left panel in Fig. 9.26). A map at 8 μm obtained by the Spitzer Space Observatory (Fig. 9.26, right panel) shows the hot dust heated by young stars, where this IR emission is clearly anti-correlated with the optical radiation. Indeed, the mid-infrared emission is strongest in the region between these two colliding galaxies; apparently, this is the location where the current star formation is most intense. Maps of the CO emission also show that these regions contain a large reservoir of molecular gas. Furthermore, the large-scale X-ray radiation in this galaxy collision shows the efficiency with which gas is heated to high temperatures by the supernova events that accompany massive star formation, with a corresponding chemical enrichment of the gas with α-elements. Many of the X-ray point sources are due to high-mass X-ray binaries.

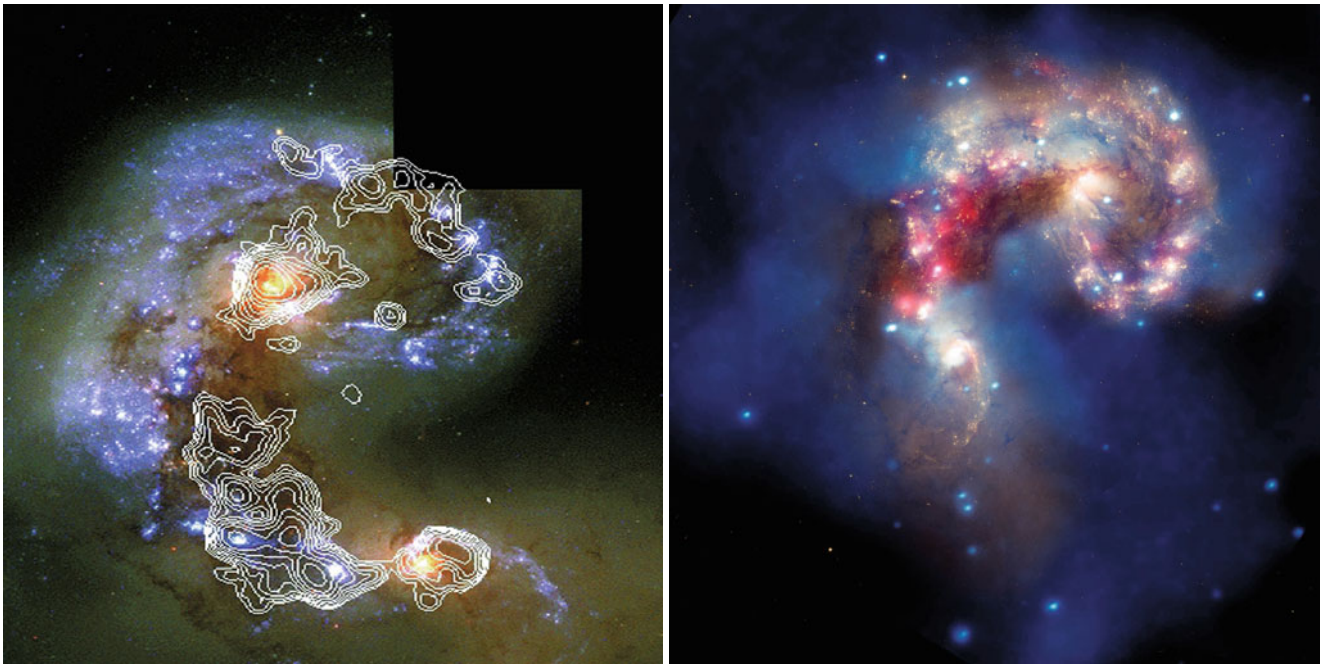


Fig. 9.26 The Antennae galaxies: In the *left panel*, superposed on the optical HST image are contours of infrared emission at 15 μm, measured by the ISO satellite. The *right panel* shows a composite: X-ray emission as observed by Chandra shown in *blue*, and the 8 μ*m* map obtained by Spitzer (*red*), is superposed on an optical three-band composite image from HST. The strongest IR emission originates in optically dark regions. A large fraction of the star formation in this galaxy pair is not visible on optical images, because it is hidden by dust absorption. Note that the orientation in the *right panel* differs from that of the other images of the Antennae, as it was taken with the HST at a different orientation angle. Credit: *Left*: *Canadian Astronomy and Astrophysics in the 21st Century*, Courtesy Christine Wilson, McMaster University. *Right*: NASA/CXC/SAO/JPL-Caltech/STScI

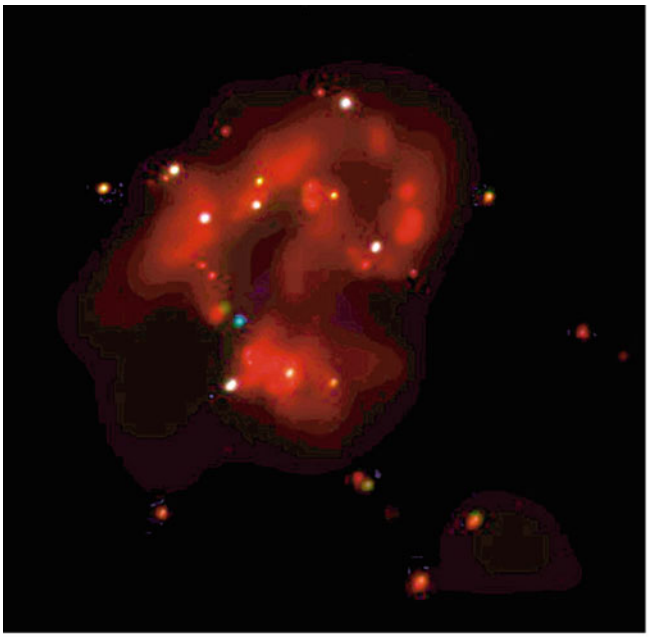
Obviously, a complete picture of star formation in such galaxies can only be obtained from a combination of optical and IR images.

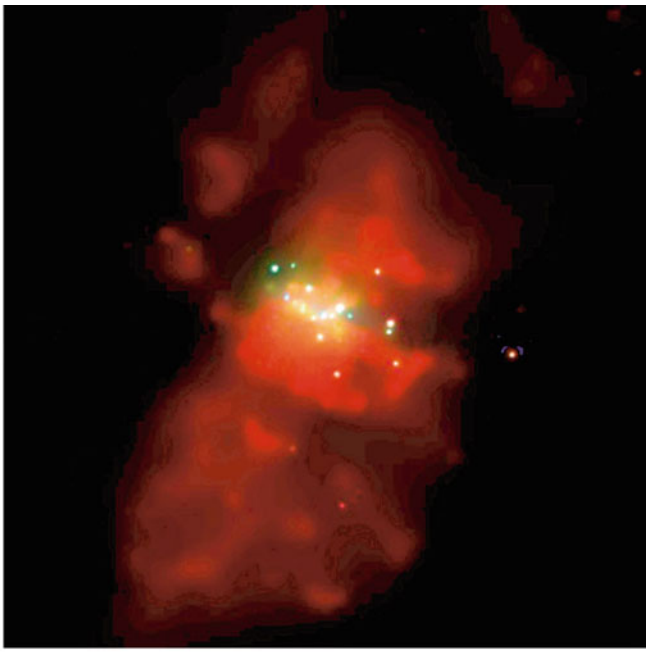
Combining deep optical and NIR photometry with MIR imaging from the Spitzer telescope, star-forming galaxies at high redshifts can be detected even if they contain an appreciable amount of dust (and thus may fail to satisfy the LBG selection criteria). These studies find that the comoving number density of ULIRGs with $L_{\rm IR} \gtrsim 10^{12} L_\odot$ at $z \sim 2$ is about three orders of magnitude larger than the local ULIRG density. These results seem to imply that the high-mass tail of the local galaxy population with $M \gtrsim 10^{11} M_\odot$ was largely in place at redshift $z \sim 1.5$ and evolves passively from there on. We shall come back to this aspect below.

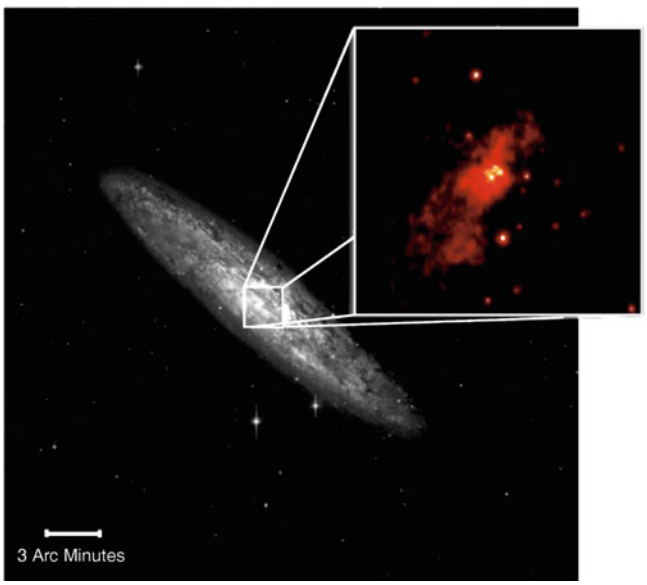
Ultra-luminous Compact X-ray Sources. Observations with the Chandra satellite have shown that starburst galaxies contain a rich population of very luminous compact X-ray sources (Ultra-luminous Compact X-ray Sources, or ULXs; see Fig. 9.27). They are formally defined to have an X-ray luminosity $> 10^{39}$ erg/s. Similar sources, though with lower luminosity, are also detected in the Milky Way, where these are binary systems with one component being a compact star (white dwarf, neutron star, or black hole). The X-ray emission is caused by accretion of matter (which we discussed in Sect. 5.3.2) from the companion star onto the compact component, and are called X-ray binaries.

Some of the ULXs in starbursts are so luminous, however, that the required mass of the compact star by far exceeds $1\,M_\odot$ if the Eddington luminosity is assumed as an upper limit for the luminosity [see (5.25)]. The detection of these ULXs in the 1980s by the Einstein observatory thus came unexpectedly, since one does not expect to form black holes with a mass larger than $\sim 10 M_\odot$ in supernova explosions. Thus, one concludes that either the emission of these sources is highly anisotropic, hence beamed towards us, or that the sources are black holes with masses of up to $\sim 200 M_\odot$. In the latter case, we may just be witnessing the formation of intermediate mass black holes in these starbursts. In fact, recently a ULX was found with a peak luminosity of 10^{42} erg/s, which, assuming that the Eddington limit is not exceeded, implies a black hole mass of $>500 M_\odot$ as the origin of this source, located ~ 4 kpc away from the center of an edge-on spiral galaxy.

This latter interpretation is also supported by the fact that the ULXs are concentrated towards the center of the galaxies—hence, these black holes may spiral into the galaxy's center by dynamical friction, and there merge to a SMBH. This is one of the possible scenarios for the







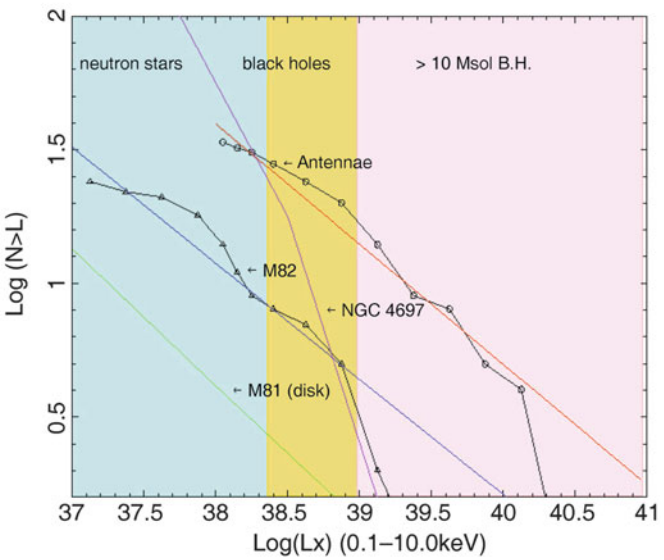


Fig. 9.27 Ultra-luminous compact X-ray sources (ULXs) in starburst galaxies. *Upper left*: The discrete X-ray sources in the Antenna galaxies; the size of the image is $4' \times 4'$. *Lower left*: Optical (image) and (inlaid) Chandra image of the starburst galaxy NGC 253. Four of the ULXs are located within one kiloparsec from the center of the galaxy. The X-ray image is $2'.2 \times 2'.2$. *Upper right*: $5' \times 5'$ Chandra image of the starburst galaxy M82; the diffuse radiation (*red*) is emitted by gas at $T \sim 10^6$ K which is heated by the starburst and flows out from the central region of the galaxy. It is supposed that M82 had a collision with its companion M81 (see Fig. 6.9) within the last 10^8 yr, by which the starburst was triggered. *Lower right*: The luminosity function of the ULXs in some starburst galaxies. The differently shaded regions indicate ranges in luminosity which correspond to Eddington luminosities of neutron stars, 'normal' stellar mass black holes, and black holes with $M \geq 10 M_\odot$. Credit: *Upper left*: NASA/SAO/CXC/G.Fabbiano et al. Bottom left: X-ray: NASA/SAO/CXC, optical: ESO. *Upper right*: NASA/SAO/G.Fabbiano et al., *Bottom right*: NASA/SAO/G.Fabbiano et al.

formation of SMBHs in the cores of galaxies, a subject to which we will return in Sect. 10.4.5. Furthermore, the similarity of the properties of ULXs with those of X-ray binaries may indicate that ULXs are just scaled-up version of these more common sources.

9.3.2 Extremely Red Objects (EROs)

As mentioned several times previously, the population of galaxies detected in a survey depends on the selection criteria. Thus, employing the Lyman-break method, one mainly

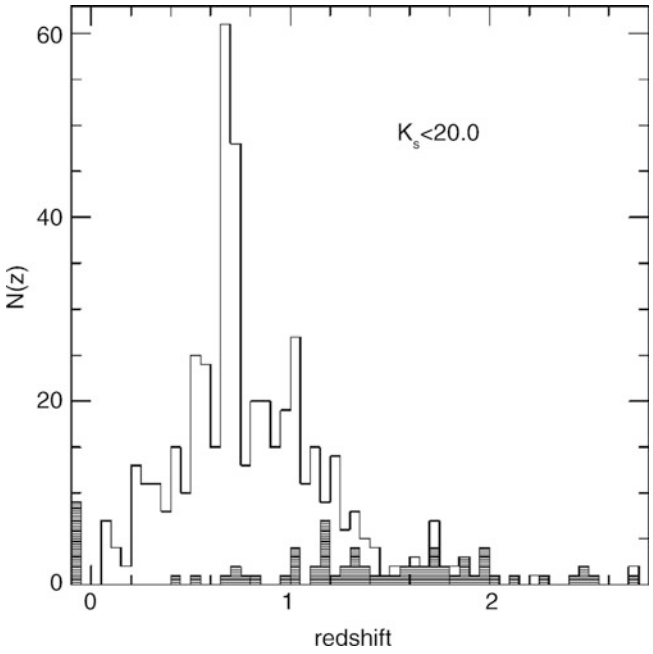


Fig. 9.28 Redshift distribution of galaxies with $K_s < 20$, as measured in the K20-survey. The *shaded histogram* represents galaxies for which the redshift was determined solely by photometric methods. The bin at $z < 0$ contains those nine galaxies for which it has not been possible to determine z. The peak at $z \sim 0.7$ is produced by two clusters of galaxies in the fields of the K20-survey. Source: A. Cimatti et al. 2002, *The K20 survey. IV. The redshift distribution of $K_s < 20$ galaxies: A test of galaxy formation models*, A&A 391, L1, p. L2, Fig. 1. ©ESO. Reproduced with permission

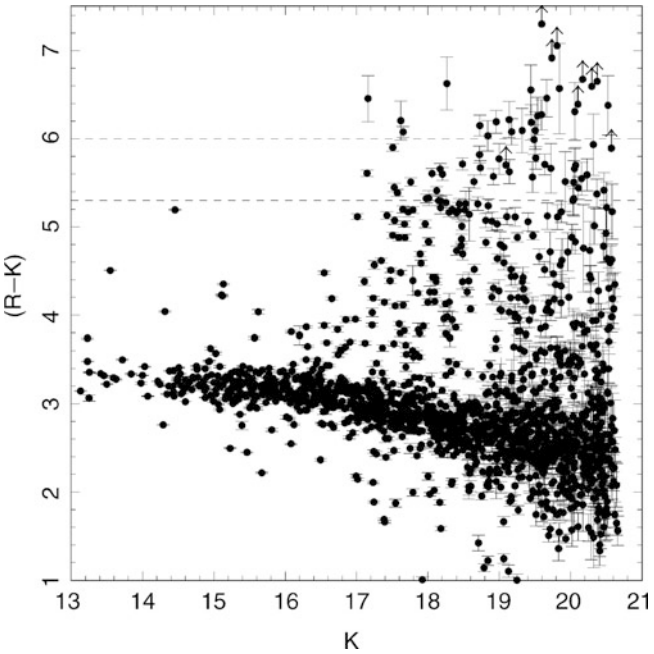


Fig. 9.29 Color-magnitude diagram, i.e., $R-K$ as a function of K, for sources in ten fields around clusters of galaxies. Most of the objects, in particular at the bright end, are located on the red sequence of early-type galaxies. We see that for faint magnitudes (roughly $K \geq 19$), a population of sources with a very red color (about $R - K \geq 5.3$) turns up. These objects are called EROs. Source: G.P. Smith 2002, *A Hubble Space Telescope Survey of X-ray Luminous Galaxy Clusters: Gravitationally Lensed Arcs and EROs*, astro-ph/0201236, Fig. 2. Reproduced by permission of the author

discovers those galaxies at high redshift which feature active star formation and therefore have a blue spectral distribution at wavelengths longwards of Lyα. The development of NIR detectors enabled the search for galaxies at longer wavelengths. Of particular interest here are surveys of galaxies in the K-band, the longest wavelength window that is reasonably accessible from the ground (with the exception of the sub-millimeter to radio domain).

The NIR waveband is of particular interest because the luminosity of galaxies at these wavelengths is not dominated by young stars. As we have seen in Fig. 3.34, the luminosity in the K-band depends rather weakly on the age of the stellar population, so that it provides a reliable measure of the total stellar mass of a galaxy.

Characteristics of EROs. Examining galaxies with a low K-band flux, one finds either galaxies with low stellar mass at low redshifts, or galaxies at high redshift with high optical (due to redshift) luminosity. But since the luminosity function of galaxies is relatively flat for $L \lesssim L_*$, one expects the latter to dominate the surveys, due to the larger volume at higher z. In fact, K-band surveys detect galaxies with a broad redshift distribution. In Fig. 9.28 the z-distribution of galaxies in the K20-survey is shown. In this survey, objects with $K_s < 20$ were selected in two fields with a combined area of 52 arcmin^2, where K_s is a filter at a wavelength slightly shorter than the classic K-band filter. After excluding stars and Type 1 AGNs, 489 galaxies were found, 480 of which have had their redshifts determined. The median redshift in this survey is $z \approx 0.8$.

Considering galaxies in a $(R-K)$ vs. K color-magnitude diagram (Fig. 9.29), one can identify a population of particularly red galaxies, thus those with a large $R - K$. These objects were named *Extremely Red Objects (EROs)*; about 10 % of the galaxies in K-selected surveys at faint magnitudes are EROs, typically defined by $R - K > 5$. Spectroscopic analysis of these galaxies poses a big challenge because an object with $K = 20$ and $R - K > 5$ necessarily has $R > 25$, i.e., it is extremely faint in the optical domain of the spectrum. However, with the advent of 10-m class telescopes, spectroscopy of these objects has become possible.

The nature of EROs: passive ellipticals versus dusty starbursts. From these spectroscopic results, it was found that the class of EROs contains rather different kinds of sources. To understand this point we will first consider the possible explanations for a galaxy with such a red spectral distribution. As a first option, the object may be an old elliptical galaxy with the 4000 Å-break redshifted to the red

side of the R-band filter, i.e., typically an elliptical galaxy at $z \gtrsim 0.8$. For these galaxies to be sufficiently red to satisfy the selection criterion for EROs, they need to already contain an old stellar population by this redshift, which implies a very high redshift for the star formation episode in these objects; it is estimated from population synthesis models that their formation redshift must be $z_{\rm form} \gtrsim 2.5$. A second possible explanation for large $R - K$ is reddening by dust. Such EROs may be galaxies with active star formation where the optical light is strongly attenuated by dust extinction. If these galaxies are located at a redshift of $z \sim 1$, the measured R-band flux corresponds to a rest-frame emission in the UV region of the spectrum where extinction is very efficient.

Spectroscopic analysis reveals that both types of EROs are roughly equally abundant. Hence, about half of the EROs are elliptical galaxies that already have, at $z \sim 1$, a luminosity similar to that of today's ellipticals, and are at that epoch already dominated by an old stellar population. The other half are galaxies with active star formation which do not show a (prominent) 4000 Å-break but which feature the emission line of [OII] at $\lambda = 3727$ Å, a clear sign of recent star formation. Further analysis of EROs by means of very deep radio observations confirms the large fraction of galaxies with high star-formation rates. Utilizing the close relation of radio emissivity and FIR luminosity, we find a considerable fraction of EROs to be ULIRGs at $z \sim 1$.

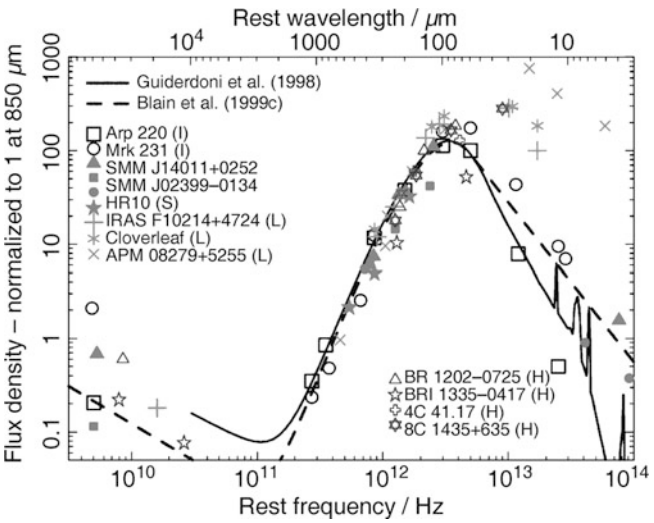


Fig. 9.30 Spectral energy distribution of some dusty galaxies with known redshift z (*symbols*), together with two model spectra (*curves*). The spectral distributions for all sources are normalized to unity at $\lambda = 850\,\mu$m. Four types of galaxies are distinguished: (I) IRAS galaxies at low z; (S) luminous sub-mm galaxies; (L) distant sources that are magnified by the gravitational lens effect and multiply imaged; (H) AGNs. Only a few sources among the lens systems (presumably due to differential magnification) and the AGNs deviate significantly from the model spectra. Source: A. Blain et al. 1999, *Submillimeter-selected galaxies*, astro-ph/9908111, Fig. 3. Reproduced by permission of the author

Spatial correlations. EROs are very strongly correlated in space. The interpretation of this strong correlation may be different for the passive ellipticals and for those with active star formation. In the former case the correlation is compatible with a picture in which these EROs are contained in clusters of galaxies or in overdense regions that will collapse to a cluster in the future. The correlation of the EROs featuring active star formation can probably not be explained by cluster membership, but the origin of the correlation may be the same as for the correlation of the LBGs.

The number density of passive EROs, thus of old ellipticals, is surprisingly large compared with expectations from the model of hierarchical structure formation that we will discuss in Chap. 10.

9.3.3 Dusty star-forming galaxies

FIR emission from hot dust is one of the best indicators of star formation. However, observations in this waveband are only possible from space, such as was done with the IRAS and ISO satellites, and more recently with Spitzer and Herschel. Depending on the dust temperature, dust emission has its maximum at about 100 μm, which is not observable from the ground. At longer wavelengths there are spectral windows longwards of $\lambda \sim 250\,\mu$m where observations through the Earth's atmosphere are possible, for instance at 450 and 850 μm in the sub-millimeter waveband. However, the observing conditions at these wavelengths are extremely dependent on the amount of water vapor in the atmosphere, so that the observing sites must by dry and at high elevations. In the sub-millimeter (sub-mm) range, the long wavelength domain of thermal dust radiation can be observed, which is illustrated in Fig. 9.30.

Developments. Since about 1998 sub-mm astronomy has experienced an enormous boom, with two instruments having been put into operation: the Sub-millimeter Common User Bolometer Array (SCUBA), operating at 450 and 850 μm, with a field-of-view of 5 arcmin^2, and the Max-Planck Millimeter Bolometer (MAMBO), operating at 1200 μm. Both are bolometer arrays which initially had 37 bolometers each, but which later were upgraded to a considerably larger number of bolometers. With the opening of the 12-m APEX (Atacama Pathfinder Experiment) telescope (see Fig. 1.28) in Chile, equipped with powerful instrumentation, a further big step in sub-mm astronomy was achieved. In addition, the Herschel satellite (Fig. 1.33), operating at wavelength between 55 and 670 μm, allowed imaging of large sky areas at these wavelengths, due to a much lower noise level than can be achieved from the ground, though with considerably worse spatial resolution

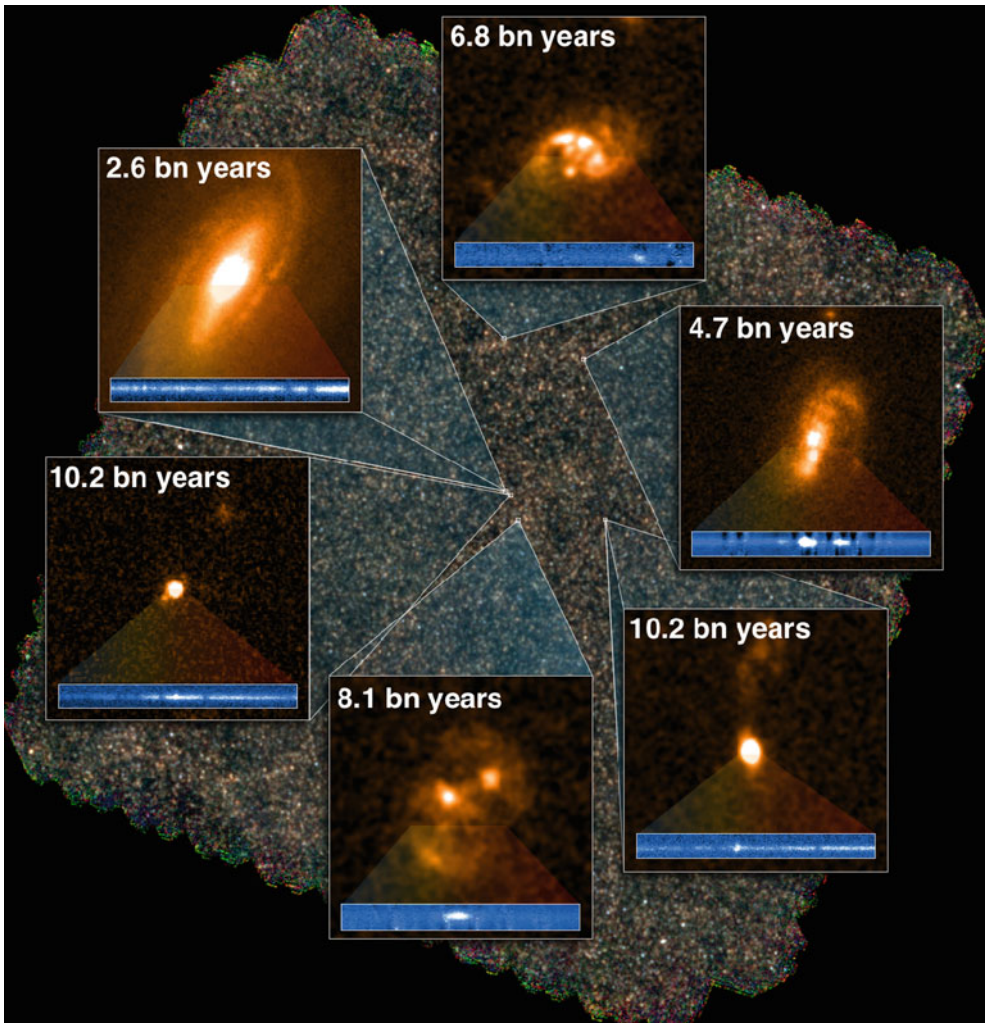


Fig. 9.31 A sub-millimeter image of the COSMOS field, taken with the SPIRE instrument on Herschel. The image is about 2 degrees on the side and is a color composite of three bands with wavelength 250, 350 and 500 μm. For six galaxies in the field, the zooms show an optical images, with the cosmic look-back time indicated as obtained by spectroscopy with the W.M. Keck telescopes. Credit: COSMOS field: ESA/Herschel/SPIRE/HerMES Key Programme; Hubble images: NASA, ESA; Keck Spectra: Caltech/W. M. Keck Observatory

due to the smaller aperture compared to ground-based sub-mm telescopes. Figure 9.31 shows a Herschel image of the COSMOS field.

Apart from single-dish observatories, interferometers operating at these wavelengths yield substantially higher angular resolution, for example the very successful IRAM Plateau de Bure interferometer in the French Alps consisting of six 15-m antennas. The Atacama Large Millimeter Array (ALMA; see Fig. 1.29) with its 54 12-m and 12 7-m antennas, inaugurated in 2013, marks a huge leap in terms of resolution and sensitivity in this waveband regime.

The negative K-correction of sub-mm sources. The emission of dust at these wavelengths is described by a Rayleigh–Jeans spectrum, modified by an emissivity function that depends on the dust properties (chemical composition, distribution of dust grain sizes); typically, one finds

$$S_\nu \propto \nu^{2+\beta} \quad \text{with} \quad \beta \sim 1 \ldots 2 \,.$$

This steep spectrum for frequencies below the peak of the thermal dust emission at $\lambda \sim 100\,\mu\text{m}$ implies a very strong negative K-correction (see Sect. 5.6.1) for wavelengths in the sub-mm domain: at a fixed observed wavelength, the rest-frame wavelength becomes increasingly smaller for sources at higher redshift, and there the emissivity is larger. As Fig. 9.32 demonstrates, this spectral behavior causes the effect that the flux in the sub-mm range does not necessarily decrease with redshift. For $z \lesssim 1$, the $1/D^2$-dependence of the flux dominates, so that up to $z \sim 1$ sources at fixed luminosity get fainter with increasing z. However, between $z \sim 1$ and $z \sim z_{\text{flat}}$ the sub-mm flux as a function of redshift remains nearly constant or even increases with z, where z_{flat} depends on the dust temperature T_d and the observed wavelength; for $T_\text{d} \sim 40\,\text{K}$ and $\lambda \sim 850\,\mu\text{m}$ one finds $z_{\text{flat}} \sim 8$. We therefore have the quite amazing situation that sources appear brighter when they are at larger distances. This is caused by the very negative K-correction which more than compensates for the $1/D^2$-decrease of the flux. Only for $z > z_{\text{flat}}$ does the flux begin to rapidly decrease with redshift, since then, due to redshift, the corresponding

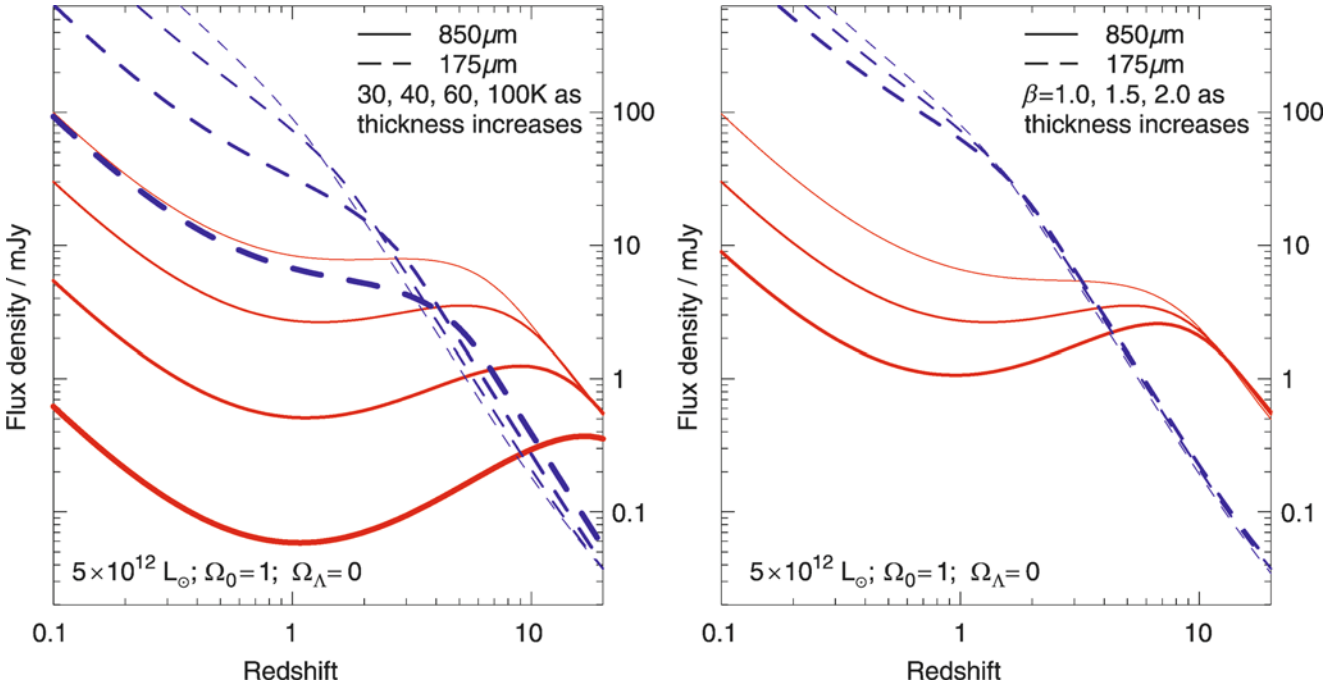


Fig. 9.32 Predicted flux from dusty galaxies as a function of redshift. The bolometric luminosity of these galaxies is kept constant. The *solid red* and the *blue dashed curves* show the flux at $\lambda = 850\,\mu\mathrm{m}$ and $\lambda = 175\,\mu\mathrm{m}$, respectively. *On the right*, the index β of the dust emissivity is varied, and the temperature of the dust $T_\mathrm{d} = 38\,\mathrm{K}$ is kept fixed. *On the left*, $\beta = 1.5$ is fixed and the temperature is varied. It is remarkable how flat these curves are over a very wide range in redshift, in particular at $850\,\mu\mathrm{m}$; this is due to the very strong negative K-correction which derives from the spectral behavior of thermal dust emission, shown in Fig. 9.30. Whereas for these model calculations an Einstein–de Sitter model was assumed, the behavior is very similar also in Λ-dominated universes. Source: A. Blain et al. 1999, *Submillimeter-selected galaxies*, astro-ph/9908111, Fig. 1. Reproduced by permission of the author

restframe frequency is shifted to the far side of the maximum of the dust spectrum (see Fig. 9.30). Hence, a sample of galaxies that is flux-limited in the sub-mm domain should have a very broad z-distribution. The dust temperature is about $T_\mathrm{d} \sim 20\,\mathrm{K}$ for low-redshift spirals, and $T_\mathrm{d} \sim 40\,\mathrm{K}$ is a typical value for galaxies at higher redshift featuring active star formation. The higher T_d, the smaller the sub-mm flux at fixed bolometric luminosity.

Counts of sub-mm sources at high Galactic latitudes have yielded a far higher number density than was predicted by early galaxy evolution models. For the density of sources as a function of limiting flux S, at wavelength $\lambda = 850\,\mu\mathrm{m}$, one obtains

$$N(>S) \simeq 7.9 \times 10^3 \left(\frac{S}{1\,\mathrm{mJy}}\right)^{-1.1} \mathrm{deg}^{-2}\,. \qquad (9.1)$$

The identification of sub-mm sources. At first, the optical identification of these sources turned out to be extremely difficult: due to the relatively low angular resolution of single-dish sub-millimeter telescopes (for example, MAMBO has a beam with FWHM of $\sim 11''$ at $\lambda = 1.2\,\mathrm{mm}$), the positions of sources can only be determined with an accuracy of several acrseconds.[5] Typically, several faint galaxies can be identified on deep optical images within an error circle of this radius. Furthermore, Fig. 9.32 suggests that these sources have a relatively high redshift, thus they should be very faint in the optical. An additional problem is reddening and extinction by the same dust that is the source of the sub-mm emission.

The identification of sub-mm sources was finally accomplished by means of their radio emission, since a majority of the sources selected at sub-mm wavelengths can be identified in very deep radio observations at 1.4 GHz. Since the radio sky is far less crowded than the optical one, and since the VLA achieves an angular resolution of $\sim 1''$ at $\lambda = 20\,\mathrm{cm}$, the optical identification of the corresponding radio source becomes relatively easy. One example of this identification process is shown in Fig. 9.33. With the accurate radio position of a sub-mm source, the optical identification can then be performed. In most cases, they are very faint optical sources indeed, so that spectroscopic analysis is difficult and

[5]The accuracy with which the position of a compact source can be determined is approximately given by the ratio of the FWHM and the signal-to-noise ratio with which this source is observed.

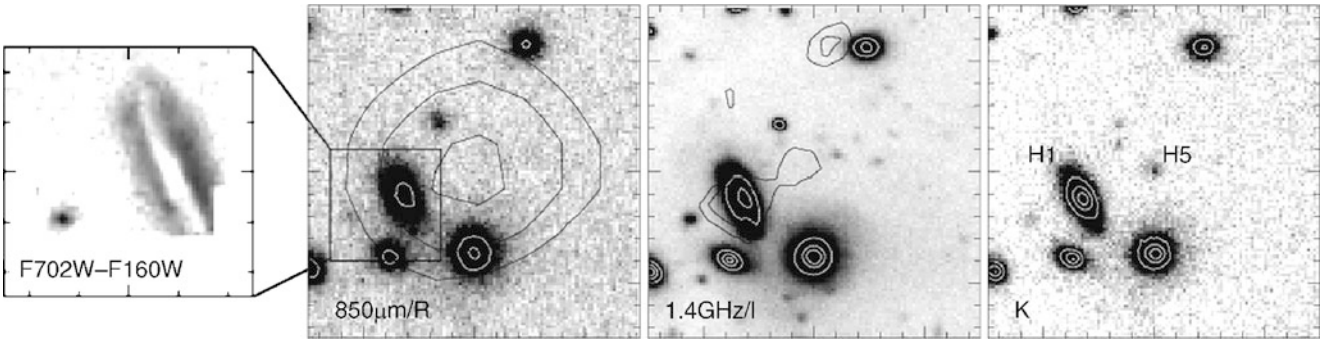


Fig. 9.33 The sub-mm galaxy SMM J09429+4658. The *three images on the right* have a side length of 30″ each, centered on the center of the error box of the 850 μm observation. The smaller image *on the left* is the difference of two HST images in red and infrared filters, showing the dust disk in the spiral galaxy H1. The *second image from the left* displays an R-band image, superposed with the contours of the SCUBA 850 μm emission. The *second image from the right* is an I-band image, superposed with the contours of radio emission at 1.4 GHz, and the *right-most panel* shows a K-band image. The radio contours show emission from the galaxy H1 ($z = 0.33$), but also weaker emission right at the center of the sub-mm map. In the K-band, a NIR source (H5) is found exactly at this position. It remains unclear which of these two sources is the sub-mm source, but the ratio of sub-mm to 1.4 GHz emission would be atypical if H1 is identified with the sub-mm source. Source: A. Blain et al. 1999, *Submillimeter-selected galaxies*, astro-ph/9908111, Fig. 4. Reproduced by permission of the author

very time-consuming. Another method for estimating the redshift results from the spectral energy distribution shown in Fig. 9.30. Since this spectrum seems to be nearly universal, i.e., not varying much among different sources, some kind of photometric redshift can be estimated from the ratio of the fluxes at 1.4 GHz and 850 μm, yielding reasonable estimates in many cases.

Redshift distribution of SMGs. The median redshift of sources with 850 μm flux larger than 5 mJy and 1.4 GHz flux above 30 μJy is about 2.2. However, at these sensitivities about half the sub-mm sources are unidentified in the radio, and hence their redshift distribution could be different from those with radio counterparts. In some of the sources with radio identification, an AGN component, which contributes to the dust heating, was identified, but in general newly born stars seem to be the prime source of the energetic photons which heat the dust. The optical morphology and the number density of the sub-mm sources suggest that we are witnessing the formation of massive elliptical galaxies in these sub-mm sources.

The great capabilities of interferometric observations enables a different method for identification and redshift determination of SMGs. The high angular resolution of the Plateau de Bure and, in particular, ALMA interferometers can pinpoint a sub-mm source very accurately, allowing the identification with optical or infrared sources. In addition, the large bandwidth of the ALMA receivers can take spectra of these sources over an appreciable range of wavelengths, and thus identify emission lines of molecules, predominantly those of the CO and water molecules, or fine-structure lines of atoms (specifically, ionized and neutral carbon), therefore removing the need for obtaining optical spectra of these optically faint sources. The potential of this method was recently established: even by observing with an incomplete array of telescopes, ALMA detected emission lines in 23 out of 26 sources selected by the South Pole Telescope (SPT; see Fig. 1.31) at 1.4 mm wavelength.

Once the precise locations of the sub-mm sources are determined with interferometric observations, they can be identified on deep optical images, and their redshifts be determined from optical spectroscopy. Hence, much of the potential bias in the redshift distribution of these sources, which are caused by the fact that only about half of the SMGs have a radio identification, are removed.[6] Indeed, if one compares the redshift distributions of both samples, shown in Fig. 9.34, one sees that they are significantly different. Using radio-identification as an intermediate step, and neglecting those SMGs for which no radio source could be identified, biases the redshift distribution of dusty star-forming galaxies low. The population extends over a significantly larger redshift range than concluded previously—in accord with expectations from the large negative K-correction. In addition, the redshifts of SMGs can be determined with the PdB and ALMA interferometers directly, using molecular line spectroscopy, without the need for optical spectroscopy. In particular the large bandwidth of the ALMA receiver allows one to cover a broad range of wavelengths, and thus of redshifts for which molecular lines can be detected and identified.

[6]The reason for this bias is the very different K-correction in the sub-mm and radio regimes, due to the very different slopes of the spectral energy distribution in these two regimes, as can be seen in Fig. 9.30: Whereas the flux in the sum-mm regime increases as a source is moved to higher redshifts, its radio flux decreases strongly, thus biasing against the detection of high-z SMGs in the radio.

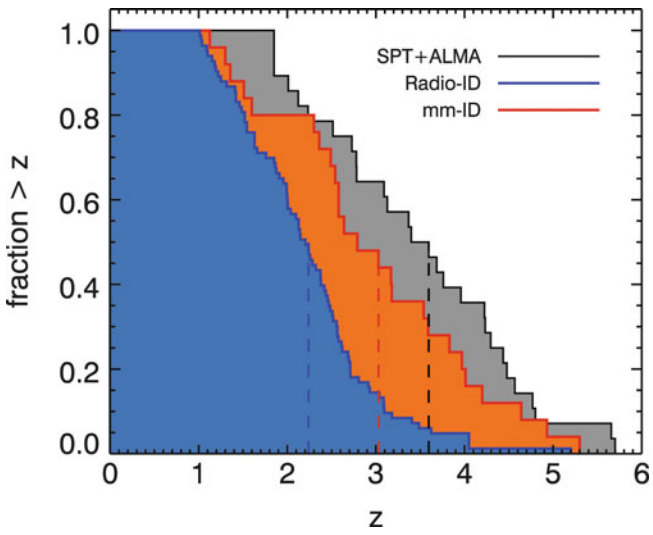


Fig. 9.34 The *grey line* shows the cumulative redshift distribution of dusty star-forming galaxies selected from the SPT survey, as determined by molecular line spectroscopy with ALMA; the distribution has a median redshift of $z_{\rm med} \sim 3.5$. The *blue curve* is the redshift distribution of SMGs that were identified with radio sources and subsequently their redshift was determined by optical spectroscopy, yielding $z_{\rm med} \sim 2.2$. The *orange curve* is the redshift distribution of SMGs identified at mm-wavelengths, with redshifts determined mainly through photometric fitting. Source: J.D. Vieira et al. 2013, *Dusty starburst galaxies in the early Universe as revealed by gravitational lensing*, Nature 495, 344, Fig. 3. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2013

Halo masses of sub-mm galaxies. As we discussed in Sect. 8.1.2, one can estimate the mass of the dark matter halo in which objects reside, by comparing their clustering properties with those of dark matter halos, as obtained from cosmological simulations. Sub-mm galaxies identified in a survey conducted by APEX (see the left panel of Fig. 9.35) allowed an estimate of the clustering length r_0 (defined such that the two-point correlation function is unity at separation r_0) of these sources, yielding $r_0 \approx (7.5 \pm 2)h^{-1}\,$Mpc. In the right panel of Fig. 9.35, this measurement of r_0 is related to that of other source populations. The clustering length of sub-mm sources is very similar to that of QSOs at the same redshift, and considerably larger than that for Lyman-break galaxies. Comparing this to the clustering length of dark matter halos with different masses, shown as dotted curves in Fig. 9.35, we conclude that SMGs live in relatively massive halos of several times $10^{12} M_\odot$ at $z \sim 2$. This results was recently confirmed by detecting a weak lensing signal around a sample of $\sim$ 600 relatively bright SMGs, which yields a characteristic halo mass of these galaxies of $\sim 10^{13} M_\odot$. This mass is comparable to that of current epoch massive elliptical galaxies, and suggests the interpretation that high-redshift SMGs evolve into present-day ellipticals. A large fraction of their stellar population is formed in the epoch at which the galaxy is seen as a SMG; by the end of this period, most of the gas in the galaxy is used up (or some fraction of it may be expelled), and in the remaining evolution little or no star formation occurs.

Additional support for this idea is provided by the fact that the sub-mm galaxies are typically brighter and redder than (restframe) UV-selected galaxies at redshifts $z \sim 2.5$. This indicates that the stellar masses in sub-mm galaxies are higher than those of LBGs.

A joint investigation of $z \sim 2$ sub-mm galaxies at X-ray, optical and MIR wavelengths yields that these sources are not only forming stars at a high rate, but that they already contain a substantial stellar population with $M \sim 10^{11} M_\odot$, roughly an order of magnitude more massive than LBGs at similar redshifts. The large AGN fraction among sub-mm galaxies indicates that the growth of the stellar population is accompanied by accretion and thus the growth of supermassive black holes in these objects. Nevertheless, the relatively faint X-ray emission from these galaxies suggests that either their SMBHs have a mass well below the local relation between $M_\bullet$ and stellar properties of (spheroidal) galaxies, or that they accrete at well below the Eddington rate. Furthermore, the typical ratio of X-ray to sub-mm luminosity of these sources is about one order of magnitude smaller than in typical AGNs, which seems to imply that the total luminosity of these sources is dominated by the star-formation activity, rather than by accretion power. This conclusion is supported by the fact that the optical counterparts of SMGs show strong signs of merging and interactions, together with their larger size compared to optically-selected galaxies at the same redshifts. This latter point shows that the emission comes from an extended region, as expected from star formation in mergers, rather than AGN activity.

Merging SMGs. The Herschel satellite found a strong sub-mm source, whose subsequent reobservations with different telescopes revealed this to be a merger of two SMGs. This source, called HXMM01, is shown in Fig. 9.36. In the near-IR, the image shows four main sources, of which the two brightest ones are galaxies at intermediate redshifts. The two fainter ones are galaxies at $z = 2.31$. Those coincide with strong sub-mm sources, as seen by the interferometric images at 880 μm, as well as through their strong molecular line emission. This pair of sources has a separation of 19 kpc, and is most likely undergoing a merger. The two foreground galaxies will cause a moderate lensing magnification of the SMGs, with an estimated $\mu \sim 1.6$. Thus, even after magnification correction, this source belongs to the brightest SMGs, with a corrected flux of $\sim$ 20 mJy at $\lambda = 850\,\mu$m. This source is extreme in several properties: Its dust temperature is estimated to be $T \approx 55\,$K, which is larger than that of most starburst galaxies. The estimated bolometric infrared luminosity is $L \sim 2 \times 10^{13} L_\odot$, with a corresponding star-formation rate of $\dot{M} \sim 2000 M_\odot$/yr. At this rate, the

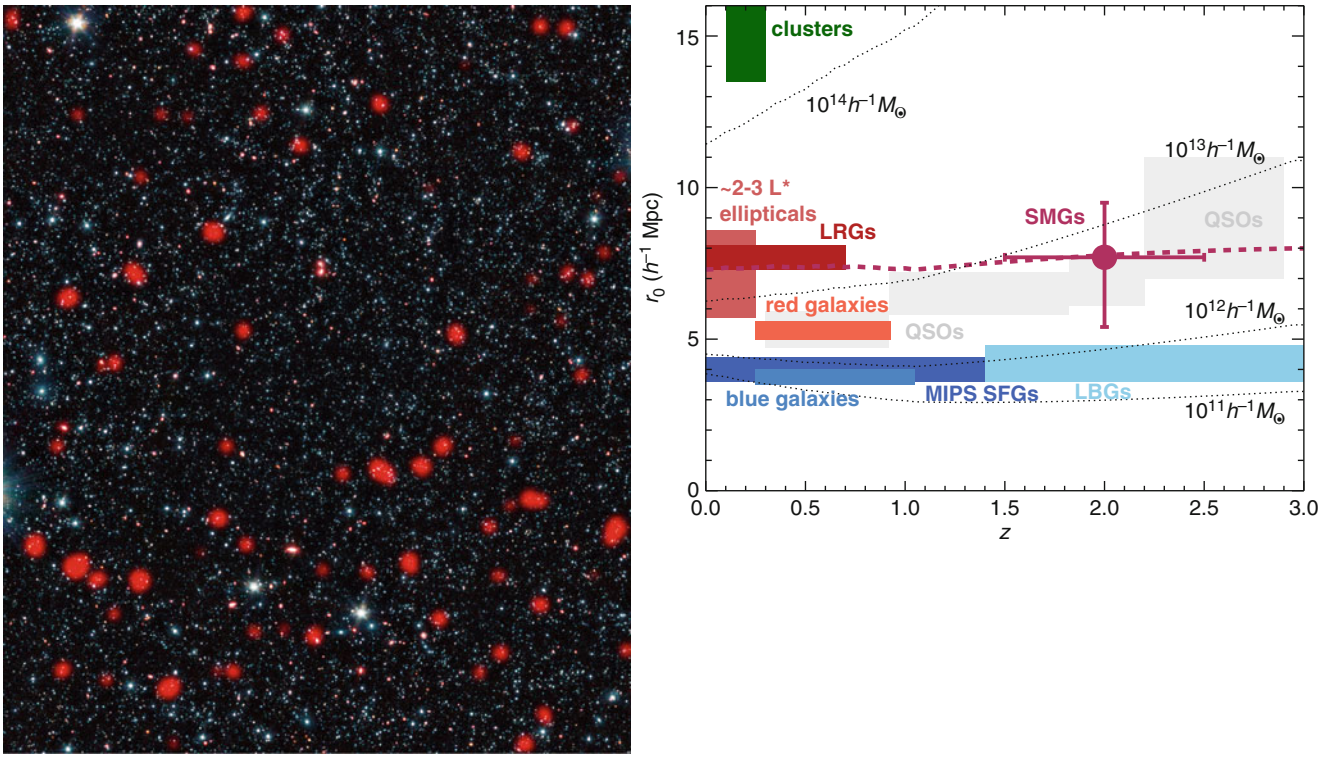


Fig. 9.35 *Left panel*: Sub-millimeter sources in the Extended Chandra Deep Field South (ECDFS). The image shows highly significant detections of sub-mm sources at 870 μm, as observed with the 12-m APEX telescope, superimposed on a Spitzer mid-IR image of the same sky region. On the 0.35 deg^2 of the ECDFS, 126 sub-mm sources were detected with a significance higher than 3.7σ. *Right panel*: Based on a clustering analysis of the sub-mm sources shown on the *left*, the clustering length r_0 of the sources was determined. The figure shows the clustering length as a function of redshift, for different types of sources: QSOs over a broad range of redshifts, local ellipticals and luminous red galaxies, local blue galaxies and star-forming galaxies at intermediate redshift, as selected by Spitzer at 24 μm wavelength, high-redshift Lyman-break galaxies, and galaxy clusters. The *dotted curves* indicate the clustering length of dark matter halos as a function of redshift, for different halo masses as indicated. Credit: *Left*: ESO, APEX (MPIfR/ESO/OSO), A. Weiss et al., NASA Spitzer Science Center. *Right*: R.C. Hickox et al. 2012, *The LABOCA survey of the Extended Chandra Deep Field-South: clustering of submillimetre galaxies*, MNRAS 421, 284, p. 291, Fig. 6b. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

molecular gas of $\sim 2.3\times10^{11}M_\odot$ will be turned into stars on a timescale of only 70 Myr. The brevity of this time interval implies that objects similar to HXMM01 should be rare. The molecular gas mass in this object is comparable with the stellar mass already present, i.e., the gas-mass fraction in this object is about 50 %. At the end of its merging process and star-formation episode, the resulting galaxy will have a stellar mass of $\sim 10^{11}M_\odot$, i.e., the stellar mass of a massive elliptical galaxy.

Although HXMM01 is not the only pair of merging SMGs yet discovered, it is brighter than the other ones by a factor of $\sim$ 2 (magnification corrected); correspondingly, its star-formation rate is also about twice that of the other merging systems. There are indications that the far-IR luminosity, and thus the star-formation rate, at fixed molecular mass (as measured by the luminosity in molecular CO lines), is higher by about a factor of three compared to galaxies that are not observed to be merging. This then is a direct hint at the possibility that merging triggers star-formation events, or bursts of star-formation.

In fact, there are several observations which suggest that the very large star-forming rates of many of the most luminous SMGs are due to merging events. Arguably the most direct one comes from interferometric observations of the molecular gas in SMGs, which shows that most of them have morphological and kinematical properties that identify them as merging systems. Hence, in accordance with the high merger rate of local ULIRGs, the extreme SMGs may also be triggered by merger events.

If dusty galaxies are at very high redshifts $z \gtrsim 5$, the peak of their spectral energy distribution shown in Fig. 9.30 is observed at wavelengths longward of 500 μm. Hence, for such sources the flux density is expected to increase with wavelength for $\lambda \lesssim 500\,\mu$m. The SPIRE instrument on the Herschel Space Observatory observed at 250, 350 and 500 μm, and blank survey fields at these frequencies can thus be used to search for very high-redshift sub-millimeter galaxies.

One such object found by this selection method is shown in Fig. 9.37. The galaxy HFLS3 is an extreme starburst

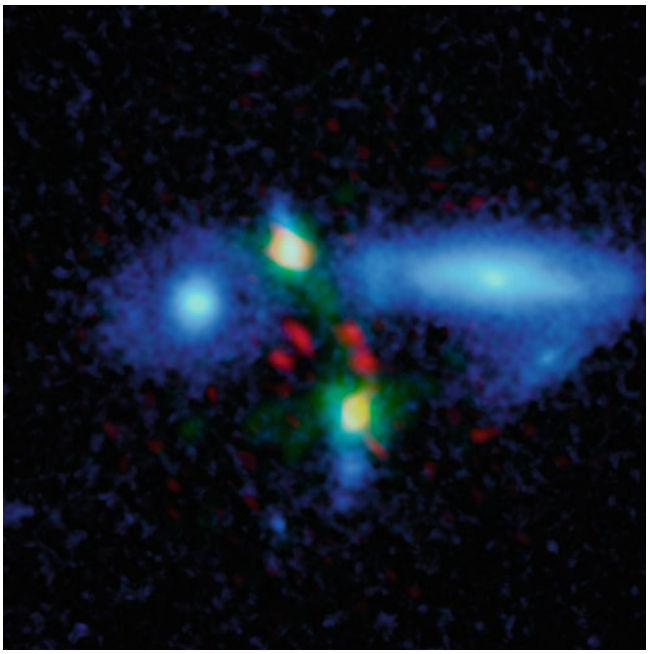


Fig. 9.36 The sub-millimeter galaxy HXMM01, detected by Herschel in the framework of the HerMES survey. In *blue*, a K-band image shows the presence of four main sources in this field; the two big ones are galaxies at $z = 0.66$ (*left*) and $z = 0.50$ (*right*). The two smaller ones, separated by $\sim 3''$, are at the same redshift of $z = 2.31$. In *red*, a sub-millimeter maps is shown, whereas *green* displays molecular emission. The *two yellow spots* show the regions where near-IR, sub-mm and molecular emission are spatially coincident; from those regions, the bulk of the flux seen by Herschel is emitted. The image has a size of about $10''$ on a side. Credit: ESA/NASA/JPL-Caltech/UC Irvine/STScI/Keck/NRAO/SAO

at $z = 6.34$, with an estimated star-formation rate of $\sim 3000\,M_\odot/\mathrm{yr}$, some factor of 20 larger than the local starburst Arp 220 (see Fig. 1.15). With an estimated gas mass of $\sim 10^{11}\,M_\odot$, this object would transform all its gas mass into stars on a time-scale of only $\sim 30\,\mathrm{Myr}$, assuming a constant star-formation rate. Spatially resolved molecular spectroscopy shows a velocity profile of this galaxy, which can be used to estimate a dynamical mass of $\sim 2.7\times10^{11}\,M_\odot$, yielding a gas-mass fraction of $\sim 40\,\%$, similar to what is found in typical $z \sim 2$ sub-millimeter galaxies.

Magnification bias of sub-millimeter sources. We mentioned before that some of the apparently most luminous sources of any kind have a large probability of being gravitationally lensed. The reason for this can be seen as follows: If we consider a population of sources, then their distribution in luminosity is most frequently described by a Schechter-like function. That means that for low luminosities, the luminosity function behaves as a power law in L, whereas for $L > L^*$, where L^* characterizes the break in the luminosity function, the density of sources decreases exponentially with L. In particular that means that there are essentially no sources with luminosity $L \gtrsim 5L^*$.

The probability that any given high-redshift source is gravitationally lensed and significantly magnified is small, of order 10^{-4}. Thus, if one picks random sources, the fraction of lensed ones among them will be similarly small. However, we can not pick random sources, but only sources above the flux limit of our observations. The situation in flux-limited samples can be quite different, since the magnification by lensing affects the mix of sources which are above the flux threshold. The probability that a source undergoes a magnification larger than μ can be shown to behave like μ^{-2}, provided the source is sufficiently small. Hence, large magnifications are correspondingly rare. But the probability for large magnifications decreases as a power law in μ—compared to the exponential decrease in the luminosity of sources for $L > L^*$. There will be a point where the (low-amplitude) power law overtakes the exponential, or in other words, where a source above a given flux threshold is more likely to be highly magnified, than having a very high intrinsic luminosity. An example of this effect are the bright $z \sim 3$ Lyman-break galaxies shown in Figs. 9.17 and 9.18, whose inferred luminosity is much larger than the L^* of this population of galaxies.

This is illustrated in Fig. 9.38, where the upper end of the 500 μm source counts are shown, together with a model decomposition of the counts into lensed and unlensed SMGs, nearby spiral galaxies and radio AGNs. As we just argued, for relatively low fluxes, the fraction of lensed sources is increasingly small. However, due to the steep decline of the unlensed counts, beyond a certain flux level they start to dominate the counts of SMGs. From that figure, one therefore expects that a SMG with $S \gtrsim 60\,\mathrm{mJy}$ at 500 μm has a fairly high probability of being lensed, whereas a SMG with $S \gtrsim 100\,\mathrm{mJy}$ almost certainly is a lensed source.

Hence, selecting sources with $S \gtrsim 100\,\mathrm{mJy}$, and cleaning the source catalog for nearby spirals (they are bright and big in the optical, and can thus easily be identified), one expects to obtain a clean lensed sample. Indeed, all five candidates selected from the survey data on which Fig. 9.38 is based, are most likely lensed, with the putative lens being identified on optical and NIR images. Furthermore, interferometric observations have confirmed the lensing nature of several of the candidates, by finding multiple images.

Using ALMA imaging of sources from the South Pole Telescope survey, selected at 1.4 and 2 mm to have the spectral index at these wavelengths corresponding to dust, and cleaning the sample for nearby objects (e.g., excluding sources detected by IRAS or in low-frequency radio catalogs), a large fraction of these sources turn out to be gravitationally lensed. In Fig. 9.39, ten of these are shown, where not only the multiple images of the sources (which are all at high redshift) are shown, but also a near-IR image of the field

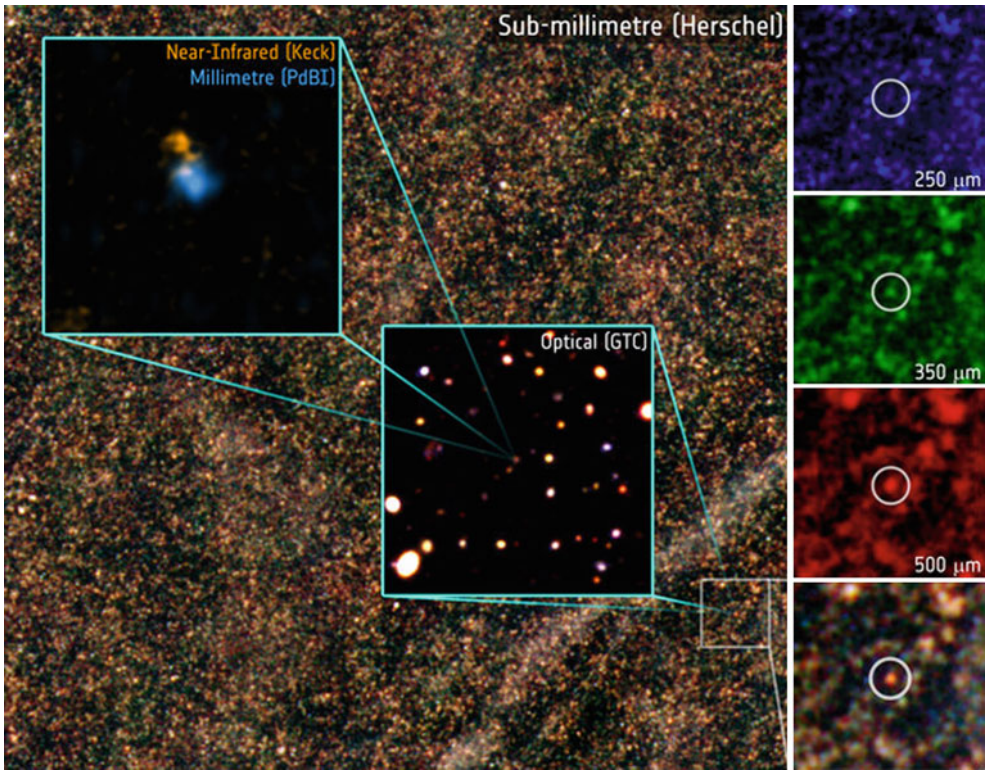


Fig. 9.37 The galaxy HFLS3 at $z = 6.34$, found by a color selection in the sub-millimeter regime. The big image shows a part of the HerMES blank field survey, with cut-outs at different wavelengths shown as small panels on the *right*, as well as an optical image and one combining near-IR and millimeter imaging. Credit: ESA/Herschel/HerMES/IRAM/GTC/W.M. Keck Observatory

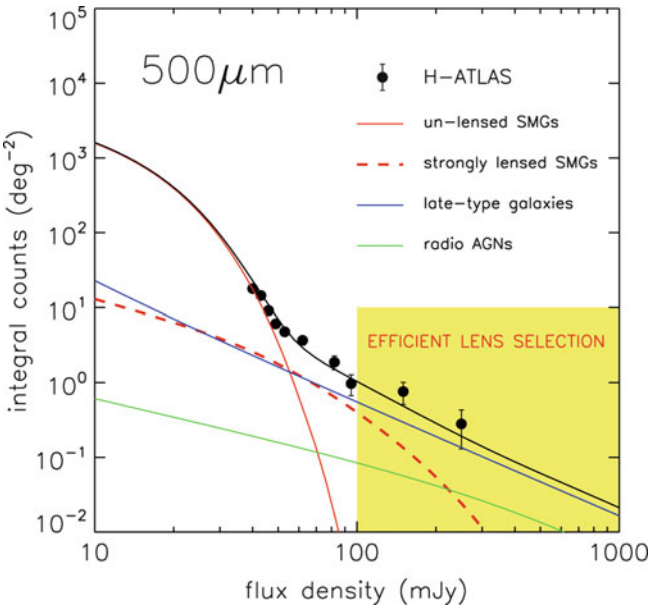


Fig. 9.38 The dominance of gravitationally-lensed magnified sub-mm galaxies at the bright end. The *dots* show the upper end of the 500 μm source counts of SMGs, obtained from the Herschel ATLAS survey. The *red curve* shows a model of the unlensed source counts, which has the shape of a Schechter-function, and drops off exponentially for large S. The *red dashed curve* shows the corresponding counts of lensed and magnified sources, whereas the *blue curve* is the contribution from low-redshift spiral galaxies. The lensed SMGs overtake the unlensed ones for fluxes $\gtrsim$ 60 mJy. Source: M. Negrello et al. 2010, *The Detection of a Population of Submillimeter-Bright, Strongly-Lensed Galaxies*, arXiv:1011.1255, Fig. 1. Reproduced by permission of the author

which clearly shows the lensing galaxy. Hence, selecting bright sources in the (sub)-millimeter regime yields a very high success rate of finding gravitational lens systems. All these lens systems would be missed in optical surveys, due to the faintness of the SMGs at optical and NIR wavelengths.

9.3.4 Damped Lyman-alpha systems

In our discussion of QSO absorption lines in Sect. 5.7, we mentioned that the Lyα lines are broadly classed into three categories: the Lyα forest, Lyman-limit systems, and damped Lyα systems, which are separated by a column density of $N_{\rm HI} \sim 10^{17}\,{\rm cm}^{-2}$ and $N_{\rm HI} \sim 2 \times 10^{20}\,{\rm cm}^{-2}$, respectively. The origin of the Lyα forest, as discussed in some detail in Sect. 8.5, is diffuse, highly ionized gas with small density contrast. In comparison, the large column density of damped Lyα systems (DLAs) strongly suggests that hydrogen is mostly neutral in these systems. The reason for this is self-shielding: for column densities of $N_{\rm HI} \gtrsim 2 \times 10^{20}\,{\rm cm}^{-2}$ the background of ionizing photons is unable to penetrate deeply into the corresponding hydrogen 'cloud', so that only its surface is highly ionized. Interestingly enough, this column density is about the same as that observed in 21 cm hydrogen emission at the optical radius of nearby spiral galaxies.

DLAs can be observed at all redshifts $z \lesssim 5$. For $z > 5$ the Lyα forest becomes so dense that these damped absorption lines are very difficult to identify. For $z \lesssim 1.6$ the Lyα

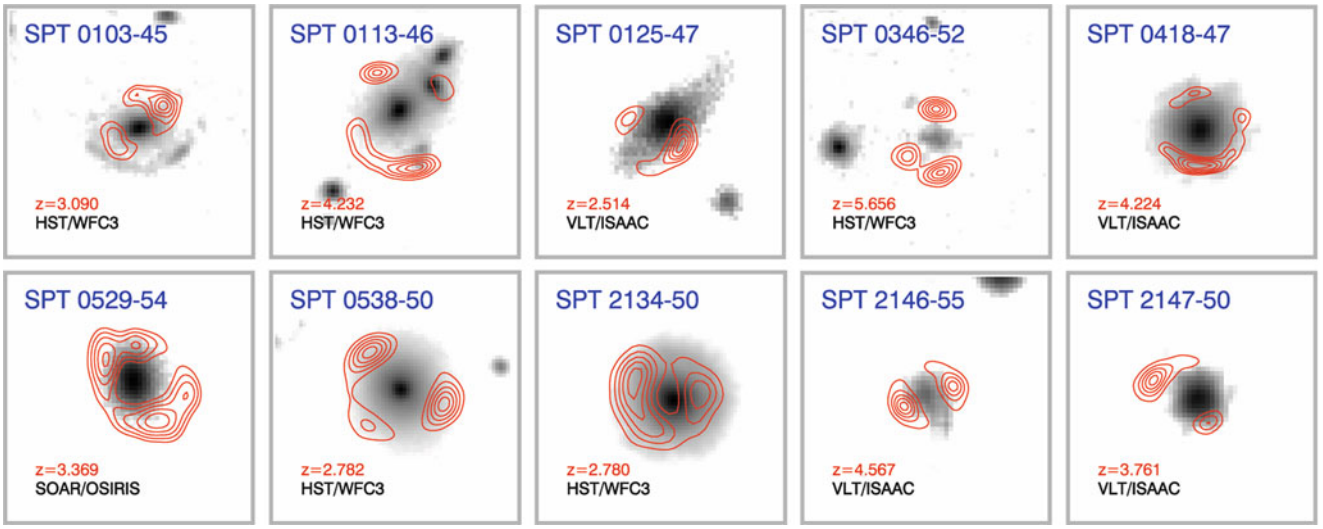


Fig. 9.39 Near-infrared (greyscale) and ALMA 870 μm images of ten high flux millimeter-selected sources from the South Pole Telescope survey. The redshifts of the SMGs are indicated in the *lower left* corner of each panel. All ten sources are obviously gravitationally lensed by a foreground galaxy. Source: J.D. Vieira et al. 2013, *Dusty starburst galaxies in the early Universe as revealed by gravitational lensing*, arXiv:1303.2723, Fig. 1. Reproduced by permission of the author

transition cannot be observed from the ground; since the apertures of optical/UV telescopes in space are considerably smaller than those on the ground, low-redshift DLAs are substantially more complicated to observe than those at higher z.

The neutral hydrogen mass contained in DLAs. The column density distribution of Lyα forest lines is a power law, given by (8.30). The relatively flat slope of $\beta \sim 1.6$ indicates that most of the neutral hydrogen is contained in systems of high column density. This can be seen as follows: the total column density of neutral hydrogen above some minimum column density $N_{\rm min}$ is

$$N_{\rm HI,tot}(N_{\rm max}) \propto \int_{N_{\rm min}}^{N_{\rm max}} {\rm d}N_{\rm HI}\; N_{\rm HI}\,\frac{{\rm d}N}{{\rm d}N_{\rm HI}} \propto \int_{N_{\rm min}}^{N_{\rm max}} {\rm d}N_{\rm HI}\; N_{\rm HI}^{1-\beta} = \frac{N_{\rm max}^{2-\beta} - N_{\rm min}^{2-\beta}}{2-\beta}\,, \tag{9.2}$$

and is, for $\beta < 2$, dominated by the highest column density systems. In fact, unless the distribution of column densities steepens for very high $N_{\rm HI}$, the integral diverges. From the extended statistics now available for DLAs, it is known that ${\rm d}N/{\rm d}N_{\rm HI}$ attains a break at column densities above $N_{\rm HI} \gtrsim 10^{21}\,{\rm cm}^{-2}$, rendering the above integral finite. Nevertheless, this consideration implies that most of the neutral hydrogen in the Universe visible in QSO absorption lines is contained in DLAs. From the observed distribution of DLAs as a function of column density and redshift, the density parameter $\Omega_{\rm HI}$ in neutral hydrogen as a function of redshift can be inferred. Apparently, $\Omega_{\rm HI} \sim 10^{-3}$ over the whole redshift interval $0 < z < 5$, with perhaps a small redshift dependence. Compared to the current density of stars, this neutral hydrogen density is smaller only by a factor ~ 3. Therefore, the hydrogen contained in DLAs is an important reservoir for star formation, and DLAs may represent condensations of gas that turn into 'normal' galaxies once star-formation sets in. Since DLAs have low metallicities, typically 1/10 of the Solar abundance, it is quite plausible that they have not yet experienced much star formation.

The nature of DLAs. This interpretation is supported by the kinematical properties of DLAs. Whereas the fact that the Lyα line is damped implies that its observed shape is essentially independent of the Doppler velocity of the gas, velocity information can nevertheless be obtained from metal lines. Every DLA is associated with metal absorption line systems, covering low- and high-ionization species (such as SiII and CIV, respectively) which can be observed by choosing the appropriate wavelength coverage of the spectrum. The profiles of these metal lines are usually split up into several components. Interpreted as ionized 'clouds', the velocity range Δv thus obtained can be used as an indicator of the characteristic velocities of the DLA. The values of Δv cover a wide range, with a median of $\sim 90\,{\rm km/s}$ for the low-ionization lines and $\sim 190\,{\rm km/s}$ for the high-ionization transitions. The observed distribution is largely compatible with the interpretation that DLAs are rotating disks with a characteristic rotational velocity of $v_{\rm c} \sim 200\,{\rm km/s}$, once random orientations and impact parameters of the line-of-sight to the QSO are taken into account.

Search for emission from DLAs. If this interpretation is correct, then we might expect that the DLAs can also be observed as galaxies in emission. This, however, is exceedingly difficult for the high-redshift DLAs. Noting that they are discovered as absorption lines in the spectrum of QSOs, we face the difficulty of imaging a high-redshift galaxy very close to the line-of-sight to a bright QSO (to quote characteristic numbers, the typical QSO used for absorption-line spectroscopy has $B \sim 18$, whereas an L_*-galaxy at $z \sim 3$ has $B \sim 24.5$). Due to the size of the point-spread function this is nearly hopeless from the ground. But even with the resolution of HST, it is a difficult undertaking. Another possibility is to look for the Lyα emission line at the absorption redshift, located right in the wavelength range where the DLA fully blocks the QSO light. However, as we discussed for LBGs above, not all galaxies show Lyα in emission, and it is not too surprising that these searches have largely failed. Only very few DLA have been detected in emission, with some of them seen only through the Lyα emission line at the trough of the damped absorption line, but with no observable continuum radiation. This latter fact indicates that the blue light from DLAs is considerably fainter than that from a typical LBG at $z \sim 3$, consistent with the interpretation that DLAs are not strong star-forming objects. But at least one DLA is observed to be considerably brighter and seems to share some characteristics of LBGs, including a high star-formation rate. In addition, a couple of DLAs have been detected by [OIII] emission lines. Overall, then, the nature of high-redshift DLAs is still unclear, due to the small number of direct identifications.

For DLAs at low redshifts the observational situation is different, in that a fair fraction of them have counterparts seen in emission. Whereas the interpretation of the data is still not unambiguous, it seems that the low-redshift population of DLAs may be composed of normal galaxies.

The spatial abundance of DLAs is largely unknown. The observed frequency of DLAs in QSO spectra is the product of the spatial abundance and the absorption cross section of the absorbers. This product can be compared with the corresponding quantity of local galaxies: the detailed mapping of nearby galaxies in the 21 cm line shows that their abundance and gaseous cross section are compatible with the frequency of DLAs for $z \lesssim 1.5$, and falls short by a factor ~ 2 for the higher-redshifts DLAs.

9.3.5 Lyman-alpha blobs

The search for high-redshift galaxies with narrow-band imaging, where the filter is centered on the redshifted Lyα emission line, has revealed a class of objects which are termed 'Lyman-α blobs'. These are luminous and very extended sources of Lyα emission; their characteristic flux in the Lyα line is $\sim 10^{44}$ erg/s, and their typical size is ~ 30 to ~ 100 kpc. Some of these sources show no detectable continuum emission in any broad-band optical filter. Hence, these sources seem to form a distinct class from the Lyman-alpha emitters discussed previously.

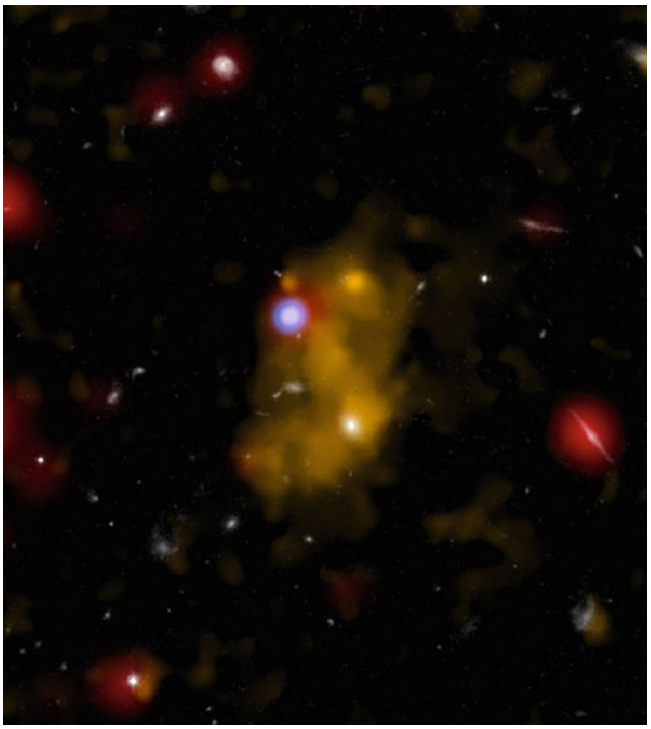


Fig. 9.40 Composite image of a Lyα blob at $z = 3.09$. The *yellow color* shows the Lyα emission, obtained by a narrow-band filter exposure. Inside the blob a galaxy is located, as seen in an optical (*white*) and infrared (8 μm, *red*) broadband image. X-ray emission (shown in *blue*) indicates that the Lyα emission is powered by an AGN. This image has a size of 38″. Credit: X-ray (NASA/CXC/Durham Univ./D.Alexander et al.); Optical (NASA/ESA/STScI/IoA/S.Chapman et al.); Lyman-alpha Optical (NAOJ/Subaru/Tohoku Univ./T.Hayashino et al.); Infrared (NASA/JPL-Caltech/Durham Univ./J.Geach et al.)

The nature of these high-redshifts objects remained unknown for a long time. Suggested explanations were wide-ranging, including a hidden QSO, strong star formation and associated superwinds, as well as 'cold accretion', where gas is accreted onto a dark matter halo and hydrogen is collisionally excited in the gas of temperature $\sim 10^4$ K, yielding the observed Lyα emission. It even seems plausible that the Lyman-α blobs encompass a range of different phenomena, and that all three modes of powering the line emission indeed occur. As a common feature, most of the Lyman-α blobs are associated with luminous galaxies, and are associated with strong infrared emission.

Chandra observations of 29 Lyα blobs detected five on them in X-rays; one of them is shown in Fig. 9.40. The X-ray sources are AGNs with $L_X \sim 10^{44}$ erg/s and rather large obscuration. Furthermore, these sources emit infrared light

from warm dust. The energy output of the AGN is sufficiently large to power the Lyα emission through photoionization. Hence, the AGN hypothesis has been verified for at least some of the sources.

Two of these Lyα blobs were discovered by narrow-band imaging of the aforementioned proto-cluster of LBGs at $z = 3.09$. Both of them are sub-mm sources and therefore star-forming objects; the more powerful one has a sub-mm flux suggesting a star-formation rate of $\sim 1000 M_\odot/\mathrm{yr}$. Spatially resolved spectroscopy extending over the full ~ 100 kpc size of one of the Lyα blobs shows that across the whole region there is an absorption line centered on the Lyα emission line. The optical depth of the absorption line suggests an HI column density of $\sim 10^{19}\,\mathrm{cm}^{-2}$, and its centroid is blueshifted relative to the underlying emission line by $\sim 250\,\mathrm{km/s}$. The spatial extent of the blueshifted absorption shows that the outflowing material is a global phenomenon in this object—a true superwind, most likely driven by energetic star formation and subsequent supernova explosions in these objects. Therefore, it appears likely that Lyα blobs are intimately connected to massive star-formation activity.

9.4 Properties of galaxies at high redshift

9.4.1 Demography of high-redshift galaxies

Being able to detect galaxies at high redshifts, we may first consider their abundance and investigate how their luminosity distribution compares with that galaxies in the local Universe. We are interested in a possible evolution of the luminosity function of galaxies with redshift, as this would clearly indicate that the galaxy population as a whole changes with redshift. The source counts from the Hubble Deep Field (Fig. 9.11) and its strong deviations from the non-evolution models provide a clear indication that the galaxy population evolves in redshift. This point will be considered here in somewhat more detail.

The most convenient way of summarizing the results is a representation of the estimated luminosity function in terms of a fit with a Schechter function (3.52). In that, the luminosity function is characterized by an overall normalization Φ^*, a characteristic luminosity L^* (or, equivalently, an absolute magnitude M^*) above which the abundance decreases exponentially, and a power-law slope α of the luminosity function at $L \ll L^*$. An evolution of Φ^* with redshift indicates that the abundance of luminous galaxies evolves. If L^* depends on z, one may conclude that the luminosity of a typical galaxy is different at higher redshifts. Finally, the faint-end slope α determines what fraction of the total luminosity of the galaxy population is emitted by the fainter galaxies—see (3.58).

The UV-luminosity function. Since high-redshift galaxies are selected using quite a variety of methods, as discussed in Sect. 9.1, and since the nature of the detected galaxies depends on their selection method, one has to consider different types of luminosity functions. For example, the Lyman-break method selects galaxies by their rest-frame UV radiation, so that from these surveys, the UV-luminosity function of galaxies can be obtained. Since the Lyman-break technique is applicable over a very wide redshift range, the UV-luminosity function has been obtained for redshifts between 2 and 7.

As already indicated by the galaxy counts shown in Fig. 9.11, the rest-frame UV-luminosity function of galaxies evolves strongly with redshift. In the redshift interval $2 \lesssim z \lesssim 4$, the characteristic luminosity L^* is about three magnitudes brighter than that of the local UV-luminosity function as determined with GALEX. This immediately shows that a typical galaxy at these redshifts is far more actively forming stars than local galaxies. Furthermore, the faint-end slope α is steeper for the high-redshift galaxies than for local ones. In fact, estimates yield $\alpha \sim -1.6$, indicating that much of the UV-luminosity density at high redshifts is emitted from rather faint galaxies—galaxies which are currently not observed due to the limited sensitivity of our instruments. Therefore, the overall abundance of UV-luminous galaxies is considerably larger in the redshift interval $2 \lesssim z \lesssim 4$ than it is today. Since the UV-radiation is produced by massive (and thus young) stars, this implies that the star-formation activity at those redshifts was much more intense than at the current epoch.

Going to even higher redshifts, the abundance of UV-selected galaxies decreases again, as can be seen in Fig. 9.41. The evolution is such that the characteristic luminosity decreases with higher z, and at the same time the faint-end slope steepens towards an estimated value of $\alpha \sim -1.8$. Hence, for these very high redshifts, most of the luminosity in the UV is emitted from faint sources. Recent attempts to find credible $z \sim 10$ galaxies using the Lyman-break technique yielded upper limits to the abundance of these objects, which yields upper limits to the luminosity function at this redshift. It appears that the decrease with z, visible in Fig. 9.41, accelerates towards even higher redshift.

In detail, these results are still burdened with quite some uncertainties, given the difficulties to identify very high redshift sources. Most of the conclusions are based on photometric redshifts only, since the spectroscopic verification of a $z \sim 7$ galaxy is extremely difficult, given that all spectral features blueward of the Lyα emission line are invisible due to intergalactic absorption, and that the radiation redward of the Lyα-line is redshifted into the near-IR. Hence, some of the sources may be misidentified and are in fact lower-redshift objects. Furthermore, since the identification of these very high redshift sources requires very deep observations,

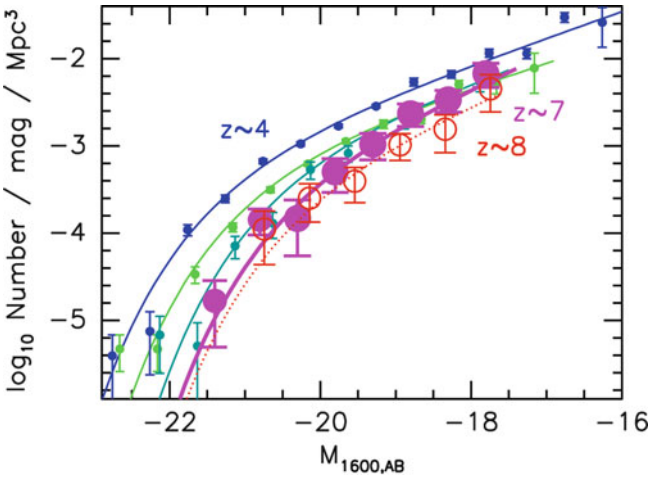


Fig. 9.41 High-redshift rest-frame UV luminosity functions of galaxies, obtained from data of the Hubble Deep Fields described in Sect. 9.2. Shown in *blue*, *green* and *cyan* are the luminosity functions at redshifts $z \sim 4$, 5, and 6, respectively, whereas the *magenta* and *red circles* show the estimated luminosity functions at $z \sim 7$ and 8. The *curves* show a Schechter-function fit to the data. Source: R.J. Bouwens et al. 2011, *Ultraviolet Luminosity Functions from 132 $z \sim 7$ and $z \sim 8$ Lyman-break Galaxies in the Ultra-deep HUDF09 and Wide-area Early Release Science WFC3/IR Observations*, ApJ 737, 90, p. 16, Fig. 12. ©AAS. Reproduced with permission

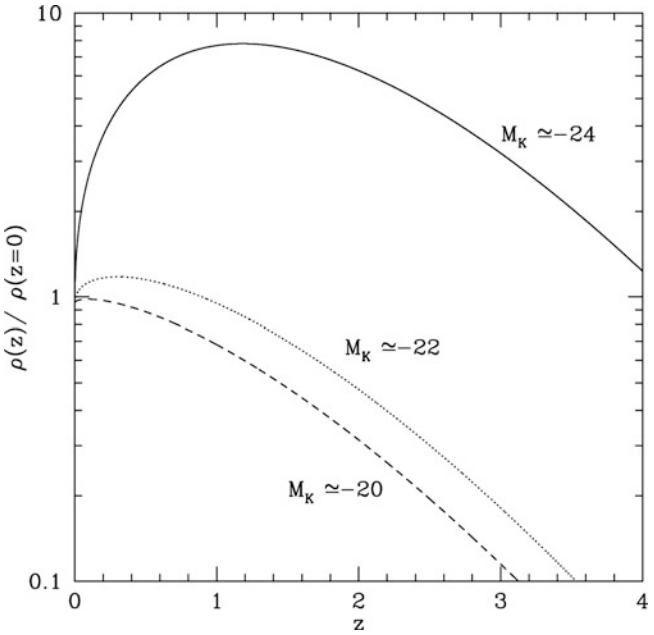


Fig. 9.42 The comoving number density of galaxies with fixed rest-frame K-band luminosity, normalized by their current space density, as a function of redshift. The very different redshift dependence of high- and low-luminosity galaxies, in the sense that the abundance gets shifted towards lower luminosity—and thus lower stellar mass—objects with cosmic time, is called downsizing. Source: Cirasuolo et al. 2010, *A new measurement of the evolving near-infrared galaxy luminosity function out to z = 4: a continuing challenge to theoretical models of galaxy formation*, MNRAS 401, 1166, p. 1173, Fig. 7. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

carried out only in a small number of fields, one must be aware of sampling variance—the fact that the distribution in a single small field may not be representative of the overall distribution. However, the general trends just discussed are established by now, providing a clear view of the evolution of the galaxy population with cosmic time.

Optical/NIR luminosity function. The rest-frame optical light is a somewhat better indicator of the total stellar mass of galaxies than is the UV-radiation. However, only when going to the NIR is the luminosity of star-forming galaxies not dominated by the radiation from newly-born stars; in addition, the K-band light is rather unaffected by dust obscuration. To assess the rest-frame K-band emission of high-redshift galaxies, one needs mid-IR observations which became possible with the Spitzer Space Telescope. The results of a combined analysis of optical, near-IR and mid-IR data show again a dramatic change of the luminosity function with redshift. The characteristic density of galaxies Φ^* decreases with redshift, as one might expect—there should be fewer galaxies around at higher redshifts. For example, at $z \sim 2$, Φ^* is about a factor 3.5 smaller than in the local Universe. In parallel to this, however, the characteristic luminosity L^* increases with z, by about one magnitude up to redshift 2. Hence it seems that a typical galaxy was brighter in the past. This phenomenon is quite counter-intuitive, given that the theory of structure formation predicts that more massive objects form in large abundance only at later redshifts—as follows from hierarchical structure formation. Another way to see this phenomenon is displayed in Fig. 9.42, which shows the comoving density of galaxies with fixed rest-frame K-band luminosity as a function of redshift, normalized to the corresponding local density. For rather low luminosities, the density decreases, but for high K-band luminosities, it increases by a factor $\sim$ 5 over a broad range in redshift, reaching a maximum at $z \sim 1.5$, and thereafter slowly decreases, but even at $z \sim 4$ the density is still higher than in the local Universe.

It thus seems that the typical galaxy at high redshift has a larger stellar mass than currently, or that the ratio of high-mass to low-mass galaxies was substantially larger at high z. This implies that with increasing cosmic time, the galaxy population becomes increasingly dominated by those with lower mass. This phenomenon has received the name downsizing. Models of galaxy evolution in a hierarchical universe need to be able to describe this effect; we will return to this in Chap. 10.

Mid-IR luminosity function. Whereas the rest-frame UV-radiation indicates the level of star formation which is unob-

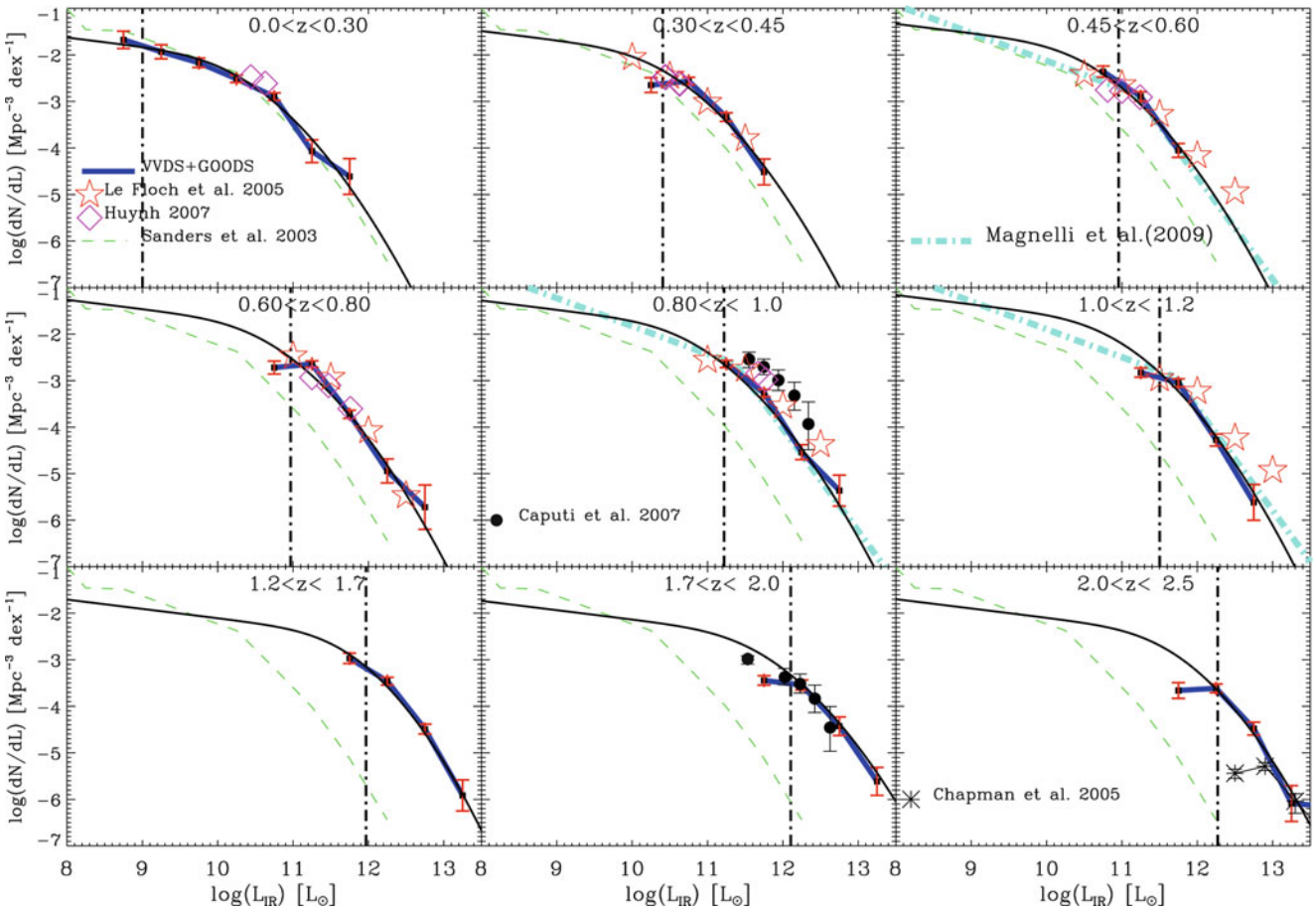


Fig. 9.43 Estimates of the luminosity function in the infrared, for different redshift intervals. The total infrared luminosity was obtained by combining optical, near- and mid-IR data with models of the spectral energy distribution, and are shown as points connected with a *thick blue curve*, whereas the *black curves* show a fit by a parametrized Schechter-like function. These redshift-dependent luminosity functions are compared to the one at $z = 0$, indicated by the *dashed green curve* in each panel. Data points from different studies are included as *different symbols*. The *vertical line* displays an estimate of the completeness limit of the samples. Source: G. Rodighiero et al. 2010, *Mid- and far-infrared luminosity functions and galaxy evolution from multiwavelength Spitzer observations up to* $z \sim 2.5$, A&A 515, A8, p. 17, Fig. 15. ©ESO. Reproduced with permission

scured by dust, it misses those star-forming galaxies which are heavily obscured by dust. Their activity can best be seen in the rest-frame mid- and far-IR. At a fixed wavelength, the emitted flux depends on the amount of heat absorbed by the dust—and reradiated as thermal dust emission—and on the dust temperature. Therefore, the most reliable indicator of the obscured star formation rate is the bolometric infrared luminosity. Whereas this is not directly observable—the combination of sensitivity and field-of-view of far-IR detectors allows one to study only relatively bright objects—the combination of observations at optical, near-IR and mid-IR can be used to estimate the dust temperature and thus to derive the bolometric IR luminosity from the Spitzer 24 μm data and the derived dust temperature.

The corresponding evolution of the IR luminosity function is shown in Fig. 9.43 for several redshift bins. Although for the higher-redshift bins only the highest luminosity sources can be observed, the figure shows a dramatic evolution towards higher redshift: The number density of luminous sources increases by a large factor compared to the local one. The trend is similar to that shown in Fig. 9.42 for the K-band luminosity function, but even stronger. The increase of very strongly star-forming galaxies with redshift is stronger than that with large stellar masses. Combined with the evolution of largely unobscured star-formation, shown in Fig. 9.41, we therefore conclude that the star-forming activities had a much higher level in earlier epochs of cosmic evolution than it has today. We will come back to this point in more detail in Sect. 9.6

Integrating the Schechter function fits shown in Fig. 9.43 over luminosity, one obtains the total infrared luminosity emitted per unit comoving volume. The redshift evolution of this IR luminosity density is shown in Fig. 9.44. It must be pointed out that these estimates carry quite some uncer-

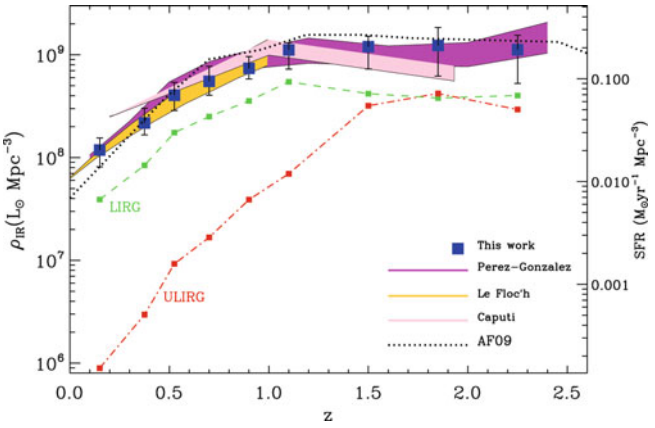


Fig. 9.44 From the luminosity functions shown in Fig. 9.43, the comoving total infrared luminosity density is obtained, and plotted as a function of redshift (left axis); on the right axis, the luminosity density is translated into an estimate of the corresponding star-formation rate density. Results from different studies are combined here, and they mutually agree quite well. The *green* and *red dashed curves* show the contribution to the luminosity density coming from galaxies with $L_{\rm IR} \geq 10^{11} L_\odot$ (LIRGs) and those with $L_{\rm IR} \geq 10^{12} L_\odot$ (ULIRGs). Source: G. Rodighiero et al. 2010, *Mid- and far-infrared luminosity functions and galaxy evolution from multiwavelength Spitzer observations up to* $z \sim 2.5$, A&A 515, A8, p. 18, Fig. 16. ©ESO. Reproduced with permission

tainty, since they require the extrapolation of the luminosity function to much fainter levels than those where data are available. In particular, the faint-end slope of the Schechter function is not at all well determined at high redshifts. Therefore, the detailed behavior of the luminosity density beyond $z \sim 1$ may be slightly different from what is shown in the figure. However, the contribution of the LIRGs (defined as $L \geq 10^{11} L_\odot$) and ULIRGs ($L \geq 10^{12} L_\odot$) to the luminosity density, also shown in Fig. 9.44, is much better determined. While the IR luminosity density increases by a factor $\sim$ 20 between today and $z \sim 1$, and stays roughly constant up to $z \sim 2.5$, the contribution from ULIRGs increases by at least a factor 100 over the same redshift range.

These results were confirmed with Herschel blank-field surveys, centered on fields for which multi-band observations were previously available (such as the GOODS fields or COSMOS). Observing in the far-IR, Herschel samples the peak of the spectral energy distribution directly, and fewer extrapolations are necessary to derive the bolometric infrared luminosity than required for using Spitzer data only. The analysis of the Herschel data showed that the evolution of the IR luminosity function with redshift is indeed dramatic. If one parametrizes the luminosity function as a Schechter function, the characteristic luminosity L^* in the infrared increases like $(1+z)^{3.5}$ for $0 \lesssim z \lesssim 2$, and $\propto (1+z)^{1.6}$ for $2 \lesssim z \lesssim 4$. The normalization Φ^* of the Schechter function decreases with redshift, with $\Phi^* \propto (1+z)^{-0.6}$ for $0 \lesssim z \lesssim 1.1$, and $\propto (1+z)^{-3.9}$ for $1.1 \lesssim z \lesssim 4$. In agreement with the results shown in Fig. 9.43, the comoving space density of very IR-luminous sources increases by huge factors from today to higher redshifts, before it starts to decline beyond redshift $z \sim 3$.

9.4.2 The color-magnitude distribution

The color bimodality, seen prominently in the local population of galaxies (see Sect. 3.1.3), has been in place at least since $z \sim 2$. As shown in Fig. 9.45, using spectroscopy of a 4.5 μm-flux limited sample of galaxies in the GOODS South field, for which deep photometry is available over a wide range of optical and infrared bands (including HST and Spitzer), the color bimodality can be clearly seen in all redshift intervals.

At even higher redshift, the sample of galaxies on the red sequence gets increasingly contaminated by dusty star-forming galaxies. However, one can account for this reddening and obtain dust-corrected colors for those galaxies. After this correction, the color bimodality can be detected out to redshifts $z \sim 3$, implying that already at young cosmic epochs, galaxies with an old stellar population coexisted with those which actively formed stars. Accordingly, the red sequence was formed early on, as can also be seen in Fig. 9.45.

This observational results implies that even at high redshifts, a large fraction of galaxies exists with a passively evolving stellar population. Whereas the star-forming galaxies—LBGs and SMGs—are the more spectacular objects at these high redshifts, many galaxies had formed their stars at even earlier epochs. From what we mentioned above—see, e.g., Fig. 9.42—the more massive galaxies seem to conclude most of the built-up of their stellar population at the highest redshifts. In parallel with the color-magnitude relation, also the local color-density relation was in place at least since $z \sim 1$.

9.4.3 The size and shape of high-redshift galaxies

The fact that the population of galaxies evolves strongly with redshift raises the expectation that the galaxies at high redshift are different from those in the local Universe. Here we will point out some of these differences.

The Hubble sequence and galaxy morphology. The majority of present day massive galaxies fall onto the morphological Hubble sequence (Fig. 3.2). In addition, we have seen that local galaxies can be classified according to their color, with most of them being either member of the red sequence or the blue cloud, as seen in Fig. 3.38, with a

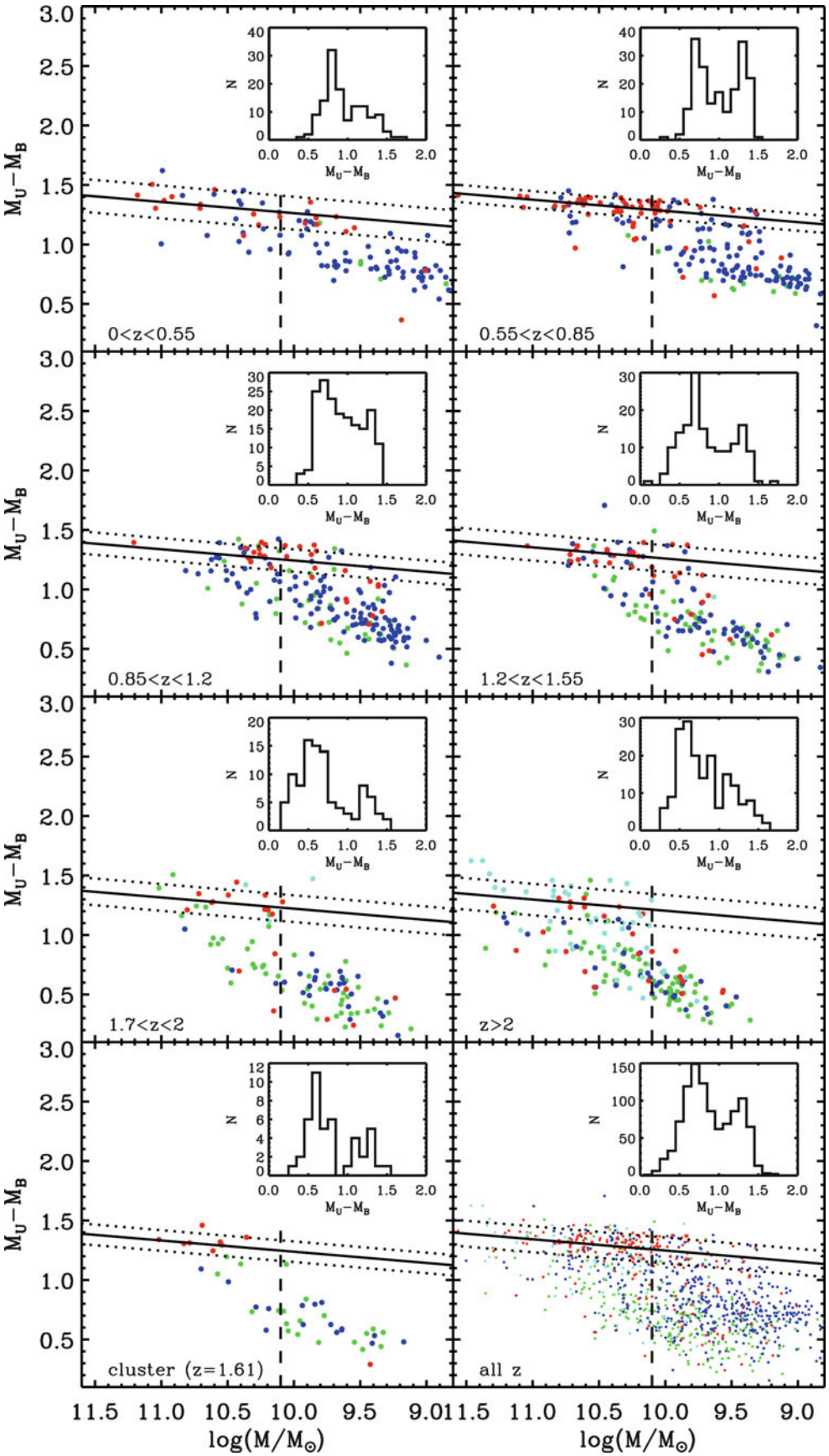


Fig. 9.45 The restframe $U - B$ color, as a function of stellar mass, in six redshifts bins, as indicated in the *six top panels*. The *two bottom panels* show the color-stellar mass diagram for a proto-cluster at $z = 1.61$, and the color-mass relation for all galaxies of the sample, irrespective of redshift. The *straight line* in each panel presents a fit to the red sequence, and the *vertical dashed line* indicates the completeness limit of the galaxy sample taken from the GMASS (Galaxy Mass Assembly ultradeep Spectroscopic Survey) project, in combination with multi-band photometry from optical to mid-infrared wavelengths in the GOODS-South field. The available high-resolution HST imaging allowed a morphological classification of the galaxies, according to which the symbols are color coded: early types (*red*), spirals (*blue*), irregulars (*green*), whereas *cyan symbols* are galaxies which are undetected in the optical bands and hence cannot be classified morphologically; those latter galaxies can appear only at the higher redshifts. The *small inset* in each panel shows the histogram of the color distribution. Source: P. Cassata et al. 2008, *GMASS ultradeep spectroscopy of galaxies at* $z \sim 2$. *III. The emergence of the color bimodality at* $z \sim 2$, A&A 483, L39, p. L40, Fig. 1. ©ESO. Reproduced with permission

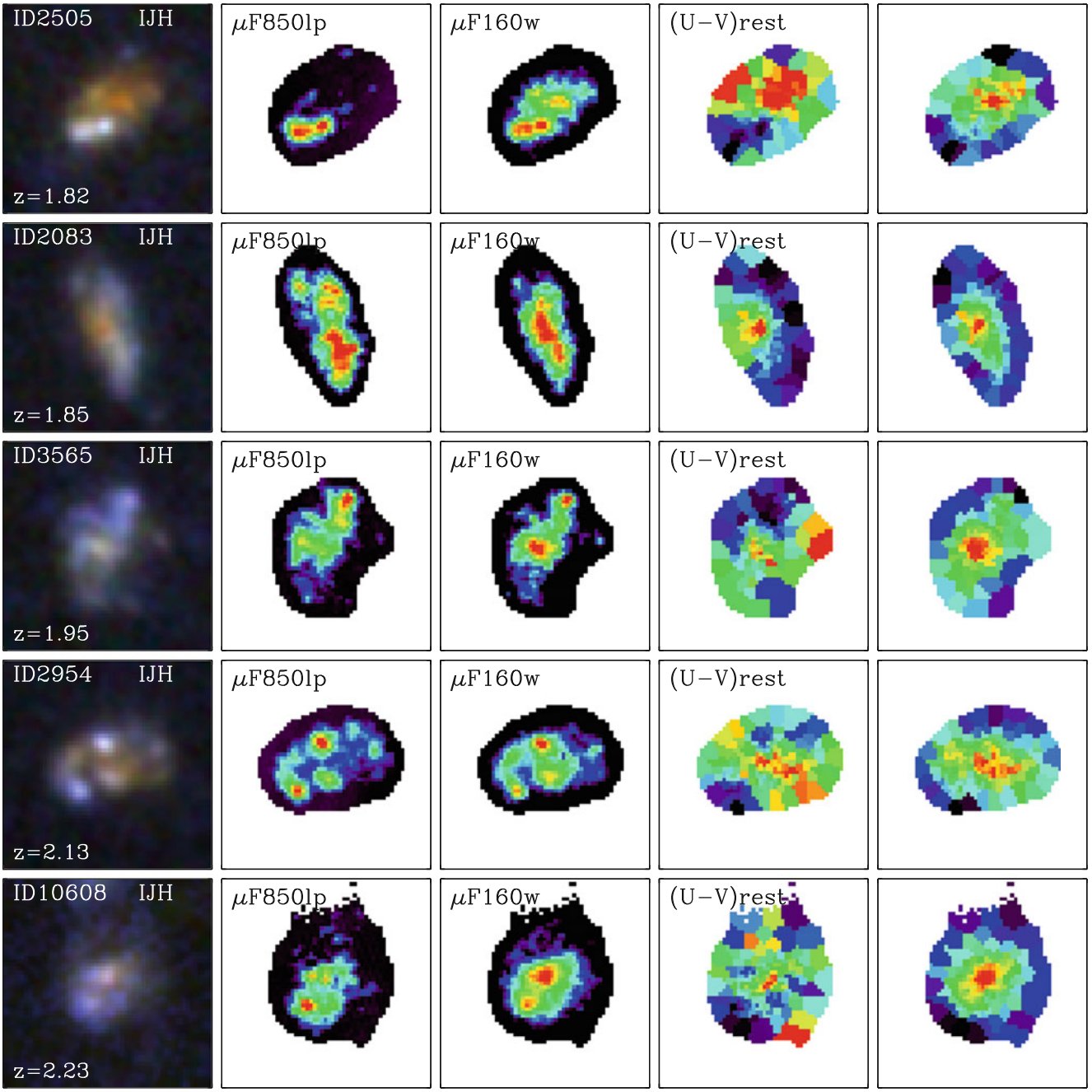
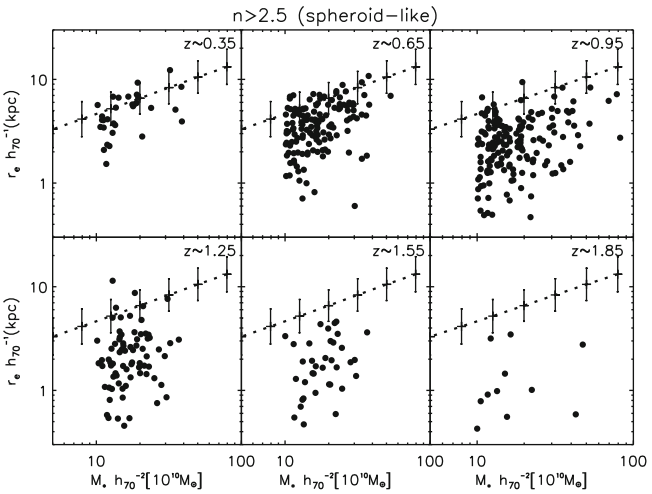


Fig. 9.46 Five $z \sim 2$ galaxies from the CANDLES survey. Shown *on the left* is an IJH color composite of the galaxy, which corresponds to the restframe UV-to-optical wavelength range. The surface brightness in two filters centered on 0.85 and 1.6 μm is shown in the *next two columns*. The *fourth column* displays the estimated $U - V$ restframe color, and the *right column* is the estimated stellar surface mass density. Source: S. Wuyts et al. 2012, *Smooth(er) Stellar Mass Maps in CANDELS: Constraints on the Longevity of Clumps in High-redshift Star-forming Galaxies*, ApJ 753, 114, p 6, Fig. 2. ©AAS. Reproduced with permission

tight correspondence between the Hubble classification and the galaxy color and morphological parameters, such as the Sérsic index.

The situation at high redshifts is quite different. At $z \gtrsim 3$, most of the galaxies are strongly star forming, with a rather small population of quiescent galaxies. The star-forming objects do not appear at all to have a regular morphology, rather, they are irregular, or clumpy. In Fig. 9.46, five $z \sim 2$ galaxies are shown as a IJH-color composite image. The irregular, knotty structure is easily seen, and many of the bright knots are clearly well separated from the center of the galaxy—these galaxies do not appear to fall on the Hubble sequence. These clumps have a characteristic size of ~ 1 kpc, and they seem to be projected onto a kind of disk galaxy.



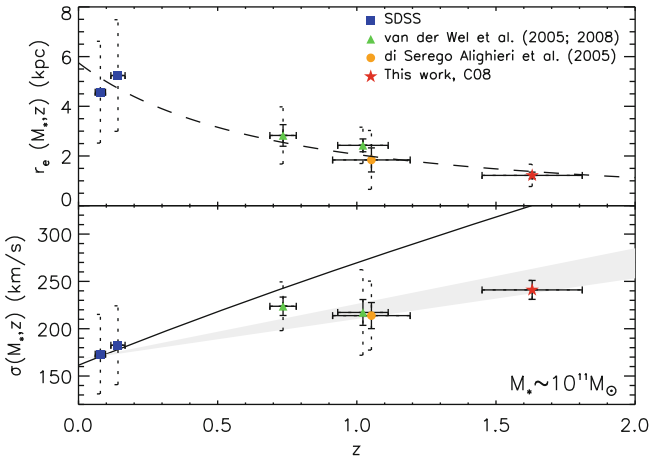


Fig. 9.47 *Left panel:* The effective radius (i.e., the radius within which half the light is emitted) versus stellar mass of galaxies with Sérsic index $n > 2.5$, representing early-type galaxies, for different redshifts. In each panel, the corresponding relation obtained in the local Universe is shown as *solid curve* with estimated uncertainties shown as error bars. *Right panel:* The mean effective radius (*top*) and velocity dispersion (*bottom*) as a function of redshift, for galaxies with stellar masses of $M_* \sim 10^{11} M_\odot$. The *solid curve* and *grey band* in the *bottom panel* shows two different models for the evolution of spheroidals. Source: *Left*: I. Trujillo et al. 2007, *Strong size evolution of the most massive galaxies since* $z \sim 2$, MNRAS 382, 109, p. 115, Fig. 7. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society. *Right*: A.J. Cenarro & I. Trujillo 2009, *Mild Velocity Dispersion Evolution of Spheroid-Like Massive Galaxies Since* $z \sim 2$, ApJ 696, L43, p. L46, Fig. 2. ©AAS. Reproduced with permission

In fact, using high angular resolution integral field spectroscopy, the rotation of several of such $z \sim 2$ galaxies could clearly be shown. But these disk are not rotating quietly, in contrast to local disk galaxies, they have a very large velocity dispersion. This fact renders the interpretation of the observed velocity field in terms of Kepler rotation more complicted than for thin, kinematically cold disk galaxies.

However, we should keep in mind that the rest-frame UV light distribution is dominated by star-forming regions. The second and third column in Fig. 9.46 show the surface brightness of these galaxies in two different filters separately, and an estimate of the rest-frame $U - V$ color is shown in the fourth column. We can see that the clumps are typically significantly bluer than the rest of the galaxy, and hence their stellar population has a younger age than the underlying disk. The stellar mass contained in the clumps make a 7 % contribution to the total stellar mass of the galaxy, but they contribute about 20 % to the star-formation rate. Finally, the right column shows the reconstructed stellar surface mass density. Now the picture is a quite different one: the stellar mass is seen to be centrally concentrated, and no prominent off-center clumps are present.

Quiescent galaxies become more abundant towards lower redshift; it is estimated that the stellar mass contained in quiescent objects increases by a factor $\sim$ 15 between redshifts 3 and 1, and by another factor of $\sim$ 3 from there until today. In other word, the number of passive red galaxies has at least doubled since $z = 1$ until today, so that many of the early-type galaxies in the current Universe arrived on the red sequence at rather low redshift. For $z > 2$, peculiar galaxies dominate the galaxy population, with some quiescent, spheroidal galaxies already present then, but a negligible disk population. At a redshift around $z \sim 2$, the abundance of spheroidal and disk galaxies together start to overtake the peculiar population, where this redshift depends on mass: at higher mass, the fraction of Hubble sequence-like galaxies is higher than at lower masses, indicating that they finish their morphological evolution earlier. Thus, starting from $z \sim 2$, the Hubble sequence is gradually built up.

Size evolution. Red, quiescent galaxies at $z \sim 2$ not only have a regular morphology compared to the clumpy star-forming galaxies, but they also are more massive and more compact. The latter aspects can be seen in Fig. 9.47, where in the left panel the effective radius is plotted as a function of stellar mass, for galaxies with Sérsic index $n > 2.5$ and different redshift bins. Compared to the local population of early-type galaxies, higher-redshift spheroidal galaxies are significantly smaller at fixed stellar mass. The effective radius as a function of redshift, for a fixed stellar mass, is shown in the upper panel on the right. The size evolution is fitted with a power law of the form $r_e \propto (1 + z)^{-1.48}$. The decrease in radius at fixed stellar mass by a factor of $\sim$ 3 at $z \sim 1.5$ implies that these galaxies have a stellar density larger by a factor $\sim$ 30 than present day early-type galaxies—these high-redshift galaxies are very different from the current population. The higher density also implies

a larger velocity dispersion, which is indeed observed, as seen in the bottom panel on the right in Fig. 9.47.

Indeed, such compact galaxies are very rare in the local Universe—that means that the typical $z \sim 2$ quiescent galaxy must have evolved significantly to fit into the local zoo of galaxies. Two principal possibility for this evolution exist: either the galaxies grow in size, at fixed stellar mass, or they accumulate more mass in their outer parts, thereby growing in mass and in size, such that they become less compact in this evolution. The latter possibility seems to be closer to the truth, as shown by simulations. Minor merging processes can yield an evolution that is compatible with the observational finding. Thus, early-type galaxies seem to grow inside-out: Their inner region was in place at earlier epochs, their outer parts were added lateron by merging processes.

9.4.4 The interstellar medium

The interstellar medium in high-redshift galaxies differs from that of local galaxies in a number of properties, of which we mention just a few here.

Metallicity. For local galaxies, there is a clear trend of increasing metallicity with increasing mass, as shown in Fig. 3.40. A similar trend is observed for Lyman-break galaxies at $z \sim 2$, except that the normalization of the mass-metallicity relation is smaller by a factor ~ 2, as can be seen in Fig. 9.48. Of course, this result does not come unexpectedly, since at earlier redshifts, there was less time to enrich the ISM.

Whereas the metallicity of star-forming galaxies is lower than at the current epoch, at least some galaxies managed to enrich their ISM to about Solar values, as can be inferred from the metallicity of the broad-line emitting gas in high-redshifts QSOs. Not only did these objects form supermassive black holes with $M_\bullet \gtrsim 10^9 M_\odot$, but the chemical evolution in these objects was already mature at a small cosmic age. Of course, luminous high-z QSOs are rare and most likely populate the most massive halos available at these epochs.

Gas content. Second, the high star-formation rate implies the presence of a large reservoir of gas. As we have seen, in the current Universe the gas-mass fraction of even late-type spiral galaxies is below $\sim 30\,\%$; in contrast to that, high-z star-forming galaxies typically have a gas-mass fraction of $\sim 50\,\%$.

Instead of considering the absolute star-formation rate $\dot{M}$, it is often meaningful to study the specific star-formation rate, $\dot{M}/M_*$, i.e., the star-formation rate per unit stellar mass. This quantity has the units of an inverse time: the inverse of the specific star-formation rate is the time it would take to built up the stellar mass present if the star-formation rate would be a constant. At a fixed gas-mass fraction, this time is very similar to the time-scale on which all the gas in a galaxy is transformed into stars. The specific star-formation rate of Lyman-break galaxies increases by a factor ~ 10 between the current epoch and $z \sim 1.5$, but then appears to stay remarkably constant out to $z \sim 7$ (where, of course, the results at the highest redshifts carry an appreciable uncertainty).

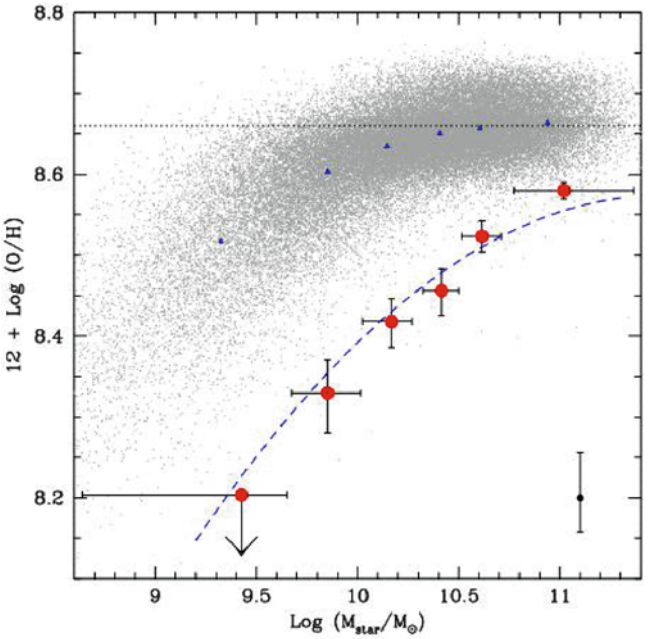


Fig. 9.48 The metallicity of UV-selected $z \sim 2$ galaxies, as a function of stellar mass (*big dots* with error bars). For comparison, the mass-metallicity relation for local SDSS galaxies is shown, as in Fig. 3.40. Whereas the shape of the mass-metallicity relation is similar at both redshifts, the normalization is lower by about a factor of two at $z \sim 2$. Source: D.K. Erb et al. 2006, *The Mass-Metallicity Relation at* $z \gtrsim 2$, ApJ 644, 813, p. 819, Fig. 3. ©AAS. Reproduced with permission

Dust. The fact that the far-IR emission is the best indicator for star formation already indicates that these galaxies must contain a significant dust abundance. The dust temperature, which determines the spectral shape in the far-IR, can vary in these dusty galaxies over quite a substantial range, $25\,\mathrm{K} \lesssim T_\mathrm{d} \lesssim 65\,\mathrm{K}$, as determined recently from Herschel observations. The impact of the dust temperature on the spectral shape means that single-band selection, e.g., the flux at $850\,\mu\mathrm{m}$, can bias against objects with hotter dust temperature.

Whereas most of the properties of $z \sim 6$ QSOs are indistinguishable from those of low-redshift QSOs, there are signs that they differ in their near-IR properties. In local QSOs, the UV/optical continuum emission is believed to be mainly due to the accretion disk, whereas the near-IR

radiation is due to hot dust, heated by the AGN. The ratio of NIR-to-optical luminosity of $z \lesssim 5$ QSOs is confined to a rather small range around ~ 1. In a sample of 21 $z \sim 6$ QSOs, there are two sources without detected NIR emission, yielding an upper limit to the NIR-to-optical flux ratio that is one order of magnitude smaller than the value typically observed. Using a control sample of more than 200 lower-z QSOs, not a single one has this flux ratio as low as the sources at $z \sim 6$. The lack of detectable NIR emission can be attributed to the lack of dust in these systems.

A clue for the origin of this lack of dust is obtained from a second finding: the NIR-to-optical flux ratio for lower-z QSOs shows no correlation with the SMBH mass as estimated from the width of broad emission lines. In contrast to that, there seems to be a strong dependence of this flux ratio on the SMBH mass in the sample of $z \sim 6$ QSO, in that the ratio increases with increasing $M_\bullet$. A simple interpretation of this result could be that these high-redshift QSOs were able to build up their SMBH and the corresponding accretion disk, but that they were unable yet to form large masses of dust. The larger $M_\bullet$, the more evolved is the AGN, and the more dust was created. It remains to be seen whether this interpretation survives further observational tests.

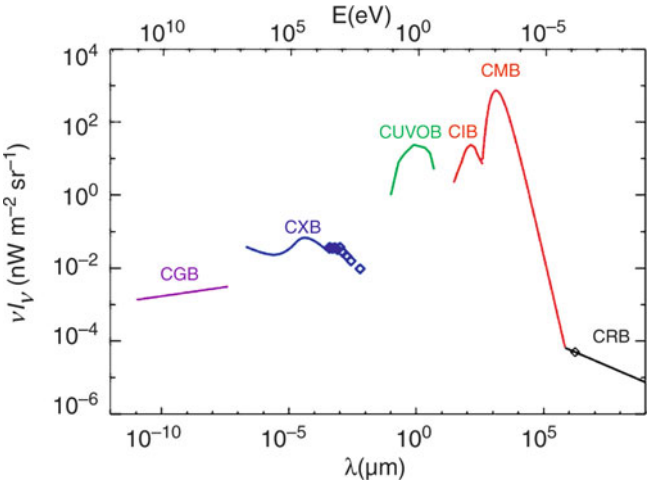


Fig. 9.49 Spectrum of cosmic background radiation, plotted as νI_ν versus wavelength. Besides the CMB, background radiation exists in the radio domain (cosmic radio background, CRB), in the infrared (CIB), in the optical/UV (CUVOB), in the X-ray (CXB), and at gamma-ray energies (CGB). With the exception of the CMB, all of these backgrounds can be understood as a superposition of the emission from discrete sources. Furthermore, this figure shows that the energy density in the CMB exceeds that of other radiation components, as was assumed when we considered the radiation density in the Universe in Chap. 4. Source: M.G. Hauser & E. Dwek 2001, *The Cosmic Infrared Background: Measurements and Implications*, ARA&A 39, 249, Fig. 1. Reprinted, with permission, from the *Annual Review of Astronomy & Astrophysics*, Volume 39 ©2001 by Annual Reviews www.annualreviews.org

9.5 Background radiation at smaller wavelengths

The cosmic microwave background (CMB) is a remnant of the early hot phase of the Universe, namely thermal radiation from the time before recombination. As we extensively discussed in Sect. 8.6, the CMB contains a great deal of information about our Universe. Therefore, one might ask whether background radiation also exists in other wavebands, which then might be of similar value for cosmology. The neutrino background that should be present as a relic from the early epochs of the Universe, in the form of a thermal distribution of all three neutrino families with $T \approx 1.9\,\mathrm{K}$ (see Sect. 4.4.3), is likely to remain undiscovered for quite some time due to the very small cross section of these low-energy neutrinos.

Indeed, apparently isotropic radiation has been found in wavelength domains other than the microwave regime (Fig. 9.49). In this figure, the background radiation measured as νI_ν is plotted against wavelength, so that the curve shows the intensity per logarithmic frequency interval. Following the terminology of the CMB, these are called background radiation as well. However, the name should not imply that it is a background radiation of cosmological origin, in the same sense as the CMB. From the thermal cosmic history (see Sect. 4.4), no optical or X-ray radiation is expected from the early phases of the Universe. Hence, for a long time it was unknown what the origin of these different background components may be.

At first, the early X-ray satellites discovered a background in the X-ray regime (cosmic X-ray background, CXB). Later, the COBE satellite detected an apparently isotropic radiation component in the FIR, the cosmic infrared background (CIB).

In the present context, we simply denote the flux in a specific frequency domain, averaged over sky position at high Galactic latitudes, as background radiation. Thus, when talking about an optical background here, we refer to the sum of the radiation of all galaxies and AGNs per solid angle. The interpretation of such a background radiation depends on the sensitivity and the angular resolution of the telescopes used. Imagine, for instance, observing the sky with an optical camera that has an angular resolution of only one arcminute. A relatively isotropic radiation would then be visible at most positions in the sky, featuring only some very bright or very large sources. Thus, the background can be decomposed into a 'resolved' component, which can be attributed to individually identified sources, and the unresolved component. On improving the angular resolution, more and more individual sources become visible, so that a larger fraction of the background radiation is resolved. At

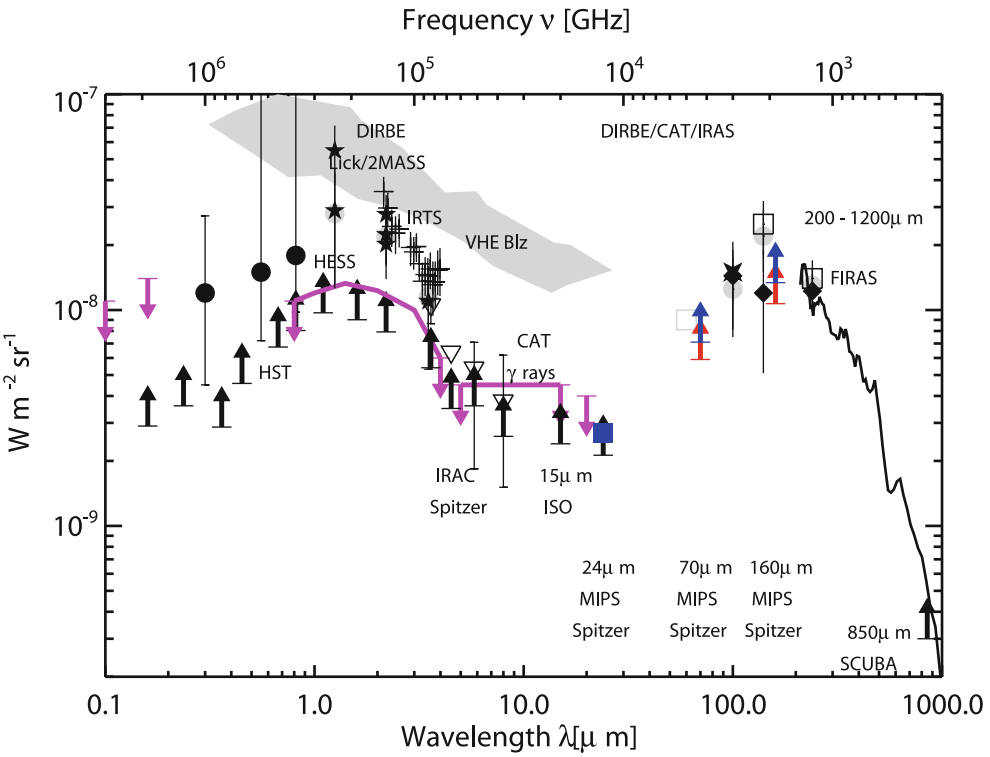


Fig. 9.50 The spectrum of the cosmic infrared background. The *black arrows* show the lower bound on the CIB, which is obtained from individually observed sources in these wavebands, i.e., by integrating the source counts at these frequencies up to the completeness limit. *Blue arrows* are lower bounds based on an extrapolation of the source counts to lower fluxes. The *magenta arrows* are upper limits, obtained from the Voyager spacecraft. The *filled circles* and *stars* with error bars are estimates, obtained with HST and several NIR instruments, as indicated. The *solid magenta line* presents an upper limit, obtained from the transparency of the Universe for very high-energy gamma rays with respect to electron-positron pair production (see Sect. 9.5.2); note that this upper limit is very close to the lower bound obtained from source counts. The *open squares* are estimates of the CIB from IRAS and the DIRBE instrument onboard COBE, while the *solid curve* at the longest wavelengths is the spectral estimate from FIRAS. Source: H. Dole et al. 2006, *The cosmic infrared background resolved by Spitzer. Contributions of mid-infrared galaxies to the far-infrared background*, A&A 451, 417, p. 426, Fig. 11. ©ESO. Reproduced with permission

optical wavebands, the Hubble Deep Fields have resolved essentially all of the background into individual sources. In analogy to this, one may wonder whether the CXB or the CIB can likewise be understood as a superposition of radiation from discrete sources.

9.5.1 The IR background

The first point to note from Fig. 9.49 is the relatively flat energy distribution between the UV- and the mm-regime. Since both, the UV-radiation and the far-IR radiation originate almost entirely from star-formation, the flat energy distribution implies that essentially half of the energetic photons emitted from newly-formed stars are absorbed by dust and reradiated in the FIR. Hence, estimates of the star-formation activity from UV-flux alone will on average be biased low by $\sim 50\,\%$.

Absolute measurements of the intensity of the background radiation are difficult to obtain, since it requires an absolute calibration of the instruments. The filled circles and stars with error bars in Fig. 9.50 show estimates of the background radiation level in the optical and near-IR. The black arrows show the integrated light of all sources which are detectable in the deepest observations with HST; within the error bars of the estimated level of the background radiation, these results are compatible with the background being solely due to the superposition of individual optical and near-IR sources, i.e., galaxies and (to a lesser degree) AGNs.

Observations of background radiation in the infrared are very difficult to accomplish, in particular due to the thermal radiation from the instruments and the zodiacal light. However, the DIRBE and FIRAS instruments onboard COBE provided a measurement (in fact, the detection) of the CIB. The question now is whether the CIB can be understood as well as being due solely to individual sources.

Confusion limit. Since mid- and far-IR observations are only possible from space, finding the answer to that question is challenging. Infrared observatories in space have a rather small aperture which, together with the long wavelength, yields a rather large point-spread function (PSF).

This implies that when one observes to low flux limits, where the mean angular separation of sources on the sky becomes comparable to the size of the PSF, these sources can not be separated. This yields a lower flux limit for the detection of individual sources, called the confusion limit. The smaller the telescope, the shallower is the confusion limit reached. For example, the flux limit down to which individual sources could be identified with the Spitzer satellite at 160 μm corresponds to only 7 % of the CIB at this wavelength. The much larger mirror on the Herschel satellite lowered the confusion limit such that individual sources can be identified which account for about 52 % of the CIB. Going to larger wavelength, the confusion limit is even more severe.

Stacking. However, one can dig deeper into the source counts with a technique called stacking. Taking the position of sources detected at some smaller wavelength (where the confusion limit is fainter), and adding up the flux in the longer wavelength band around all these positions, one obtains the mean long-wavelength flux of these sources. With this method, one will miss all fainter sources which do not have a detected counterpart in the short-wavelength input catalog, so that wavelength should be selected carefully. Given the characteristic spectrum of FIR-bright sources shown in Fig. 9.30, one expects that most of the sources radiating in the FIR will have an appreciable flux at 24 μm. Since Spitzer was particularly sensitive at this wavelength, the corresponding source catalog is best for a stacking analysis. Furthermore, if the redshifts of the sources selected at 24 μm is known, the stacking analysis can be used to determine the redshift distribution of the contributions to the CIB in the FIR. With stacking, the source counts can be followed to about three times lower flux than the confusion limit of individual sources permits.

The state of the art is defined by deep fields observed with the Herschel observatory, owing to its large aperture and sensitive instrumentation. From observing the well-studied GOODS, COSMOS, and ECDFS fields, where detailed multi-waveband data from other observatories are available, Herschel was able to attribute between 65 and 89 %, depending on wavelength, of the estimated CIB level between 100 and 500 μm to resolved sources or sources seen at 24 μm. A moderate extrapolation of the source counts to fainter flux limits then shows that the bulk, if not all, of the CIB comes from galaxies or AGNs.

In addition, the redshift distribution of the CIB could be determined. At wavelengths below $\sim$ 160 μm, more than half of the CIB radiation comes from sources at $z < 1$, whereas at longer wavelength, the source distribution shifts to increasingly higher redshifts. The major fraction of the CIB is due to galaxies with infrared luminosities in the range 10^{11} to $10^{12} L_\odot$, i.e., due to LIRGs.

9.5.2 Limits on the extragalactic background light from γ-ray blazars

The added flux of sources, either individually detected or obtained from a stacking analysis, yields a lower limit to the extragalactic background light, which in the UV, optical and near-IR regime is smaller than estimates of the total intensity of the background, as seen in Fig. 9.50, although these latter measurements have fairly larger error bars. Hence, the question arises whether there are other sources of the background light not identified as individual sources—for example, very low surface brightness galaxies that could escape detection. In fact, this question can be answered from observations of blazars (see Sect. 5.2.6) at energies in the GeV and TeV regime, as will be described next.

Attenuation of γ-rays: Condition for pair production. High-energy photons from distant sources propagate through the extragalactic background radiation field. If the photon energy is high enough, then by colliding with one of the background photons, it may produce an e^+e^--pair, in which case it does not reach the Earth. Thus, this pair production attenuates the flux from the source.

In order for pair production to occur, the product of the energies of the background-light photon (ϵ) and the photon from the source (E_γ) must be sufficiently high. If the two photons propagate in opposite direction (head-on collision), then the threshold condition is $\epsilon\, E_\gamma > (m_e c^2)^2$. In general, if the photon directions enclose an angle θ, this gets modified to

$$\epsilon\, E_\gamma > \frac{2(m_e c^2)^2}{1-\cos\theta}\ . \tag{9.3}$$

We see that the threshold energy is smallest for head-on collisions, where $\theta = \pi$. The cross-section for this process is small for photon energies very close to the threshold, reaches its maximum at about twice the threshold energy (9.3), and decreases again for larger energies. At the maximum of the cross-section, the relation between the two photon energies can be written in practical units,

$$\left(\frac{E_\gamma}{1\,\mathrm{TeV}}\right) = \frac{0.86}{1-\cos\theta}\left(\frac{\lambda}{1\,\mu\mathrm{m}}\right)\ , \tag{9.4}$$

from which we see that for the attenuation of TeV photons, extragalactic background photons in the near-IR are most efficient, whereas radiation in the tens-of-GeV-regime can be attenuated by UV-photons.

Optical depth. In order to derive the efficiency of the attenuation, one needs to calculate the optical depth $\tau_{\gamma\gamma}(E_\gamma, z)$, which depends on the energy of the γ-ray and the source redshift. To obtain $\tau_{\gamma\gamma}(E_\gamma, z)$, the pair-production cross section

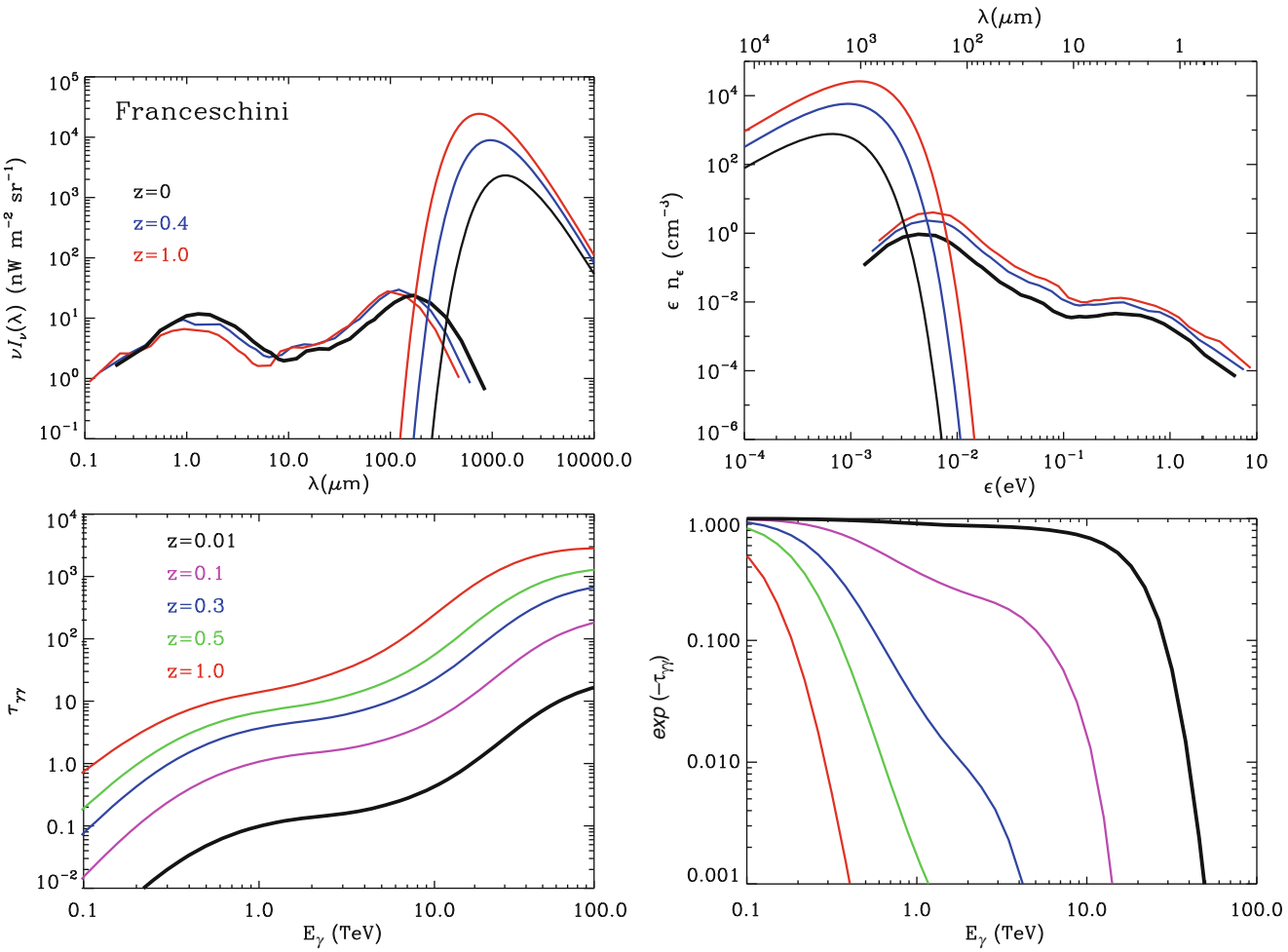


Fig. 9.51 *Top left:* Model of the extragalactic background light, at three different redshifts. The high-amplitude peak at long wavelengths is the CMB and thus shows the evolution of the Planck spectrum with redshift. *Top right:* The proper photon number density per logarithmic energy interval, for the same three redshifts. *Bottom left:* The optical depth $\tau_{\gamma\gamma}(E_\gamma, z)$ for pair production, as a function of the γ-ray energy. *Bottom right:* The attenuation factor $\exp[-\tau_{\gamma\gamma}(E_\gamma, z)]$ as a function of the γ-ray energy. Source: E. Dwek & F. Krennrich 2012, *The Extragalactic Background Light and the Gamma-ray Opacity of the Universe*, arXiv:1209.4661, Fig. 12. Reproduced by permission of the author

needs to be integrated along the line-of-sight to the source, multiplied by the spectral energy density of the background radiation; for this, the redshift evolution of the background radiation needs to be accounted for. Since the extragalactic background light can be observed only at the current redshift, one needs to model its redshift evolution, based on what is known about the source population. A particular model is shown in the top left panel of Fig. 9.51, which also includes the CMB, and the corresponding photon number density per logarithmic photon energy interval is shown in the top right panel. Based on the background light model, the optical depth for pair production can be calculated, which is shown in the bottom left panel of the same figure. $\tau_{\gamma\gamma}(E_\gamma, z)$ is a strong function of both, the γ-ray energy and the redshift. The plateau in $\tau_{\gamma\gamma}$ at energies ~ 2 TeV is due to the minimum of the background radiation spectrum at $\sim 10\,\mu$m.

The observed flux $S_{\rm obs}(E_\gamma)$ is related to the intrinsic (i.e., non-attenuated) flux $S_{\rm int}(E_\gamma)$ by

$$S_{\rm obs}(E_\gamma) = S_{\rm int}(E_\gamma)\, \exp[-\tau_{\gamma\gamma}(E_\gamma, z)]\,, \qquad (9.5)$$

where the attenuation factor $e^{-\tau_{\gamma\gamma}}$ is plotted in the bottom right panel of Fig. 9.51. We can see that the attenuation factor has a very steep decline with photon energy; for example, based on this model we would not expect to see 20 TeV photons from any source at $z \gtrsim 0.1$. This steep, exponential decline implies that the attenuation factor is very sensitive to the model of the extragalactic background light; conversely, if observations yield constraints on the attenuation factor, then strong constraints on the background light can be obtained, at wavelengths depending on the detection of the attenuation, according to (9.4).

Observational constraints on the attenuation. As mentioned in Sect. 5.5.4, blazars can emit at GeV and TeV energies, most likely caused by their jet pointing towards us. The Fermi Gamma-Ray Space Telescope, operating in the energy range between 200 MeV and 300 GeV, and the air Cherenkov observatories H.E.S.S., MAGIC and VERITAS, observing in the range between $\sim$ 50 GeV and 100 TeV, have detected more than 1000 blazars in the GeV-range and more than 30 at TeV energies. Whereas blazars in the GeV-range are observed out to redshifts $z \geq 1$, essentially all the TeV blazars are at low redshift, most of them having $z \lesssim 0.2$. This is indeed what one expects, based on the results in Fig. 9.51.[7] In principle, these observations of the spectral energy distribution could be used to determine the attenuation factor; however, in order to employ (9.5), one needs some knowledge about the intrinsic flux distribution $S_{\rm int}(E_\gamma)$.

There are various ways how realistic estimates for $\tau_{\gamma\gamma}$ can be obtained from the observations. The first of these is to base the intrinsic flux distribution on models of the γ-ray emission, and constrain these models by observations at somewhat lower photon energies. However, the models are sufficiently uncertain to preclude very accurate predictions, and thus the corresponding results on $\tau_{\gamma\gamma}$ are correspondingly uncertain. Second, since the relativistic electrons responsible for the inverse Compton effect that presumably causes the γ-ray emission, are expected to result from acceleration by shock fronts, as mentioned in Sect. 5.1.3, the slope of the electron distribution can not be arbitrarily flat, and thus the resulting inverse Compton radiation is also limited in slope; in the notation of Sect. 5.1.3, $s \gtrsim 2$ and thus $\alpha \gtrsim 0.5$. Assuming this value of the spectral index as a limit for the intrinsic flux distribution, the observed energy distribution can be translated into upper bounds on the attenuation. An even weaker assumption is used by a third methods, where one requires that the intrinsic flux $S_{\rm int}(E_\gamma) = S_{\rm obs}(E_\gamma)\,\exp[\tau_{\gamma\gamma}(E_\gamma, z)]$ does not (exponentially) increase with photon energy, as would be the case if the background light intensity would be overestimated.

Results. Depending on which of these methods are used, the results will differ slightly. A particular result is shown in Fig. 9.50, where the magenta curve indicates the upper bound on the extragalactic background light obtained from the high-energy observations of blazars. This upper bound is almost coincident with the lower bound obtained from the resolved source counts in the UV, optical and near-IR regime, strongly arguing that there are no other significant contributions of the background light than the observed galaxies and AGN. It thus seems that the spectral intensity of the background light in this spectral regime is now rather well determined. This conversely implies that the optical depth $\tau_{\gamma\gamma}$ is very well constrained, which in turn allows us to derive the intrinsic flux distribution from the observed one. In the future, this method will therefore yield more detailed constraints on the emission mechanism for high-energy radiation from blazars and other AGNs.

[7] There a few TeV blazars at higher redshift, but as we discussed in Sect. 5.2.6, the featureless spectrum of most blazars renders the determination of a secure redshift sometimes uncertain.

9.5.3 The X-ray background

The first X-ray experiment in astronomy, a balloon flight in 1962, discovered a diffuse X-ray emission across the sky, confirmed by the first X-ray satellites which discovered not only a number of extragalactic X-ray sources (such as AGNs and clusters of galaxies), but also an apparently isotropic radiation component. The spectrum of the cosmic X-ray background (CXB) is a very hard (i.e., flat) power law, cut off at an energy above $\sim$ 40 keV, which can roughly be described by

$$I_\nu \propto E^{-0.3}\,\exp\left(-\frac{E}{E_0}\right)\,, \qquad (9.6)$$

with $E_0 \sim 40$ keV. A recent estimate of the spectrum of the CXB is shown in Fig. 9.52. The estimates from different instruments agree in general, though differences in the level are clearly visible. These differences can have a number of origins, including cosmic variance (the spectral shape of the CXB is usually determined from rather small fields, so there could be variations from field to field), stray light entering the telescope, and remaining calibration uncertainties of the instruments. Together, the CXB is known with an uncertainty of $\sim 20\,\%$.

Initially, the origin of this radiation was unknown, since its spectral shape was different from the spectra of sources that were known at that time. For example, it was not possible to obtain this spectrum by a superposition of the spectra of known AGNs.

ROSAT, with its substantially improved angular resolution compared to earlier satellites (such as the *Einstein* observatory), conducted source counts at much lower fluxes, based on some very deep images. From this, it was shown that at least 80 % of the CXB in the energy range between 0.5 and 2 keV is emitted by discrete sources, of which the majority are AGNs. Hence it is natural to assume that the total CXB at these low X-ray energies originates from discrete sources, and observations by XMM-Newton and Chandra have confirmed this.

However, the X-ray spectrum of normal AGNs is different from (9.6), namely it is considerably steeper

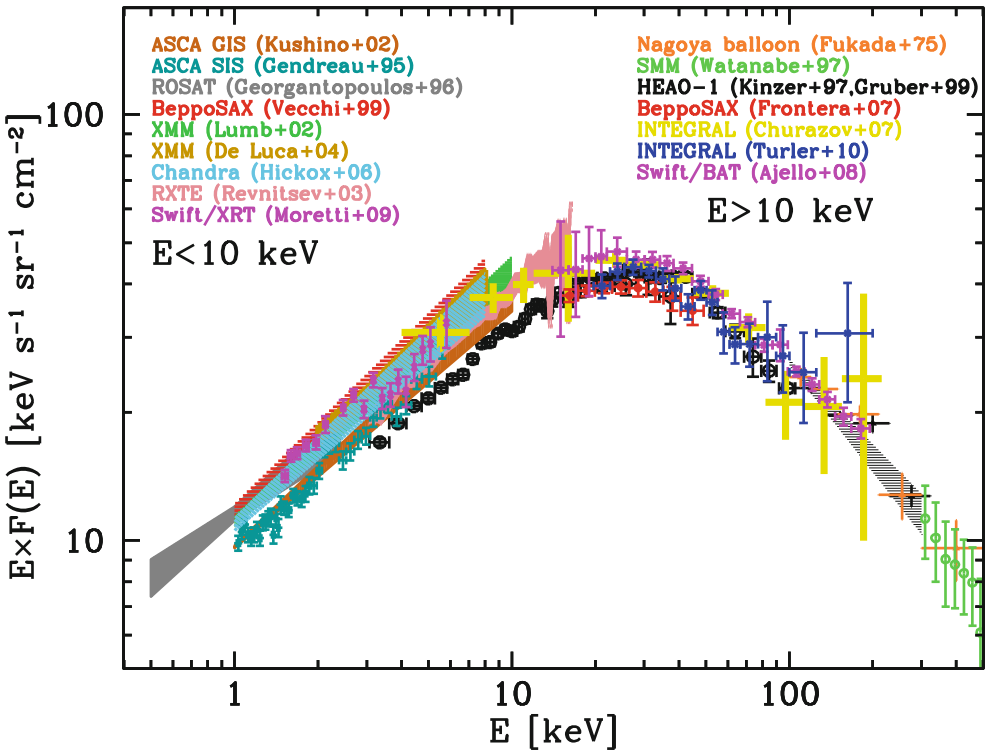


Fig. 9.52 Measurement of the cosmic X-ray background over a wide range of photon energies, measured by different satellites and instruments. Source: R. Gilli 2013, *The cosmic X-ray background: abundance and evolution of hidden black holes*, arXiv:1304.3665, Fig. 1. Reproduced by permission of the author

(about $S_\nu \propto \nu^{-0.7}$). Therefore, if these AGNs contribute the major part of the CXB at low energies, the CXB at higher energies cannot possibly be produced by the same AGNs. Subtracting the spectral energy of the AGNs found by ROSAT from the CXB spectrum (9.6), one obtains an even harder spectrum, resembling very closely that of thermal bremsstrahlung. Therefore, it was supposed for a long time that the CXB is, at higher energies, produced by a hot intergalactic gas at temperatures of $k_\mathrm{B}T \sim 30\,\mathrm{keV}$.

This model was excluded, however, by the precise measurement of the thermal spectrum of the CMB by COBE, showing that the CMB has a perfect blackbody spectrum. If a postulated hot intergalactic gas were able to produce the CXB, it would cause significant deviations of the CMB from the Planck spectrum, namely by the inverse Compton effect (the same effect that causes the SZ effect in clusters of galaxies—see Sect. 6.4.4). Thus, the COBE results clearly ruled out this possibility.

Deep observations with Chandra and XMM (e.g., in the CDFS shown in Fig. 9.14) have finally resolved most of the CXB also at higher energies, as seen in Fig. 9.53. From source counts performed in such fields, more than 75 % of the CXB in the energy range of $2\,\mathrm{keV} \le E \le 10\,\mathrm{keV}$ could be resolved into discrete sources. Again, most of these sources are AGNs, but typically with a significantly harder (i.e., flatter) spectrum than the AGNs that are producing the low-energy CXB. Such a flat X-ray spectrum can be produced by photoelectric absorption of an intrinsically steep power-law spectrum, where photons closer to the ionization energy are more efficiently absorbed than those at higher energy. According to the classification scheme of AGNs discussed in Sect. 5.5, these are Type 2 AGNs, thus Seyfert 2 galaxies and QSOs with strong intrinsic self-absorption. We should recall that Type 2 QSOs have only been detected by Chandra—hence, it is no coincidence that the same satellite has also been able to resolve the high-energy CXB.

However, at even higher energies most of the CXB was still unaccounted for—even the observed Type-2 AGNs could not account for it. It thus seems that there is a population of sources in the Universe which dominate the X-ray emission at high energies, still escape the observations at low X-ray frequencies. These could be heavily obscured AGNs, where only the hard X-rays manage to escape the emitting region. With the X-ray telescope onboard the Swift satellite, a significant number of such heavily obscured AGNs were found. Their estimated number density, together with their spectral energy distribution, make it plausible that they are the missing population of 'hidden black holes' responsible for the hard CXB.

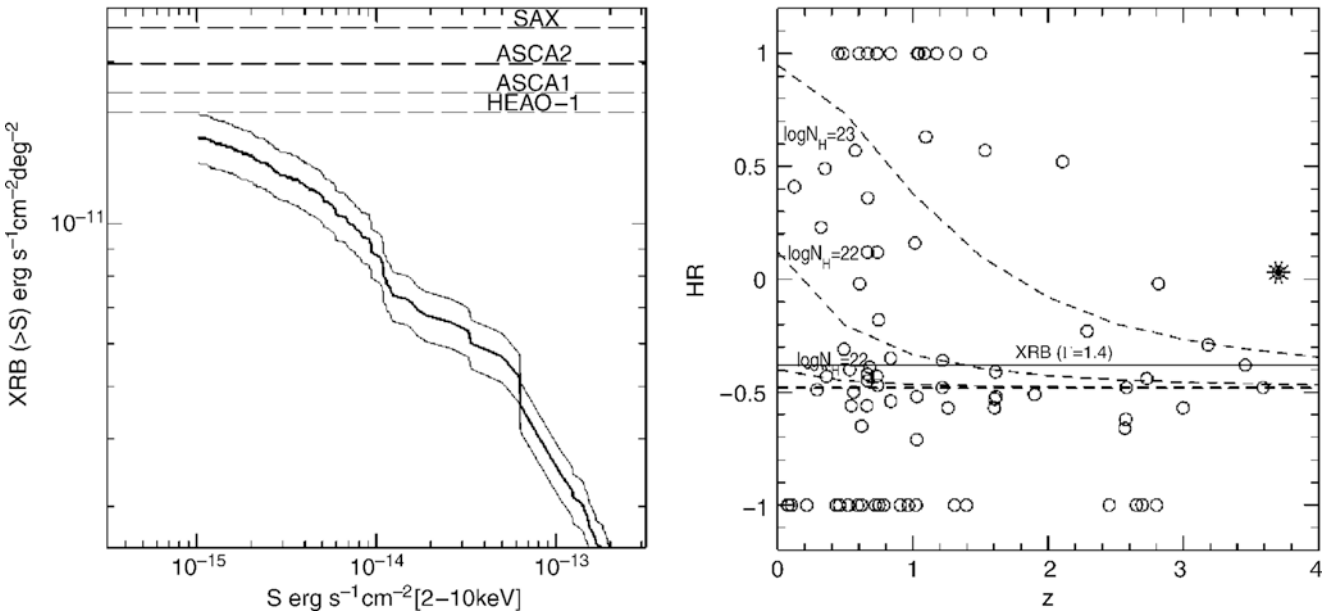


Fig. 9.53 In the *left panel*, the total intensity of discrete sources with an individual flux $> S$ in the energy range $2\,\mathrm{keV} \le E \le 10\,\mathrm{keV}$ is plotted (*thick curve*), together with the uncertainty range (between the two *thin curves*). Most of the data are from a 3×10^5 s exposure of the Chandra Deep Field. The *dashed lines* show different measurements of the CXB flux in this energy range; depending on which of these values is the correct one, between 60 and 90 % of the CXB in the Chandra Deep Field at this energy is resolved into discrete sources. In the *right panel*, the hardness ratio HR—specifying the ratio of photons in the energy range $2\,\mathrm{keV} \le E \le 10\,\mathrm{keV}$ to those in $0.5\,\mathrm{keV} \le E \le 2\,\mathrm{keV}$, $\mathrm{HR} = (S_{>2\mathrm{keV}} - S_{<2\mathrm{keV}})/(S_{>2\mathrm{keV}} + S_{<2\mathrm{keV}})$—is plotted as a function of redshift, for 84 sources in the Chandra Deep Field with measured redshifts. This plot indicates that the HR decreases with redshift; this trend is expected if the X-ray spectrum of the AGNs is affected by intrinsic absorption. The *dashed curves* show the expected value of HR for a source with an intrinsic power-law spectrum $I_\nu \propto \nu^{-0.7}$, which is observed through an absorbing layer with a hydrogen column density of N_H, by which these curves are labeled. Since low-energy photons are more strongly absorbed by the photoelectric effect than high-energy ones, the absorption causes the spectrum to become harder, thus flatter, at relatively low X-ray energies. This implies an increase of the HR (also see the *bottom panel* of Fig. 6.22). This effect is smaller for higher redshift sources, since the photon energy at emission is then larger by a factor of $(1 + z)$. Source: P. Tozzi et al. 2001, *New Results from the X-Ray and Optical Survey of the Chandra Deep Field-South: The 300 Kilosecond Exposure. II.*, ApJ 562, 42, p. 48, 49, Figs. 7, 9. ©AAS. Reproduced with permission

9.6 The cosmic star-formation history

The evolution of the galaxy luminosity function with redshift, discussed in Sect. 9.4.1, provides clear evidence that the galaxy population is strongly changing with cosmic epoch. In particular, the z-dependence of the luminosity functions in the UV and the IR shows that the rate at which new stars form in the Universe must be a function of time—such that the average star-formation rate (SFR) was considerably larger at high redshifts. In this section, we consider this evolution of the SFR, together with the evolution of the stellar density. Of course, these two quantities are related: The stellar density at redshift z is the integral of the SFR per unit volume over time, from the earliest epochs to the one corresponding to z. The combination of sensitive space observatories with large ground-based telescopes equipped with modern instruments allows us to trace the SFR up to very high redshifts.

We define the star-formation rate (SFR) as the mass of the stars that form per unit time in a galaxy, typically given in units of $M_\odot/\mathrm{yr}$. For our Milky Way, we find a SFR of $\sim 3 M_\odot/\mathrm{yr}$. Furthermore, we define the star-formation rate density as the mass of stars that are formed per unit time and per unit (comoving) volume, expressed in units of $M_\odot\,\mathrm{yr}^{-1}\,\mathrm{Mpc}^{-3}$.

The importance of the initial mass function. Since the observable signatures for star formation are obtained only from massive stars, their formation rate needs to be extrapolated to lower masses to obtain the full SFR, by assuming an IMF (initial mass function; see Sect. 3.5.4). Typically, a Salpeter-IMF is chosen between $0.1 M_\odot \le M \le 100 M_\odot$. However, there are clear indications that the IMF may be flatter for $M \lesssim 1 M_\odot$ than described by the Salpeter law, and several descriptions for such modified IMFs have been developed over the years, mainly based on observations and interpretation of star-forming regions in our Milky Way or in nearby galaxies. The total stellar mass, obtained by integration over the IMF, is up to a factor of ~ 2 lower in these modified IMFs than for the Salpeter IMF. Thus, this factor provides a characteristic uncertainty in the determination of the SFR from observations; a similar, though somewhat

smaller uncertainty applies to the stellar mass density whose estimation also is mainly based on the more massive stars of a galaxy which dominate the luminosity. Furthermore, the IMF need not be universal, but may in principle vary between different environments, or depend on the metallicity of the gas from which stars are formed. Whereas there has not yet been unambiguous evidence for variations of the IMF, this possibility must always be taken into account.

9.6.1 Indicators of star formation

We will start by discussing the most important indicators of star formation.

Emission in the far infrared (FIR). This is radiation emitted by warm dust which is heated by hot young stars. Observations yield for the approximate relation between the FIR luminosity and the SFR

$$\frac{\mathrm{SFR}_{\mathrm{FIR}}}{M_\odot/\mathrm{yr}} \sim \frac{L_{\mathrm{FIR}}}{5.8\times10^9 L_\odot} \,.$$

For this relation it is assumed that all the energetic photons from newly born hot stars are absorbed locally and heat the dust; more generally, this expression yields the SFR that is dust enshrouded. The wavelength range over which L_{FIR} is determined should be large, covering the two decades from 8 μm to 1 mm, so that this luminosity is essentially independent of the dust temperature. However, in most cases the observation do not cover such a broad spectral region, so that interpolation and extrapolation of the spectral behavior is required to determine L_{FIR}, based on template spectra constructed from very well observed sources. This procedure therefore carries an intrinsic uncertainty in the determination of the SFR. For large samples of sources, one often uses the 24 μm flux to obtain an estimate of L_{FIR}; this wavelength is seen as a good compromise between the need to go to large wavelengths and the decreasing sensitivity and field-of-view of infrared instrumentation. The Spitzer Space Telescope played a very important role in the studies of star formation at these wavelengths.

Radio emission by galaxies. A very tight correlation exists between the radio luminosity of galaxies and their luminosity in the FIR, over many orders of magnitude of the corresponding luminosities. Since L_{FIR} is a good indicator of the star-formation rate, this should apply for radiation in the radio as well (where we need to disregard the radio emission from a potential AGN component). The synchrotron radio emission of normal galaxies originates mainly from relativistic electrons accelerated in supernova remnants (SNRs). Since SNRs appear shortly after the beginning of star formation, caused by core-collapse supernovae at the end of the life of massive stars in a stellar population, radiation from SNRs is a nearly instantaneous indicator of the SFR. Once again from observations, one obtains

$$\frac{\mathrm{SFR}_{1.4\,\mathrm{GHz}}}{M_\odot/\mathrm{yr}} \sim \frac{L_{1.4\,\mathrm{GHz}}}{8.4\times10^{27}\,\mathrm{erg\,s^{-1}\,Hz^{-1}}} \,.$$

Hα emission. This line emission comes mainly from the HII-regions that form around young hot stars with $M \gtrsim 10 M_\odot$. As an estimate of the SFR, one uses

$$\frac{\mathrm{SFR}_{\mathrm{H}\alpha}}{M_\odot/\mathrm{yr}} \sim \frac{L_{\mathrm{H}\alpha}}{1.3\times10^{41}\,\mathrm{erg\,s^{-1}}} \,.$$

For redshifts $z \gtrsim 2$, the observed Hα lines moves into the near-IR part of the spectrum and is thus much more difficult to observe. One can also employ other emission lines, such as Hβ or Lyα. However, whereas Hβ can be observed in the optical to higher redshifts, due to its shorter wavelength, it is a weaker line. For Lyα, the uncertainty to convert line flux into a SFR is much larger, since it is a transition to the ground state of hydrogen (a so-called resonance line). Because of that, a Lyα photon is easily absorbed by neutral hydrogen, which is then excited and reemits a Lyα photon in a random direction. Effectively, thus, this process is a scattering of the photon. In order to leave the interstellar medium of a galaxy, a Lyα photon may be subject to many such scatters, which implies that is path inside the ISM is very much enhanced. This longer path then also increases the probability for the photon to become absorbed by dust. Therefore, the conversion of Lyα flux that leaves a galaxy and the SFR is burdened with a large uncertainty. Alternatively, one can also consider recombination lines of hydrogen coming from transition between higher energy states of the atom, such a Bracket γ, a transition that occurs in the NIR of the spectrum, or transition lines that have millimeter or radio wavelengths.

UV radiation. This is mainly emitted by O and B stars, i.e., by hot young stars thus indicating the SFR in the most recent past, with

$$\frac{\mathrm{SFR}_{\mathrm{UV}}}{M_\odot/\mathrm{yr}} \sim \frac{L_{\mathrm{UV}}}{7.2\times10^{27}\,\mathrm{erg\,s^{-1}\,Hz^{-1}}} \,.$$

This relation assumes that the UV-flux can leave the galaxy without being attenuated by dust absorption (and it neglects that there may be an AGN contribution to the UV luminosity). However, in most galaxies this is not a reasonable assumption, and the observed L_{UV} must be corrected for this effect. Due to the wavelength-dependence of dust absorption, extinction is always connected to reddening, thus affecting the spectral slope of the UV-radiation. One therefore expects

that the redder the UV-spectrum, the larger are the effects of dust obscuration. To quantify this effect, one has to assume an intrinsic, unobscured spectral shape, and to make assumptions about the dust properties, specifically regarding its wavelength dependence of the extinction coefficient (see Fig. 2.6). The unobscured spectral shape can be obtained from the shape of the IMF at the high-mass end (i.e., over that mass region of stars from which the UV-radiation is emitted), whereas there is considerable uncertainty about the variation of dust properties between different objects.

As a result of this procedure, one finds that only a small fraction of UV-photons actually leaves an UV-selected galaxy. A typical value for this escape fraction in a Lyman-break galaxy is ~ 0.2, which means that the observed UV-flux has to be corrected by a factor ~ 5 to obtain the corresponding SFR.

X-ray luminosity. We have seen (Fig. 9.15) that non-active galaxies are X-ray emitters. Most of the X-ray emission is due to high-mass X-ray binaries which are members of a young stellar populations. About 25 % of the X-ray emission from a normal galaxy is due to bremsstrahlung from a hot interstellar medium; since its heating is provided by star-formation activity, it should also scale with the star-formation rate. Hence, if a contribution from an AGN can be excluded, the X-ray luminosity should be a good indicator for the star-formation rate. One finds a tight relation,

$$\frac{\mathrm{SFR}_{\mathrm{X-ray}}}{M_\odot/\mathrm{yr}} \sim \frac{L_{\mathrm{X-ray}}}{3.5\times 10^{39}\,\mathrm{erg\,s^{-1}}}\ ,$$

where the X-ray luminosity is integrated from 0.5 to 8 keV, and where the scatter in this relation is estimated to be less than a factor of 1.5.

Comparison. Applied to individual galaxies, each of these estimates is quite uncertain, which can be seen by comparing the resulting estimates from the various methods (see Fig. 9.54). For instance, Hα and UV photons are readily absorbed by dust in the interstellar medium of the galaxy or in the star-formation regions themselves. Therefore, the relations above should be corrected for this self-absorption, which is possible when the reddening can be obtained from multi-color data. It is also expected that the larger the dust absorption, the stronger the FIR luminosity will be, causing deviations from the linear relation $\mathrm{SFR}_{\mathrm{FIR}} \propto \mathrm{SFR}_{\mathrm{UV}}$. After the appropriate corrections, the values for the SFR derived from the various indicators are quite similar on average, but still have a relatively large scatter.

There are also a number of other indicators of star formation. The fine-structure line of singly ionized carbon at $\lambda = 157.7\,\mu\mathrm{m}$ is of particular importance as it is one of the brightest emission lines in galaxies, which can account for a fraction of up to 1 % of their total luminosity. The large abundance of carbon, together with the fact that this transition can be collisionally excited even at low temperature, result in this line to have a major role in the cooling of the neutral interstellar medium, and thus signifies the presence of star-forming regions. At its wavelength, this line is difficult to observe and until recently has been detected only in star-forming regions in our Galaxy and in other local galaxies. However, more recently this line was detected from the host galaxies of high-redshift QSOs and SMGs, where it shifted into more accessible spectral windows.

9.6.2 Redshift dependence of the star formation: The Madau diagram

The density of star formation, ρ_{SFR}, is defined as the mass of newly formed stars per year per unit (comoving) volume, typically measured in $M_\odot\,\mathrm{yr}^{-1}\,\mathrm{Mpc}^{-3}$. Therefore, ρ_{SFR} as a function of redshift specifies how many stars have formed at any time. By means of the star-formation density we can examine the question, for instance, of whether the formation of stars began only at relatively low redshifts, or whether the conditions in the early Universe were such that stars formed efficiently even at very early times.

Investigations of the SFR in galaxies, by means of the above indicators, and source counts of such star-forming galaxies, allow us to determine ρ_{SFR}. The plot of these results (Fig. 9.55) is sometimes called "Madau diagram". In about 1996, Piero Madau and his colleagues accomplished, for the first time, an estimate of the SFR at high redshifts from Lyman-break galaxies in the Hubble Deep Field North. For these early results, the intrinsic extinction was neglected. In order to correct for this extinction, the progress in FIR and sub-millimeter astronomy was extremely important, as we saw in Sect. 9.3.3.

There is a strong increase in ρ_{SFR} from the current epoch to $z = 1$ by about a factor 10, a further slight increase towards $z \sim 2$, and a decrease at redshifts beyond $z \sim 3$. These results have more recently been confirmed by investigations with the Spitzer and Herschel satellites, observing a large sample of galaxies at FIR wavelengths. Whereas the star-formation rate density at low redshifts is dominated by galaxies which are not very prominent at FIR wavelength, this changes drastically for redshifts $z \gtrsim 0.7$, above which most of the star-formation activity is hidden from the optical view by dust.[8] The increasing importance of dust-obscured star formation is concluded from the very strong redshift

[8] We recall that the roughly equal energy in the optical and FIR extragalactic background radiation shows that about half of the cosmic star formation occurs in dust-obscured regions.

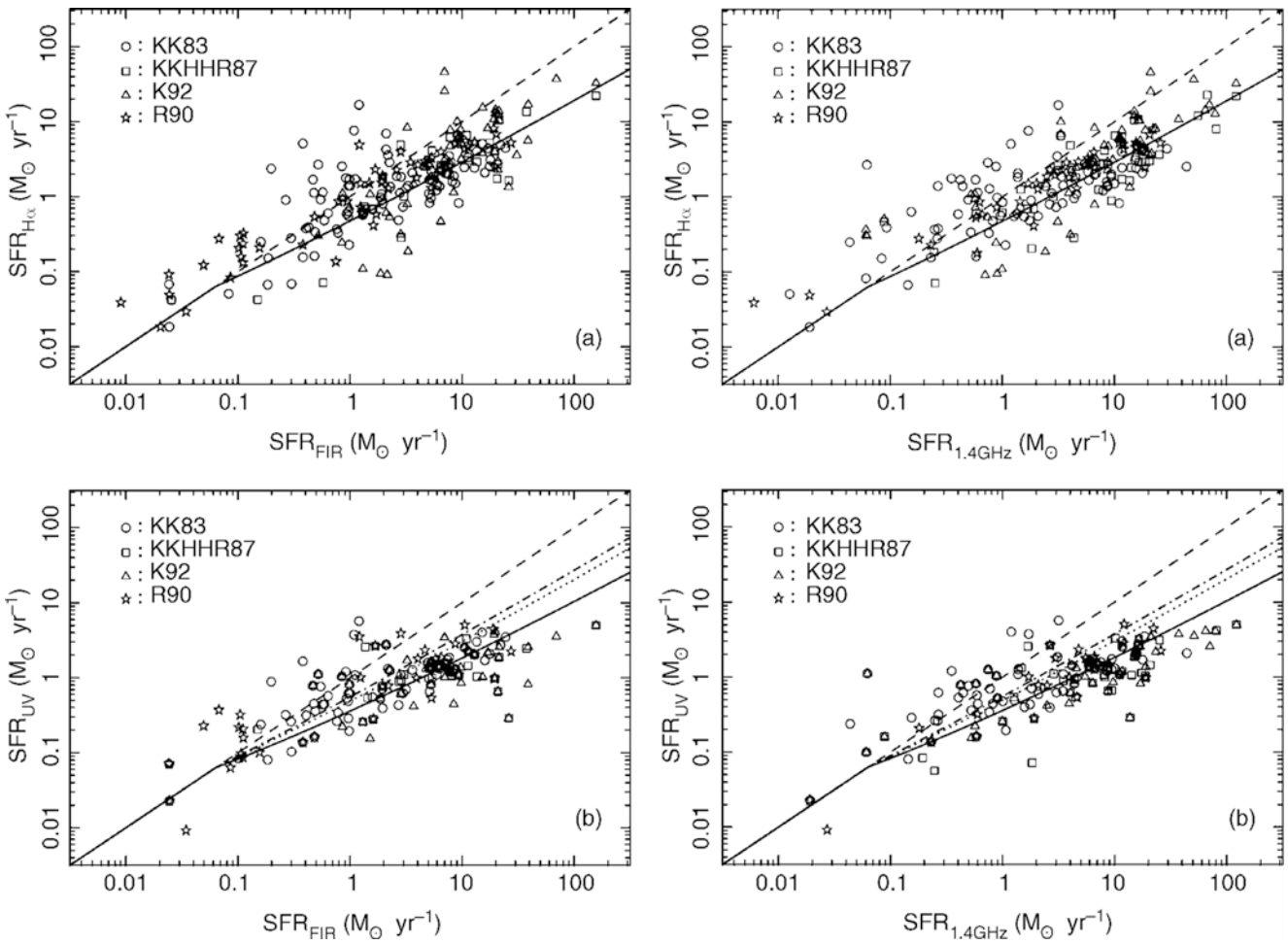


Fig. 9.54 Correlations of the star formation rates in a sample of galaxies, as derived from observation in different wavebands. In all four diagrams, the *dashed line* marks the identity relation $\mathrm{SFR}_i = \mathrm{SFR}_j$; as is clearly seen, using the Hα luminosity and UV radiation as star-formation indicators seems to underestimate the SFR. Since radiation may be absorbed by dust at these wavelengths, and also since the amount of warm dust probably depends on the SFR itself, this effect can be corrected for, as shown by the *solid curves* in the four panels. Source: A.M. Hopkins et al. 2001, *Toward a Resolution of the Discrepancy between Different Estimators of Star Formation Rate*, AJ 122, 288, p. 291, Figs. 2, 3. ©AAS. Reproduced with permission

evolution of the infrared luminosity function of galaxies shown in Fig. 9.43, and the corresponding evolution of the infrared luminosity density (Fig. 9.44).

Integrating the star-formation density over cosmic time, one obtains the stellar mass density as a function of redshift, shown in Fig. 9.56. From this we conclude that most stars in the present-day Universe were already formed at high redshift: star formation at earlier epochs was considerably more active than it is today. Although the redshift-integrated star-formation rate and the mass density of stars determined from galaxy surveys slightly deviate from each other, the degree of agreement is quite satisfactory if one recalls the assumptions that are involved in the determination of the two quantities: besides the uncertainties discussed above in the determination of the star-formation rate, we need to mention in particular the shape of the IMF of the newly formed stars for the determination of the stellar mass density. In fact, Fig. 9.56 shows that we have observed the formation of essentially the complete current stellar density.

The difference between the observed stellar mass density and the one predicted from Fig. 9.55 is largest at high redshifts, which may be due to the uncertainties with which ρ_{SFR} is determined at high redshifts. With the more recent very deep fields observed with HST, the unobscured star-formation rate can be estimated at larger redshifts, based on the luminosity function of Lyman-break galaxies (Fig. 9.41). The result is shown in Fig. 9.57, which shows a steep decline of ρ_{SFR} towards higher redshift.[9] Hence, the bottom line is

[9]The derivation of the star-formation rate as a function of redshift is largely drawn from galaxy surveys which are based on color selection, such as LBGs, EROs and sub-mm galaxies. The possibility cannot be excluded that additional populations of galaxies which are luminous but do not satisfy any of these photometric selection criteria are present at high redshift. Such galaxies can be searched for by spectroscopic

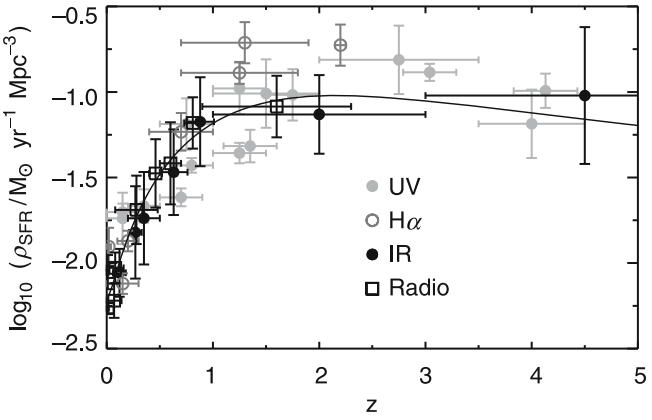


Fig. 9.55 The comoving star-formation density $\rho_{\rm SFR}$ as a function of redshift, where the *different symbols* denote different indicators used for the determination of the star-formation rate. This plot shows the history of star formation in the Universe. Clearly visible is the strong decline for $z < 1$; towards higher redshifts, $\rho_{\rm SFR}$ seems to remain approximately constant. The *curve* is an empirical fit to the data. Source: E.F. Bell 2004, *Galaxy Assembly*, astro-ph/0408023, Fig. 1. Reproduced by permission of the author

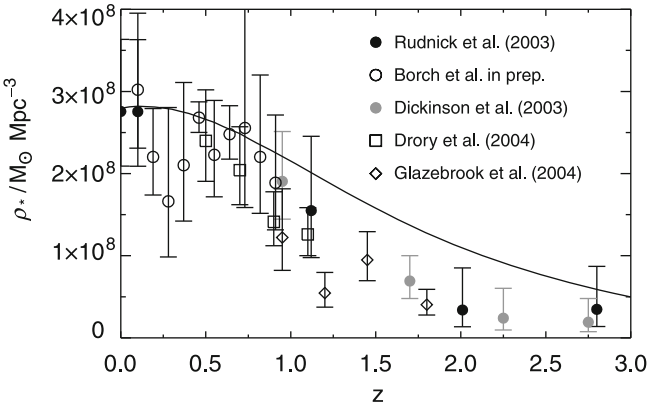


Fig. 9.56 Redshift evolution of the mass density in stars, as measured from various galaxy surveys. The *solid curve* specifies the integrated star-formation density from Fig. 9.55. Source: E.F. Bell 2004, *Galaxy Assembly*, astro-ph/0408023, Fig. 2. Reproduced by permission of the author

that there was an epoch in the Universe, between redshifts ~ 1 and ~ 4, where the star-formation activity was largest. This epoch coincides with the period in our Universe where the QSO activity was highest—indicating that the built-up of the supermassive black hole mass in the Universe happened in parallel to the formation of the stellar population. The close correlation between SMBH mass and properties of the stellar population discussed in Sect. 3.8.3 may thus find a first explanation in this parallel evolution.

Different modes of star formation. Whereas most of the star formation in the local Universe occurs in spiral and irregular galaxies at a modest rate (so-called quiescent star formation), the star-formation activity at higher redshifts was dominated by bursts of star formation, as evidenced in the sub-mm galaxies and in LBGs. At a redshift $z \sim 1$, the latter has apparently ceased to dominate, yielding the strong decline of the star-formation rate density from then until today. This behavior may be expected if bursts of star formation are associated with the merging of galaxies; the merger rate declines strongly with time in models of the Universe dominated by a cosmological constant. This transition may also be responsible for the onset of the Hubble sequence of galaxy morphologies around $z \sim 1$.

The starburst-AGN connection. The just-mentioned coincidence of the 'QSO epoch' with the peak of the star-formation activity can either have a statistical origin, or there can be a connection object-by-object, in the sense that QSO are hosted in galaxies with active star formation. For physical reasons, one would in fact expect such a direct connection, since both processes, star formation and AGN fueling, need a gas supply.

Indeed, in a large fraction of QSOs, clear signs of star-formation activity are found, in some cases at a level where the QSO host galaxy appears as a ULIRG. Conversely, in many of the low-redshift star-forming galaxies, signs of the presence of an AGN are seen. Furthermore, a direct connection between these two processes in a given galaxy does not necessarily have to be observable: it is conceivable that a fresh supply of gas (say, from a major merging of two galaxies) first leads to a strong star-formation activity, and that the accretion onto the central black hole occurs with some delay—or in reverse order.

Statistical studies based on the MIR and FIR emission properties of X-ray selected AGNs suggest that for low-luminosity AGNs, there is no correlation between AGN luminosity and star-formation rate. However, at high AGN luminosity, such a correlation is indeed found, providing

surveys, extending to very faint magnitude limits. This opportunity now arises as several of the 10-m class telescopes are now equipped with high multiplex spectrographs which can thus take spectra of many objects at the same time. One of them is VIMOS at the VLT, another is DEIMOS on Keck. With both instruments, extensive spectroscopic surveys are being carried out on flux-limited samples of galaxies. Among the first results of these surveys is the finding that there are indeed more bright galaxies at redshift $z \sim 3$ than previously found, by about a factor of 2, leading to a corresponding correction of the star-formation rate at high redshifts. In a color-color diagram, these galaxies are preferentially located just outside the selection box for LBGs (see Fig. 9.4). Given that this selection box was chosen such as to yield a high reliability of the selected candidates, it is not very surprising that a non-negligible fraction of galaxies lying outside, but near to it are galaxies at high redshift with similar properties.

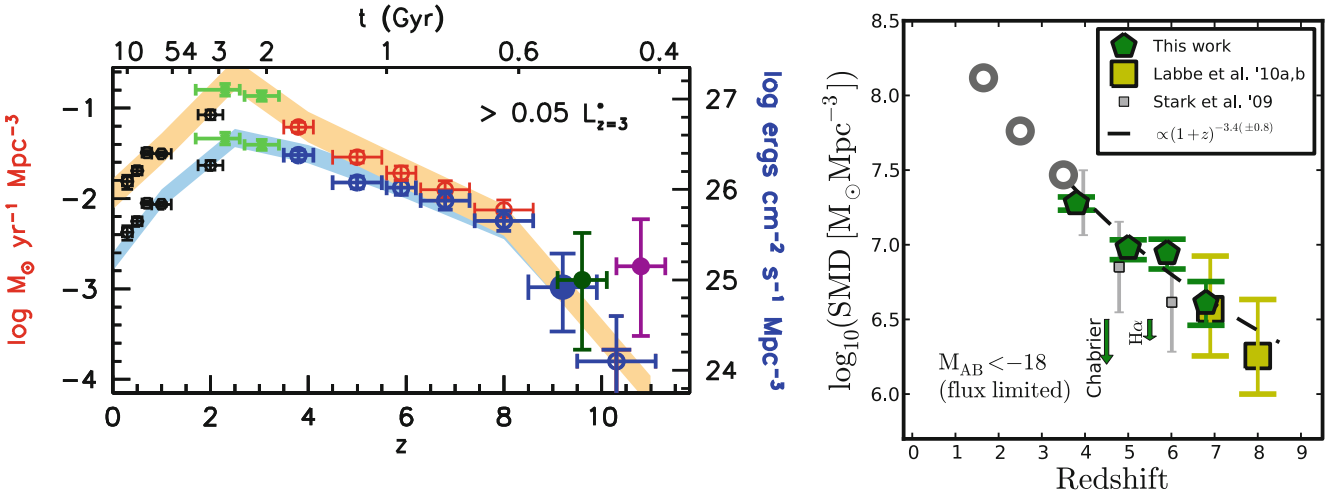


Fig. 9.57 Based on deep HST imaging in the optical and near-infrared, the star-formation density $\rho_{\rm SFR}$ and stellar mass density was estimated to higher redshifts. *Left panel: Blue points* (and right axis) show the UV-luminosity density as a function of redshift, which is proportional to the unobscured star-formation rate density. The *red points* (left axis) show the star-formation rate density where a correction for obscured star formation is included; this correction is assumed to be negligible for $z \gtrsim 7$. The contribution from galaxies with $M_{\rm UV} \leq -18$ in AB magnitudes are added up for these estimates. *Black* and *green points* are estimates of $\rho_{\rm SFR}$ and the UV-luminosity density for $z \leq 3$. The points at $z \sim 10$ is based on LBG candidates detected behind lensing clusters of the CLASH survey. *Right panel:* The stellar mass density as a function of redshift. For high redshifts, ρ_* decreases approximately $\propto (1+z)^{-3.4}$, as indicated by the *dashed grey curve*. Source: *Left*: R.J. Bouwens et al. 2012, *A Census of Star-Forming Galaxies in the* $z \sim 9-10$ *Universe based on HST+Spitzer Observations Over 19 CLASH clusters: Three Candidate* $z \sim 9-10$ *Galaxies and Improved Constraints on the Star Formation Rate Density at* $z \sim 9.2$, arXiv:1211.2230, Fig. 10. *Right*: V. Gonzalez et al. 2011, *Evolution of Galaxy Stellar Mass Functions, Mass Densities, and Mass-to-light Ratios from* $z \sim 7$ *to* $z \sim 4$, ApJ 735, L34, p. 6, Fig. 4. ©AAS. Reproduced with permission

a strong hint for the connection between the built-up of the black hole mass and the stellar population in individual sources.

9.6.3 Summary: High-redshift galaxies

In this chapter, we have considered various aspects of galaxies in the high-redshift Universe. Our discussion of this very quickly evolving field is not complete, but concentrates on some of the central issues. Before we move to another class of high-redshift sources, we want to summarize some of the points mentioned before:

- High-redshift galaxies can be selected by a number of different methods, the most famous one being the Lyman-break technique, other multi-(optical and NIR) band selections, narrow-band imaging targeting highly-redshifted Lyα emitters, mid-infrared selection, and far-infrared/(sub-)millimeter selection. Spectroscopic confirmation of these candidates can be quite challenging, in particular for very dusty galaxies which can be very faint in the optical and NIR spectral regime, and when the source redshift approaches ~ 7, so that the Lyα line is shifted out of the optical window.
- As is true for other situations as well, the properties of the galaxy sample obtained depend on the selection method. A comparison of different samples can therefore be difficult, and must proceed with great care. Lyman-break galaxies at $z \sim 3$ have a stellar mass smaller by a factor ~ 10 than sub-millimeter galaxies, but larger masses than Lyman-alpha emitters.
- The galaxy population at high redshift is distinctly different from that in the current Universe. Most galaxies at $z > 2$ do not fit into Hubble's morphological classification, but show irregular light distributions. The star-formation activity in the Universe was far more intense in the past than it is now. At $z \sim 2.5$, some 10 % of all stars had been formed, and about 50 % of the stars in the local Universe were in place at $z \sim 1$. Correspondingly, the average star-formation rate of distant galaxies is much higher than that of local galaxies. This is reflected in the strong evolution of the galaxy luminosity function in wavebands which strongly respond to the star-formation activity—most notably at mid-IR, far-IR and (sub-)millimeter wavelengths. Similarly, the star-formation rate density is a strongly evolving function of redshift, with a more than tenfold increase between today and $z \sim 1$, an extended period of redshift lasting to $z \sim 3$ or 4, where the star-formation density stays at a high rate, before declining towards even higher redshifts.
- On the other hand, even at $z \sim 2.5$, about half of the most massive galaxies are quiescent, that is, they must have formed their large stellar population at even higher redshifts. From the evolution of the luminosity function

with redshift, it appears that the most massive galaxies formed most of their stars early on, and lower-mass galaxies finish most of their evolution at lower redshifts. This trend has been termed 'downsizing' in the literature.
- The mean metallicity of galaxies evolves with redshift. At a fixed stellar mass, the metallicity of galaxies at $z = 2$ is about smaller by a factor $\sim$ 2 than today, and a further factor of $\sim$ 2 decrease is found at $z \sim 3.5$. On the other hand, the gas of high-redshift QSOs seems to be fairly enriched with metals, approaching Solar metallicity. The dust content of galaxies appears to decrease towards the highest available redshifts, with dust-poor and almost dust-free QSOs detected at $z \sim 6$.
- Except for the CMB, which is a relic of the Big Bang, the radiation in the Universe can be understood by the cumulative emission from active and inactive galaxies in the Universe; there are no clear signs of additional source of the extragalactic background radiation. A large fraction of the background radiation can be resolved into individual sources.

9.7 Gamma-ray bursts

Discovery and phenomenology. In 1967, surveillance satellites for the monitoring of nuclear test ban treaties discovered γ-flashes similar to those that are expected from nuclear explosions. However, these satellites found that the flashes were not emitted from Earth but from the opposite direction—hence, these γ-flashes must be a phenomenon of cosmic origin. Since the satellite missions were classified, the results were not published until 1973. The sources were named *gamma-ray bursts* (GRB).

The flashes are of very different duration, from a few milliseconds up to $\sim$ 100 s, and they differ strongly in their respective light curves (see Fig. 9.58). They are observed in an energy range from $\sim$ 100 keV up to several MeV, sometimes to even higher energies.

The nature of GRBs had been completely unclear initially, because the accuracy with which the location of the bursts was determined by the satellites was totally insufficient to allow an identification of any corresponding optical or X-ray source. The angular resolution of these γ-detectors was many degrees (for some, a 2π solid angle). A more precise position was determined from the time of arrival of the bursts at the location of several satellites, but the error box was still too large to search for counterparts of the source in other spectral ranges.

Early models. The model favored for a long time included accretion phenomena on neutron stars in our Galaxy. If their distance was $D \sim 100$ pc, the corresponding luminosity would be about $L \sim 10^{38}$ erg/s, thus about the Eddington luminosity of a neutron star. Furthermore, indications of absorption lines in GRBs at about 40 and 80 keV were found, which were interpreted as cyclotron absorption corresponding to a magnetic field of $\sim 10^{12}$ Gauss—again, a characteristic value for the magnetic field of neutron stars. Hence, most researchers before the early 1990s thought that GRBs occur in our immediate Galactic neighborhood.

The extragalactic origin of GRBs. A fundamental breakthrough was then achieved with the BATSE experiment on-board the Compton Gamma Ray Observatory, which detected GRBs at a rate of about one per day over a period of 9 years. The statistics of these GRBs shows that GRBs are isotropically distributed on the sky (see Fig. 9.59), and that the flux distribution $N(>S)$ clearly deviates, at low fluxes, from the $S^{-1.5}$-law. These two results meant an end to those models that had linked GRBs to neutron stars in our Milky Way, which becomes clear from the following argument.

Neutron stars are concentrated towards the disk of the Galaxy, hence the distribution of GRBs should feature a clear anisotropy—except for the case that the typical distance of the sources is very small ($\lesssim$100 pc), much smaller than the scale-height of the disk. In the latter case, the distribution might possibly be isotropic, but the flux distribution would necessarily have to follow the Euclidean law $N(>S) \propto S^{-3/2}$, as expected for a homogeneous distribution of sources, which was discussed in Sect. 4.1.2. Because this is clearly not the case, a different distribution of sources is required, hence also a different kind of source.

The only way to obtain an isotropic distribution for sources which are typically more distant than the disk scale-height is to assume sources at distances considerably larger than the distance to the Virgo cluster, hence $D \gg 20$ Mpc; otherwise, one would observe an overdensity in this direction. In addition, the deviation from the $N(>S) \propto S^{-3/2}$-law means that we observe sources up to the edge of the distribution (or, more precisely, that the curvature of spacetime, or the cosmic evolution of the source population, induces deviations from the Euclidean counts), so that the typical distance of GRBs should correspond to an appreciable redshift. This implies that the total energy in a burst has to be $E \sim 10^{51}$ to 10^{54} erg. This energy corresponds to the rest mass Mc^2 of a star. The major part of this energy is emitted within $\sim$ 1 s, so that GRBs are, during this short time-span, more luminous than all other γ-sources in the Universe put together.

We note that the estimated energy of a GRB assumes that the relation between observed flux and luminosity is given by $L = 4\pi D_L^2 S$. This relation is valid only for source which emit isotropically. We have seen that this assumption breaks down for some classes of objects, for example blazars, for which relativistic beaming plays a major role.

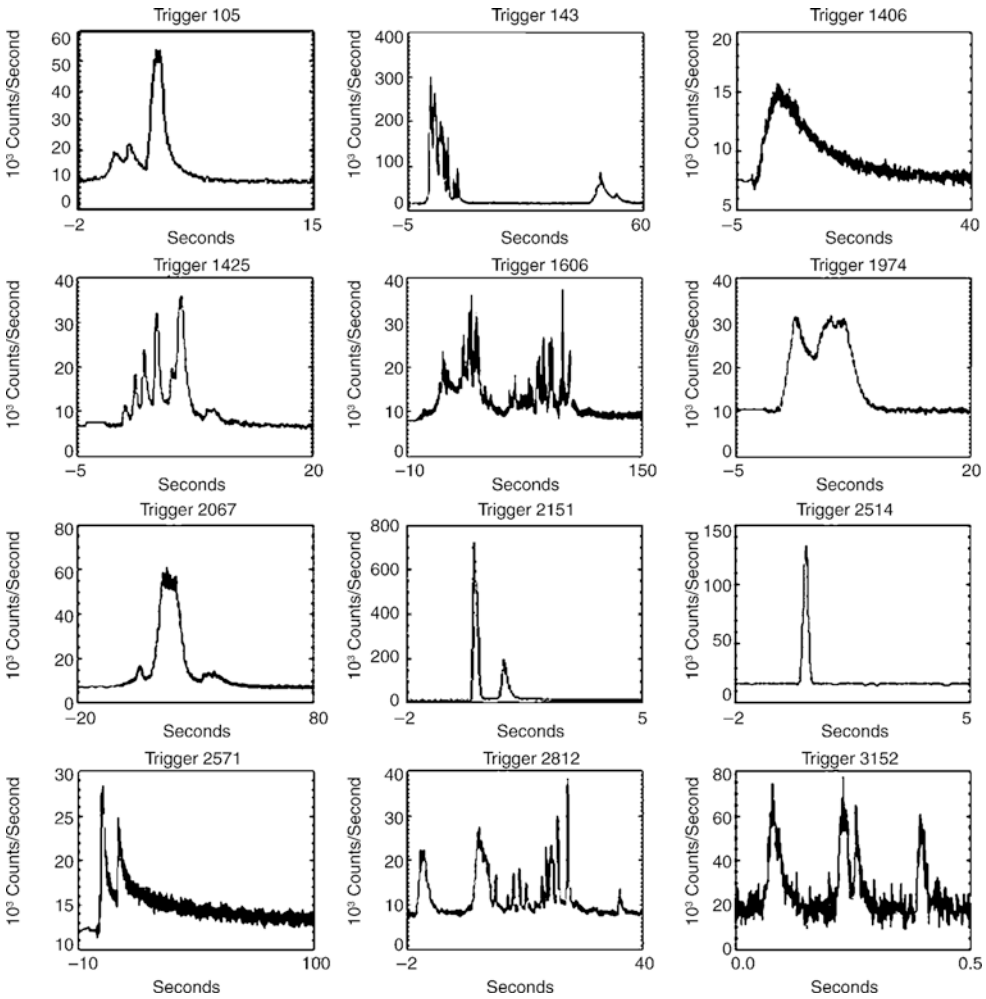


Fig. 9.58 Gamma-ray light curves of various gamma-ray bursts; the different time-scales on the x-axis should be particularly noted. All these light curves appear to be very dissimilar. Credit: J.T. Bonnell, GLAST Science Support Center, NASA/Goddard Space Flight Center, Greenbelt, Maryland, USA

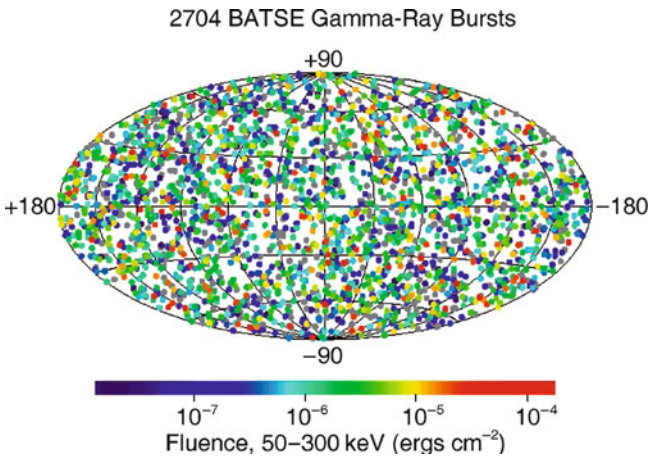


Fig. 9.59 Distribution of gamma-ray bursts on the sphere as observed by BATSE, an instrument on-board the CGRO-satellite, during the about 9 year mission; in total, 2704 GRBs are displayed. The *color of the symbols* represents the observed strength (fluence, or energy per unit area) of the bursts. One can see that the distribution on the sky is isotropic to a high degree. Credit: G. Fishman et al., BATSE, CGRO, NASA

Identification and afterglows. In February 1997, the first identification of a GRB in another wavelength band was accomplished by the X-ray satellite Beppo-SAX. Within a few hours of the burst, Beppo-SAX observed the field within the GRB error box and discovered a transient source, by which the positional uncertainty was decreased to a few arcminutes. In optical observations of this field, a transient source was then detected as well, very accurately defining the position of this GRB. The optical source was identified with a faint galaxy. Optical spectroscopy of the source revealed the presence of absorption features at redshift $z = 0.835$; hence, this GRB must have a redshift equal or larger than this. For the first time, the extragalactic nature of GRBs was established directly. In fast progression, other GRBs could be identified with a transient optical source, and some of them show transient radiation also at other wavelengths, from the radio band up to X-rays. The lower-energy radiation of a GRB after the actual burst in gamma-rays is called an *afterglow*.

With the launch of the SWIFT satellite in November 2004, the observations of GRBs entered a new phase. This satellite is equipped with three instruments: a wide-field gamma-ray telescope to discover the GRBs, an X-ray telescope, and an optical/UV telescope. Within a few seconds of the discovery of a GRB, the satellite targets the location of the burst, so that it can be observed by the latter two telescopes, obtaining an accurate position. This information is then immediately transmitted to the ground, where other telescopes can follow the afterglow emission and obtain spectroscopic information. In its first 8 years of operation, SWIFT discovered some 700 GRBs, of which $\sim$ 200 have their redshift determined. The afterglow could be studied in a much more homogeneous way than before. The prompt γ-ray emission carries about the same amount of total energy as the afterglow emission in the X-ray regime, whereas the total energy of the optical afterglow is smaller by a factor $\sim$ 100.

Relativistic motion. A GRB detected in May 1997 showed an afterglow also at radio frequencies. In the first $\sim$ 20 days, its radio light curve varied erratically, before it settled into a smoother behavior, with flux declining in time. The flux fluctuations in the initial light curve were interpreted as being due to scintillation in the inhomogeneous interstellar medium, very much like the scintillation in the Earth atmosphere.[10] The end of the fluctuating period is then interpreted as being due to the growing size of the emitting source: Just like planets are not scintillating due to their large angular size on the sky, interstellar scintillations are observable only for sufficiently small sources. Hence, these observations provided a clear evidence for an expanding source responsible for the radio afterglow, as well as an estimate of the source size at the end of the scintillation period, of order a few light-weeks. Hence, the expansion velocity of the source must be of order of the speed of light.

There is another independent argument for the presence of relativistic motion in GRBs. The short time-scale of the γ-ray emission, together with its large flux, implies that the density of γ-ray photons in the source must be extremely high. In such a situation, the γ-rays are subject to a large opacity for e^+e^--pair production—in other word, the γ-rays cannot escape from the source region, but are efficiently transformed into pairs. To escape this conclusion, Doppler boosting needs to be employed (cf. Sect. 5.5.2). Allowing for relativistic velocities along our line-of-sight, the radiation density in the source declines significantly. Furthermore, the estimated source size, based on variability argument, increases if Doppler boosting is at work. With a Lorentz factor of the bulk motion of $\gamma = [1-(v/c)^2]^{-1/2} \sim 10^2$ or larger, the pair-production opacity constraints can be avoided, and source sizes of order 10^{13} cm can be accomplished.

Fireball model. Hence, GRBs are associated with a relativistic phenomenon, but the question of their nature still remained unanswered. One model of GRBs quite accurately describes their emission characteristics, including the afterglow. In this fireball model, the radiation is released in the relativistic outflow of electron-positron pairs with a Lorentz-factor of $\gamma \geq 100$. This radiation is not isotropic, but most likely concentrated in a rather narrow beam, resembling the jets in AGNs. In order to form such collimated outflows, one needs a strong energy source, and presumably strong rotation whose rotation axis defines the preferred directions into which the jets flow. To collimate the jets, the presence of magnetic fields are probably also required.

Short vs. long-duration bursts. GRBs can be broadly classified into short- and long-duration bursts, with a division at a duration of $t_{\rm burst} \sim 2$ s. The spectral index of the short-duration bursts is considerably harder at γ-ray energies than that of long-duration bursts. Until 2005, only afterglows from long-duration bursts had been discovered. Long-duration bursts typically occur in galaxies at high redshift, with a mean of $z \sim 2.5$. Also GRBs with very high redshift were discovered, with at least three having redshifts $z > 6$. One GRB redshift of $z = 8.2$ has been spectroscopically obtained, and there are indications that an even higher-redshift burst was observed. In one case, an optical burst was discovered about 30 seconds after the GRB, with the fantastic brightness of $V \sim 9$, at a redshift of $z = 1.6$. For a short period of time, this source was apparently more luminous than any quasar in the Universe. In March 2008, a GRB at $z = 0.937$ occurred which has a peak optical brightness of $m = 5.7$—i.e., this source was visible for a very short period to the naked eye (it is not known, though, whether anyone peeked at the right position of the sky at that moment). Thus, during or shortly after the burst at high energies, GRBs can also be *very* bright in the optical.

Counterparts of long-duration GRBs: Hypernovae. In April 1998, the positional error box of a GRB contained a supernova, hinting for a possible connection. This has been verified subsequently, by finding that the light-curve of some optical afterglows were described by the sum of a declining power law in time plus the light-curve of a luminous supernova. For a GRB in March 2003, the presence of a supernova in the spectrum of the optical afterglow was identified, proving the direct connection between SNe and GRBs. Since most of the GRBs are located at high redshifts,

[10] Recall that atmospheric scintillations are due to a space and time dependent refractive index of the air. For propagating radio waves, the same is true, except that the refractive index here is determined by the electron density of the ionized plasma in the ISM.

the corresponding SN cannot be identified for them, but for more nearby long-duration bursts, the association is clearly established.

Long-duration GRBs are located in star-forming regions of galaxies, and their redshift distribution is similar to that of the star-formation rate density in the Universe.[11] This observation yields a close connection of the GRB phenomenon to star formation, and thus the associated supernovae are due to young massive stars. Not every core-collapse SN yields a GRB, though. The current picture is that GRBs are produced in the core-collapse process of very massive stars, giving rise to extraordinarily energetic explosions, so-called hypernovae. The combination of stellar rotation and an internal magnetic field can form a highly relativistic bi-directional outflow after the collapse event, when the stellar material falls onto the newly formed compact remnant, a black hole. Even if the emission is highly anisotropic, as expected from the fireball model, the corresponding energy released by the hypernovae is very large.

Counterparts of short-duration GRBs. SWIFT has allowed the identification of afterglow emission from short-duration GRBs. In contrast to the long-duration bursts, some of these seem to be associated with elliptical galaxies; this essentially precludes any association with (core-collapse) supernova explosions. In fact, for one of these short burst, very sensitive limits on the optical brightness explicitly rules out any contribution from a supernova explosion. Furthermore, the host galaxies of short bursts are at substantially lower redshift, $z \lesssim 0.5$. Given that both kinds of GRBs have about the same observed flux (or energy), this implies that short-duration bursts are less energetic than long-duration ones, by approximately two orders of magnitude. All of these facts clearly indicate that short- and long-duration GRBs are due to different populations of sources. The lower energies of short bursts and their occurrence in early-type galaxies with old stellar populations are consistent with them being due to the merging of compact objects, either two neutron stars, or a neutron star and a black hole.

[11] Indeed, it seems that the distribution of GRBs extends further out in redshift than that of the star formation density. This observational fact is most likely related to the finding that GRBs are found in host galaxies with small metallicity. It is possible that the metal enrichment of galaxies suppresses GRBs at later redshifts. The connection to the metallicity may have its origin on a possible metallicity-dependent star formation, i.e., allowing for higher-mass stars from metal-poor gas.

Galaxy evolution

10

After having described the cosmological model in great detail, as well as the objects that inhabit our Universe at low and high redshifts, we will now try to understand how these objects can be formed and how they evolve in cosmic time.

The extensive results from observations of galaxies at high redshift which were presented earlier might suggest that the formation and evolution of galaxies is quite well understood today. We are able to examine galaxies at redshifts up to $z \sim 7$ (and find plausible candidates at even higher redshifts) and therefore observe galaxies at nearly all epochs of cosmic evolution. This seems to imply that we can study the evolution of galaxies directly. However, this is true only to a certain degree. Although we observe the galaxy population throughout 90 % of the cosmic history, the relation between galaxies at different redshifts is not easily understood. We cannot suppose that galaxies seen at different redshifts represent various subsequent stages of evolution of the same kind of galaxy. The main reason for this difficulty is that different selection criteria need to be applied to find galaxies at different redshifts.

We shall explain this point with an example. Actively star-forming galaxies with $z \gtrsim 2.5$ are efficiently detected by applying the Lyman-break criterion, but only those which do not experience much reddening by dust. Actively star-forming galaxies at $z \sim 1$ are discovered as extremely red objects (EROs) if they are sufficiently reddened by dust, and at $z \sim 2.5$ as sub-millimeter galaxies. The relation between these galaxy populations depends, of course, on how large the fraction of galaxies is whose star-formation regions are enshrouded by dense dust. To determine this fraction, one would need to find Lyman-break galaxies (LBGs) at $z \sim 1$, or EROs at $z \sim 3$. Both observations are very difficult today, however. For the former, this is because the Lyman break is then located in the UV domain of the spectrum and thus can not be observed with ground-based telescopes. For the latter it is because the rest wavelength corresponding to the observed R-band lies in the UV where the emission of EROs is very small, so that virtually no optical radiation from such objects would be visible, rendering spectroscopy of these objects impossible. In addition to this, there is the problem that galaxies with $1.3 \lesssim z \lesssim 2.5$ are difficult to discover because, for objects at those redshifts, hardly any spectroscopic indicators are visible in the optical range of the spectrum—both the 4000 Å-break and the $\lambda = 3727$ Å line of [OII] are redshifted into the NIR, as are the Balmer lines of hydrogen, whereas the Lyman lines of hydrogen are located in the UV part of the spectrum. For these reasons, this range in redshift is also called the 'redshift desert'.[1] Thus, it is difficult to trace the individual galaxy populations as they evolve into each other at the different redshifts. Do the LBGs at $z \sim 3$ possibly represent an early stage of today's ellipticals (and the passive EROs at $z \sim 1$), or are they an early stage of spiral galaxies? Or do some galaxies form the bulk of their stellar population at $z \sim 3$, whereas others do it at some later epoch?

The difficulties just mentioned are the reasons why our understanding of the evolution of the galaxy population is only possible within the framework of models, with the help of which the different observational facts are being interpreted. We will discuss some aspects of such models in this chapter.

Another challenge for galaxy evolution models are the observed scaling relations of galaxy properties. We expect that a successful theory of galaxy evolution can predict the Tully–Fisher relation for spiral galaxies, the fundamental plane for ellipticals, as well as the tight correlation between galaxy properties and the central black hole mass. This latter point also implies that the evolution of AGNs and galaxies must be considered in parallel, since the growth of black holes with time is expected to occur via accretion, i.e., during phases of activity in the corresponding galaxies. The hierarchical model of structure formation implies that high-mass galaxies form by the merging of smaller ones

[1] Spectroscopy in the NIR is possible in principle, but the high level of night-sky brightness and, in particular, the large number of atmospheric transition lines renders spectroscopic observations in the NIR much more time consuming than optical spectroscopy.

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_10,

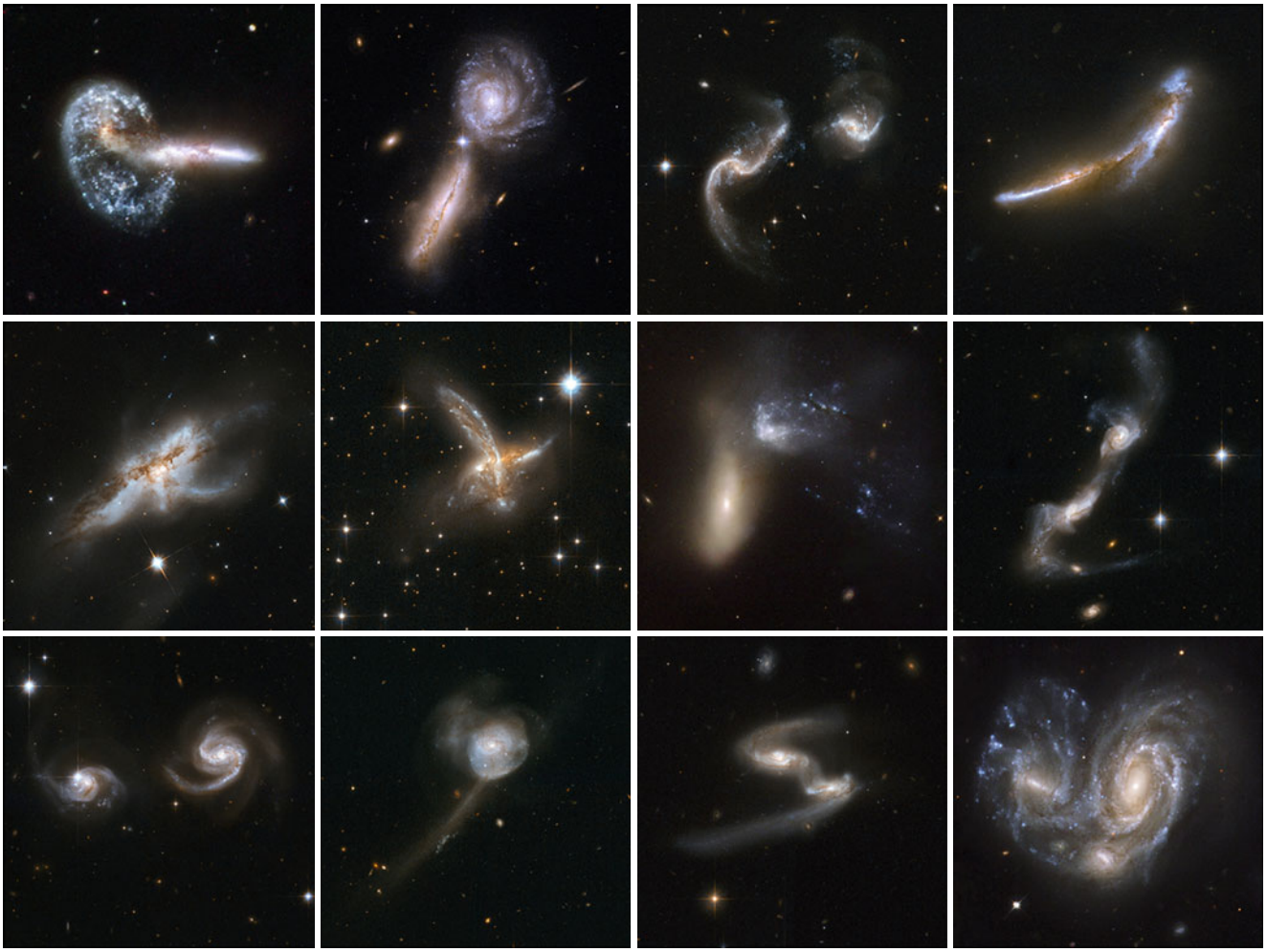


Fig. 10.1 A collection of interacting and peculiar galaxies, as obtained by the Hubble Space Telescope. Such interactions and mergers are partly responsible for the formation of the current population of galaxies. *Top row*: Arp 148, UGC 9618, Arp 256, NGC 6670. *Middle row*: NGC 6240, ESO 593-8, NGC 454, UGC 8335. *Bottom row*: NGC 6786, NGC 17, ESO 77-14, NGC 6050. Credit: NASA, ESA, the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration, and A. Evans (University of Virginia, Charlottesville/NRAO/Stony Brook University)

(Fig. 10.1); if the aforementioned scaling relations apply at high redshifts (and there are indications for this to be true, although with redshift-dependent pre-factors that reflect the evolution of the stellar population in galaxies), then the merging process must preserve the scaling laws, at least on average.

10.1 Introduction and overview

Key questions. In this final chapter we shall outline some of the current ideas on the formation and evolution of galaxies, their large-scale environment and their central black holes. We start with a list of questions a successful model is expected to provide answers for:

- Why are galaxies the dominant objects in the Universe? We have seen that most of the stars live in galaxies with a luminosity which lies within a factor of $\sim$ 10 of the characteristic luminosity L^* of the Schechter function [(3.52); see also (3.59)]; what defines this characteristic luminosity (and mass) scale?
- Can the model of structure evolution in the Universe, which is based on gravitational instability and mainly driven by dark matter inhomogeneities, explain the formation of galaxies?
- What is the reason for the existence of two main galaxy populations, the early-types or ellipticals, and the late-type or spirals? Do they have a different evolutionary history? Can we actually understand the relative abundance of these two populations, and even their distribution in luminosity or stellar mass?
- Why is the shape of the galaxy luminosity function different from the shape of the dark matter halo mass function (see Fig. 10.2)? In other words, what causes the different mass-to-light ratios, or stellar-to-total mass ratios, of halos?

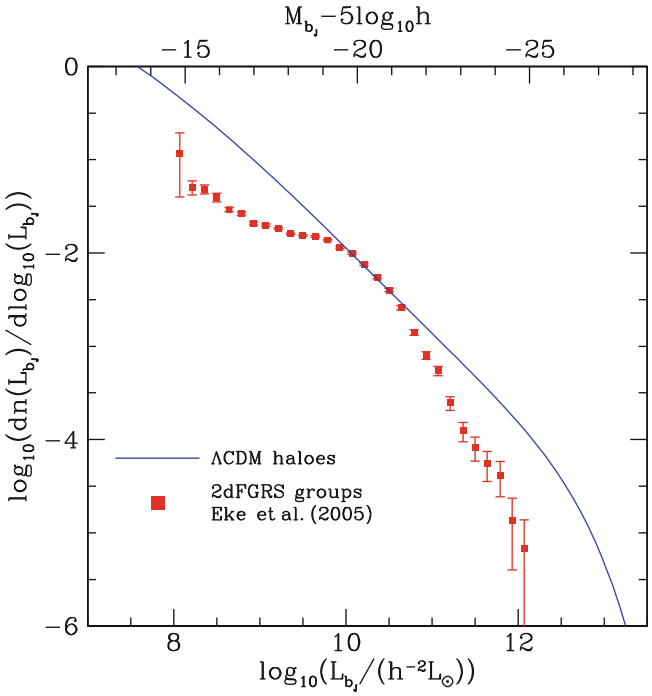


Fig. 10.2 The *red points* with error bars show the luminosity function of galaxies and groups of galaxies as measured from the two-degree field galaxy redshift survey. In comparison, the *solid curve* shows the abundance of dark matter halos as predicted in a ΛCDM model, assuming a fixed mass-to-light ratio. This mass-to-light ratio is chosen such that the curves touch at one point, yielding $M/L \sim 80hM_\odot/L_\odot$. The corresponding halo mass is $M \sim 10^{12}h^{-1}M_\odot$. There is an obvious discrepancy between the shape of the observed luminosity function and that expected if all halos had the same mass-to-light ratio. This implies that halos of different mass have different efficiency with which baryons are converted into stars. In other words, the mass-to-light ratio is smallest for halos of mass $\sim 10^{12}h^{-1}M_\odot$, and the efficiency of turning baryons into stars is suppressed for higher and lower mass halos. Source: C.M. Baugh 2006, *A primer on hierarchical galaxy formation: the semi-analytical approach*, arXiv:astro-ph/0610031, Fig. 6. Reproduced by permission of the author

- The properties of the galaxy population and its relative abundance depend on the environment. Why are red galaxies the dominant population in high-density regions such as clusters, whereas blue galaxies dominate the field population?
- Can we understand the dependence of the star-formation rate on redshift, such as is displayed in the Madau diagram (see Sect. 9.6.2)? Why has the star-formation activity declined so strongly over the past five billion years?
- Why is the mass of the supermassive black hole in the center of galaxies so tightly linked to the stellar properties of the galaxies? Which mechanism yields the co-evolution of the mass of the central black hole and the growth of the stellar mass? Do the stellar properties determine the mass of the black hole, or reversely, does the black hole affect the evolution of the stellar population—or are both of them jointly affected by the processes of galaxy growth?
- What is the role of active galactic nuclei in the evolution of the galaxy population? Why are some galaxies very active, some are not, and why is the fraction of active galaxies such a strong function of redshift?
- How are the special kinds of galaxies seen at high redshift related to the local galaxy population? What is the fate of an object which we can see as sub-millimeter galaxy at $z \sim 2$, or a Lyman-break galaxy at $z \sim 3$—into what kind of object have they developed?
- Why are galaxies at high redshift significantly different from local ones, in terms of size and morphology?

As we shall see, for most of these question plausible answers can be given in the framework of the cosmological model. The panchromatic view of cosmic sources, provided to us by a suite of superb telescopes and instruments, allows us to link together the evidence about the physical nature of objects obtained from very different wavelengths. These observational results are used to build models of the evolution of galaxies which attempt to account for as many of them as possible. These models differ in detail, but we currently have a rather coherent picture of the key features that govern the formation and evolution of galaxies, although many important issues are still to be clarified.

Overview. The evolution of structure in the Universe is seeded by density fluctuations which at the epoch of recombination can be observed through the CMB anisotropy. Hence, we have direct observational evidence about the fluctuation spectrum at $z \sim 1100$. Cosmological N-body simulations predict the evolution of the dark matter distribution as a function of redshift, in particular the formation of halos and their merger processes. Before recombination, baryons were coupled tightly to the photons and thus subject to a strong pressure which prevented them to fall into the potential wells formed by dark matter inhomogeneities (see Sect. 7.4.3). After recombination the baryonic matter decoupled from radiation, became essentially pressure free, and soon followed the same spatial distribution as the dark matter. However, baryonic matter is subject to physical processes like dissipation, friction, heating and cooling, and star formation. Since dark matter is not susceptible to these processes, the behavior of baryons and the dark matter is expected to differ in the ongoing evolution of the density field.

In the cold dark matter universe, small density structures formed first, which means that low-mass dark matter halos preceded those of higher mass. This 'bottom-up' scenario of structure formation follows from the shape of the power spectrum of density fluctuations, which itself is determined by the nature of dark matter—namely cold dark matter. The gas in these halos is compressed and heated, the source of heat being the potential energy. If the gas is able to cool by radiative processes, i.e., to get rid of some of its

thermal energy and thus pressure, it can collapse into denser structures, and eventually form stars. In order for this to happen, the potential wells have to have a minimum depth, so the resulting kinetic energy of atoms is sufficient to excite the lowest-lying energy levels whose de-excitation then leads to the emission of a photon which yields the radiative cooling. We shall see that this latter aspect is particularly relevant for the first stars to form, since they have to be made of gas of primordial composition, i.e., only of hydrogen and helium.

Once the first stars form in the Universe, the baryons in their cosmic neighborhood get ionized. This reionization at first happens locally around the most massive dark matter halos that were formed; lateron, the individual ionized regions begin to overlap, the remaining neutral regions become increasingly small, until the process of reionization is completed, and the Universe becomes largely transparent to radiation, i.e., photons can propagate over large distances in the Universe. The gas in dark matter halos is denser than that in intergalactic space; therefore, the recombination rate is higher there and the gas in these halos is more difficult to ionize. Probably, the ionizing intergalactic radiation has a small influence on the gas in halos hosting a massive galaxy. However, for lower-mass halos, the gas not only maintains a higher ionization fraction, but the heating due to ionization can be appreciable. As a result, the gas in these low-mass halos finds it more difficult to cool and to form stars. Thus, the star-formation efficiency—or the fraction of baryons that is turned into stars—is expected to be smaller in low-mass galaxies.

The mass of halos grows, either by merging processes of smaller-mass halos or by accreting surrounding matter through the filaments of the large-scale density field. The behavior of the baryonic matter in these halos depends on the interplay of various processes. If the gas in a halo can cool, it will sink towards the center. One expects that the gas, having a finite amount of angular momentum like the dark matter halo itself, will initially accumulate in a disk perpendicular to the angular momentum of the gas, as a consequence of gas friction—provided a sufficiently long time of quiescent evolution for this to happen. The gas in the disk then reaches densities at which efficient star formation can set in. In this way, the formation of disk galaxies, thus of spirals, can be understood qualitatively.

As soon as star formation sets in, it has a feedback on the gas: the most massive stars very quickly explode as supernovae, putting energy into the gas and thereby heating it. This feedback then prevents that all the gas turns into stars on a very short time-scale, providing a self-regulation mechanism of the star-formation rate. In the accretion of additional material from the surrounding of a dark matter halo, also additional gas is accreted as raw material for further star formation.

When two dark matter halos with their embedded galaxies merge, the outcome depends mainly on the mass ratio of the halos: if one of them is much lighter than the other, its mass is simply added to the more massive halo; the same is true for their stars. More specific, the small-mass galaxy is disrupted by tidal forces, in the same way as the Sagittarius dwarf galaxy is currently destroyed in our Milky Way, with the stars being added to the Galactic halo. If, on the other hand, the masses of the two objects are similar, the kinematically cold disks of the two galaxies are expected to be disrupted, the stars in both objects obtain a large random velocity component, and the resulting object will be kinematically hot, resembling an elliptical galaxy. In addition, the merging of gas-rich galaxies can yield strong compression of the gas, triggering a burst of star formation, such as we have seen in the Antennae galaxies (see Fig. 9.25). Merging should be particularly frequent in regions where the galaxy density is high, in galaxy groups for instance. From the example shown in Fig. 6.68, a large number of such merging and collision processes are detected in galaxy clusters at high redshift.

In parallel, the supermassive black holes in the center of galaxies must evolve, as clearly shown by the tight scaling relations between black hole mass and the properties of the stellar component of galaxies (Sect. 3.8.3). The same gas that triggers star formation, say in galaxy mergers, can be used to 'feed' the central black hole. If, for example, a certain fraction of infalling gas is accreted onto the black hole, with the rest being transformed into stars, the parallel evolution of black hole mass and stellar mass could be explained. In those phases where the black hole accretes, the galaxy turns into an active galaxy; energy from the active galactic nucleus, e.g., in the form of kinetic energy carried by the jets, can be transmitted to the gas of the galaxy, thereby heating it. This provides another kind of feedback regulating the cooling of gas and star formation.

When two halos merge, both hosting a galaxy with a central black hole, the fate of the black holes needs to be considered. At first they will be orbiting in the resulting merged galaxy. In this process, they will scatter off stars, transmitting a small fraction of their kinetic energy to these stars. As a result, the velocity of the stars on average increases and many of them will be 'kicked out' of the galaxy. Through these scattering events, the black holes lose orbital energy and sink towards the center of the potential. Finally, they form a tight binary black hole system which loses energy through the emission of gravitational waves (see Sect. 7.9), until they merge. With the planned space-based laser interferometer LISA, one expects to detect these coalescing black hole events almost throughout the observable Universe.

The more massive halos corresponding to groups and clusters only form in the more recent cosmic epoch. In those

regions of space where at a later cosmic epoch a cluster will form, the galaxy-mass halos form first—the larger-scale overdensity corresponding to the proto-cluster promotes the formation of galaxy-mass halos, compared to the average density region in the Universe; this is the physical origin of galaxy bias. Therefore, one expects the oldest massive galaxies to be located in clusters nowadays, explaining why most massive cluster galaxies are red. In addition, the large-scale environment provided by the cluster affects the evolution of galaxies, e.g., through tidal stripping of material.

In the rest of this chapter, we will elaborate on the various processes which are essential for our understanding of galaxy formation and evolution. In Sect. 10.2 we study the behavior of gas in a dark matter halo, in particular consider its heating and cooling properties; the latter obviously is most relevant for its ability to form stars. We then turn in Sect. 10.3 to the first generation of stars and consider their ability to reionize the Universe; we will also briefly discuss observational evidence for approaching the reionization epoch for the highest redshift objects known.

The formation of disk and elliptical galaxies is studied in Sects. 10.4 and 10.5, respectively. Here we will stress the importance of cooling processes on the one hand, and feedback processes that leads to gas heating on the other hand. We will also discuss the impact of mergers on the evolution of galaxies, the evolution of supermassive black holes, and the fate of these black holes in the aftermath of mergers. The final two sections are dedicated to modeling the formation and evolution of galaxies, both in the framework of numerical simulations which include the properties of the baryons (Sect. 10.6), and with somewhat simplified 'semi-analytic' models (Sect. 10.7) which, due to their great flexibility, have guided much of our understanding of galaxy evolution over the past two decades.

10.2 Gas in dark matter halos

We have seen in Sect. 7.5.1 how density fluctuations in the dark matter distribution evolve into gravitationally bound and virialized systems, the dark matter halos, through the process of gravitationally instability. In order to understand the formation of galaxies, we need to study the behavior of the baryons in these dark matter halos—the baryons out of which stars form.

10.2.1 The infall of gas during halo collapse

Gas heating. As long as the fractional overdensities are small, the spatial distribution of baryons and dark matter are expected to be very similar. In the language of the spherical collapse model, initially the radial distribution of an overdense sphere is the same for dark matter and baryons, scaled by their different mean cosmic density. However, when the sphere collapses, the behavior of both components must be very different: dark matter is collisionless, and the dark matter particles can freely propagate through the density distribution, crossing the orbits of other particles. Baryons, on the other hand, are collisional, which means that friction prevents gas from crossing through a gas distribution. Thus, as the halo collapses, the potential energy of the gas is transformed into heat through the frictional processes. Furthermore, the pressure of the gas can prevent it from falling into the dark matter potential well, depending on the gas temperature and the depth of the potential well, i.e., the halo mass. As we shall see below, this pressure effect is important for low-mass halos at high redshifts. But first, we assume that the gas initially is sufficiently cold such that this effect can be neglected in the halo collapse.

In the case of (approximate) spherical symmetry, one can picture this as follows: In the inner part of the halo, gas has already settled down into a quasi-hydrostatic state, where gas pressure balances the gravitational force. As the outer part of the halo collapses, gas falls onto this gas distribution. The infall speed is much higher than the sound velocity of the (cold) infalling gas, i.e., the gas falls in supersonically. This is the situation in which a shock front develops, i.e., a zone in which gas density, pressure, and velocity varies rapidly with position and in which the dissipation of kinetic energy (given by the infall velocity) into heat occurs. Inside this shock front, the gas is hot, and (almost) all of its kinetic energy gets converted into heat.

Virial temperature. We can now calculate the temperature of the gas inside a halo of (total) mass M. For that, we assume that the gas temperature T_g is uniform. According to the virial theorem, half of the potential energy of the infalling gas is converted into kinetic energy, which in turn is transformed into heat. We can therefore equate the thermal energy per unit volume to one half of the potential energy of the gas per unit volume,

$$\frac{3}{2} n k_\mathrm{B} T_\mathrm{g} = \frac{3}{2} \frac{\rho_\mathrm{g} k_\mathrm{B} T_\mathrm{g}}{\mu m_\mathrm{p}} = \frac{\nu}{2} \rho_\mathrm{g} \frac{GM}{r} \,, \tag{10.1}$$

where μm_p is the mean mass per particle in the gas, and the factor $\nu \sim 1$ depends on the assumed density profile of the halo of mass M and radius r. Note that the final term in (10.1) is just the square of the circular velocity, V_c^2. Ignoring factors of order unity, the gas temperature will thus be similar to the *virial temperature* T_vir, defined as

$$T_\mathrm{vir} := \frac{\mu m_\mathrm{p}}{2 k_\mathrm{B}} V_\mathrm{c}^2 \approx 3.6 \times 10^5 \mathrm{K} \left(\frac{V_\mathrm{c}}{100\,\mathrm{km/s}} \right)^2 . \tag{10.2}$$

Thus, a collapsed halo contains hot gas, with a temperature depending on the halo radius and mass. Such a hot gas is seen in galaxy groups and clusters though its X-ray emission (see Sect. 6.4). For galaxy-mass halos, this hot gas is much more difficult to observe: (10.2) predicts a characteristic gas temperature for galaxy-mass halos of $\sim 10^6$ K. At these temperatures, gas is very difficult to observe. The temperature is too low for being observable in X-rays—the corresponding X-ray energies are ~ 0.1 keV, for which the interstellar medium of the Milky Way is essentially opaque. Furthermore, at these temperatures most atoms are fully ionized, so there is little diagnostics of this gas from optical or UV line radiation. Nevertheless, some highly (but not fully) ionized species (such as five times ionized oxygen) exist, and their presence can be seen through absorption lines, e.g., in the spectrum of quasars. In this way, the presence of hot gas surrounding our Milky Way has been established. Nevertheless, some significant fraction the hot gas in galaxy-mass halos does not stay hot, but must cool, otherwise stars can not form. We shall turn to cooling processes next.

10.2.2 Cooling of gas

In order to form stars in a halo, the gas needs to compress to form dense clouds in which star formation occurs. The pressure of the hot gas prevents gas from condensing further, unless the gas can cool and thereby, at fixed pressure, increases its density.

Cooling processes. Optically thin gas can cool by emitting radiation—i.e., it gets rid of some of its energy in form of photons. There are several relevant processes by which internal energy can be transformed into radiation. In an ionized gas, the scattering between electrons and nuclei causes the emission of bremsstrahlung (free-free emission), as we discussed in Sect. 6.4.1. Collisions between atoms and electrons can lead to a transition of an atom into an excited state (collisional excitation). When the excited state decays radiatively, the energy difference between the ground level and the excited state is radiated away. Collisions can also lead to (partial) ionization of atoms, and subsequent recombination is again related to the emission of photons.

Cooling function. Common to all these processes is that they depend on the square of the gas density: they all are two-body processes due to collisions of particles. If we define the cooling rate C as the energy radiated away per unit volume and unit time, then $C \propto n_{\rm H}^2$, with $n_{\rm H}$ being the number density of hydrogen nuclei (i.e., the sum of neutral and ionized hydrogen atoms). The constant of proportionality is called the *cooling function*, defined as

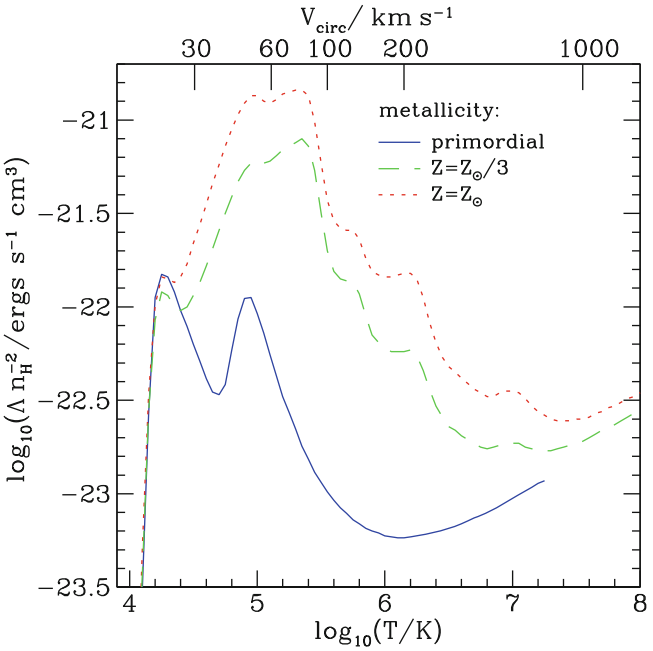


Fig. 10.3 The cooling function for gas with primordial composition (*blue solid curve*), 1/3 of the Solar metallicity (*green dashed curve*) and Solar metallicity (*red dotted curve*). On the top axis, the temperature is converted into a circular velocity, according to (10.2). To obtain such a cooling function, one needs to assume an equilibrium state of the gas. Here it is assumed that the gas is in thermodynamical equilibrium, where the fraction of ionization states of any atom depends just on T. The total cooling function shown here is a superposition of different gas cooling processes, including atomic processes and bremsstrahlung, the latter of which dominating at high T where the gas is fully ionized. Source: C.M. Baugh 2006, *A primer on hierarchical galaxy formation: the semi-analytical approach*, arXiv:astro-ph/0610031, Fig. 9. Reproduced by permission of the author

$$\Lambda(T) := \frac{C}{n_{\rm H}^2} \,, \tag{10.3}$$

which depends on the gas temperature and its chemical composition. Figure 10.3 shows the cooling function for three different values of the gas metallicity.

The relative importance and efficiency of the various cooling processes depend on the density and temperature of the gas, as well as on its chemical composition. At very high temperatures, all atoms are fully ionized, and thus the processes of collisional excitation and ionization are no longer of relevance. Then, bremsstrahlung becomes the dominant effect, with $\Lambda(T) \propto T^{1/2}$ [see also (6.32)]. This behavior is seen in Fig. 10.3 for a pure hydrogen plus helium gas at $T \gtrsim 10^6$ K; for gas with non-zero metallicity, bremsstrahlung starts to dominate the cooling at somewhat higher temperatures.

For gas with primordial abundance, we see two clear peaks in the cooling function in Fig. 10.3, one at $T \sim 2 \times 10^4$ K, the other at $T \sim 10^5$ K. The former one is due to

the fact that for gas at this temperature, hydrogen is mostly neutral, and many particles in the gas have an energy sufficient for the excitation of higher energy levels in hydrogen atoms; note that the lowest lying excited state of hydrogen has an energy corresponding to the Lyman-α transition, i.e., 10.2 eV, corresponding to a temperature of $T \sim 10^5$ K. Thus, collisional excitation is efficient. At slightly higher temperatures, also collisional ionization (and subsequent recombination) is very effective, but with increasing T, the cooling function drops, because then hydrogen becomes mostly ionized. The second peak has the same origin, except that now the helium atom is the main coolant. Since the lowest energy level and the ionization energy in helium is higher than for hydrogen, the helium peak is simply shifted. Once helium is fully ionized, atomic cooling shuts off, and only at higher temperatures the bremsstrahlung effect takes over.

Although elements heavier than helium have a small abundance in number, they can dominate the gas cooling, due to the rich energy spectrum of many-electron atoms. The cooling function for gas with Solar metallicity is larger by more than an order of magnitude than that of primordial gas, over a broad range of temperatures. Hence, more enriched gas finds it easier to cool.

Atomic gas cannot cool efficiently for temperatures $T \lesssim 10^4$ K, due to the lack of charged particles (electrons and ions) in the gas. However, in chemically enriched gas, the few free electrons present at $T \lesssim 10^4$ K can excite low-energy states (the so-called fine-structure levels) of ions like that of oxygen or carbon. Molecules, on the other hand, have a rich spectrum of energy levels at considerably smaller energies, and can therefore lead to efficient cooling towards lower temperatures. This is the reason why star formation occurs in molecular clouds, where gas can efficiently cool and thereby compress to high densities.

Cooling time. Once we know the rate at which gas loses its energy, we can calculate the cooling time, the time it takes the gas (at constant cooling rate) to lose all of its energy:

$$t_{\rm cool} = \frac{3nk_{\rm B}T}{2C} = \frac{3nk_{\rm B}T}{2n_{\rm H}^2\Lambda(T)} \,. \qquad (10.4)$$

If this cooling time is longer than the age of the Universe, then the gas essentially stays at the same temperature and is unable to collapse towards the halo center. We have seen in Sect. 6.4.3 that for most regions in clusters, this is indeed the case; only in the central regions of clusters cooling can be effective.

Free-fall time. On the other hand, if the cooling time is sufficiently short, gas can compress towards the halo center. What 'sufficiently short' means can be seen if we compare

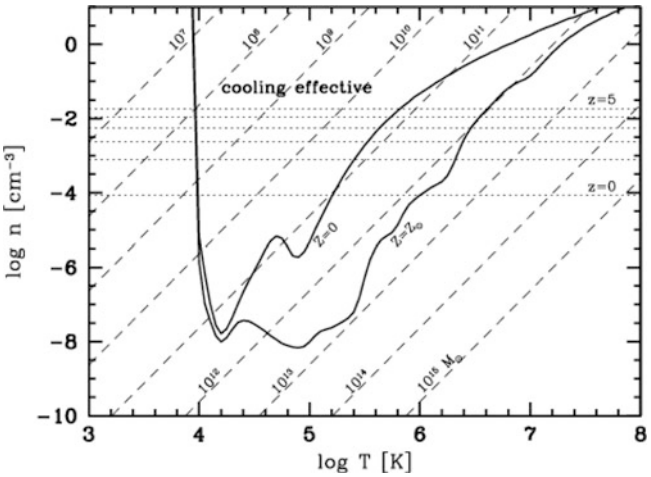


Fig. 10.4 The *solid curves* in this cooling diagram show the density as a function of temperature, for which $t_{\rm cool} = t_{\rm ff}$, both for gas with primordial abundance ($Z = 0$) and with Solar abundance ($Z = Z_\odot$). Note that this condition yields $n \propto f_{\rm g}^{-1}\,[T/\Lambda(T)]^2$, and so these curves are similar to the inverse of the cooling function in Fig. 10.3. Here, the gas fraction was chosen to correspond to its cosmic mean $f_{\rm g} = 0.15$. *Dotted horizontal lines* indicate the density of halos which form at the indicated redshifts, which is determined by the fact that the mean density of a halo is $\sim$ 200 times the critical density of the Universe at this epoch. The *diagonal dashed lines* show the n-T relation for fixed gas mass $M_{\rm g}$, which is obtained from (10.2), $r \propto M/T$ and the fact that $M_{\rm g} \propto r^3 n$. Eliminating r from these two relations yields $n \propto f_{\rm g}^{-1} M_{\rm g}^{-2} T^3$. Source: H. Mo, F. van den Bosch & S. White 2010, *Galaxy Formation and Evolution*, Cambridge University Press, p. 386. Reproduced by permission of the author

the cooling time with the free-fall time, i.e., the time it takes a freely falling particle at some radius r in the halo to reach the center. The free-fall time depends only on the mean total mass density (i.e., dark matter plus baryons) inside r (see problem 4.7) and is given by

$$t_{\rm ff} = \sqrt{\frac{3\pi}{32G\rho}} = \sqrt{\frac{3\pi f_{\rm g}}{32Gn\mu m_{\rm p}}} \,, \qquad (10.5)$$

where we used the gas-mass fraction $f_{\rm g} = \rho_{\rm g}/\rho$ to convert the total density to the gas density, which was then expressed in terms of the particle number density n by $\rho_{\rm g} = n\mu m_{\rm p}$.

Conditions for efficient cooling. If the cooling time is shorter than the free-fall time, then gas falls freely towards the center, essentially unaffected by gas pressure. If, on the other hand, the cooling time is much longer than the free-fall time, the gas at best sinks to the center at a rate given by the cooling rate—this is similar to the cooling flows discussed in Sect. 6.4.3. Hence, in this case, cooling is rather inefficient.

Thus, the condition $t_{\rm cool} = t_{\rm ff}$ separates situations in which gas can easily fall inside the halo and form denser gas concentrations from those where gas compression is prevented. In Fig. 10.4, this condition is shown as solid curves,

both for primordial gas and gas with Solar abundance. For gas densities and temperature above the curves, cooling is efficient, whereas the cooling time is longer than the free-fall time below the curves.

The dotted horizontal lines in Fig. 10.4 indicate the mean gas density of halos collapsed at the redshift indicated, assuming a halos gas-mass fraction of $f_g = 0.15$, i.e., about the cosmic average (recall that a halo has about 200 times the critical density of the Universe at the epoch of halo formation). Thus, for the cooling of gas in halos, only the region above the dotted lines is relevant. For each redshift, there is a range in temperatures for which gas can cool efficiently.

Finally, the dashed diagonal lines indicate the density n as a function of temperature, for a fixed mass M_g as indicated. For $M_g \gtrsim 10^{13} M_\odot$, the dashed line lies below the solid curves for all T; hence, gas in halos with mass $M \gtrsim 10^{13}/f_g\, M_\odot$ cannot cool. Even for $M_g \sim 10^{12} M_\odot$, the dashed curve lies mostly outside the region where cooling is efficient, even for Solar abundance, except below the dotted lines, i.e., at densities which are smaller than the mean densities of halos. But if gas cannot cool, gas condensation and star formation is inefficient.

The difference between galaxies and groups/clusters. From Fig. 10.4, we can thus draw a first important conclusion: In sufficiently massive halos with $M_g \gtrsim 10^{12} M_\odot$, the small efficiency of gas cooling prevents gas from collapsing to the center and forming stars there. At smaller masses, cooling is effective to enable rapid gas collapse. This dividing line in mass is about the mass which distinguishes galaxies from groups and clusters. In the latter, only a small fraction of the baryons is turned into stars, and these are contained in the galaxies within the group; the group halo itself does not contain stars, with the exception of the intracluster light. But as we discussed in Sect. 6.3.4, these stars most likely have been stripped from galaxies in groups through interactions. In contrast, a large fraction of baryons in galaxies is concentrated towards the center, as visible in their stellar distribution. Thus, the difference between galaxies and groups/clusters is their efficiency to turn baryons into stars, and this difference is explained with the different cooling efficiency shown in Fig. 10.4.

This effect partly answers one of the questions posed at the start of this chapter. The mass-to-light ratio of very massive halos is much larger than that of galaxies (see Fig. 10.2) because of the much longer cooling time of the gas. In groups and clusters, most of the gas is present in the form of a hot gaseous halo.

Low-mass halos. Another conclusion we might want to draw from this cooling diagram is the behavior of halos at the low-mass end. A halo with gas mass $\sim 10^{7.5} M_\odot$ lies inside the cooling curve only at *very* high redshift, i.e., when the corresponding density in a halo is very high. Therefore, gas can cool, and stars form, in halos of this mass only if they formed early enough. We therefore expect that the stars in such low-mass halos are very old. We will soon find that there are additional effects which further strengthen this conclusion. Combined, these effects provide a natural explanation for the 'missing satellite' problem discussed in Sect. 7.8.

Cold accretion vs. hot accretion. The cooling diagram in Fig. 10.4 is very useful to discuss such properties qualitatively. Of course, the assumptions made to derive it are quite simple and idealized, such as the consideration of just the mean gas density, instead of a density profile, and the neglect of further effects, such as merging of halos.

A more realistic consideration needs to account for the fact that the gas is not homogeneous. The quasi-hydrostatic density profile implies that the gas density increases towards the center. In the inner part, it may be dense enough for cooling to be effective. In such halos, we therefore expect to have a central concentration of cold gas, surrounded by a hot gaseous halo with a temperature close to the virial temperature.

Furthermore, the implicit assumption of spherical symmetry made above may be misleading. From simulations of structure formation (see Sect. 7.5.3) we have seen that dark matter halos are embedded in a network of sheets and filaments, with massive halos forming at the intersection of filaments. Once formed, such halos accrete further matter, both dark and baryonic matter. In case of spherical symmetry, the gas would fall in and be heated through an accretion shock, as described before. However, the infall of matter occurs predominantly along the directions of the filaments connected to the halo, forming streams of gas which can reach the central regions of the halo without being strongly heated. Hydrodynamic simulations have identified this mode of accretion as an important route for halos to attain or replenish their gas.

10.3 Reionization of the Universe

After recombination at $z \sim 1100$, the intergalactic gas became neutral, with a residual ionization fraction of only $\sim 10^{-4}$. Had the Universe remained neutral we would not be able to receive any photons that were emitted bluewards of the Lyα line of a source, because the absorption cross section for Lyα photons is too large [see (8.27)]. Since such photons are observed from QSOs, as can be seen for instance in the spectra of the $z > 5.7$ QSOs in Fig. 10.5, and since an appreciable fraction of homogeneously distributed neutral gas in the intergalactic medium can be

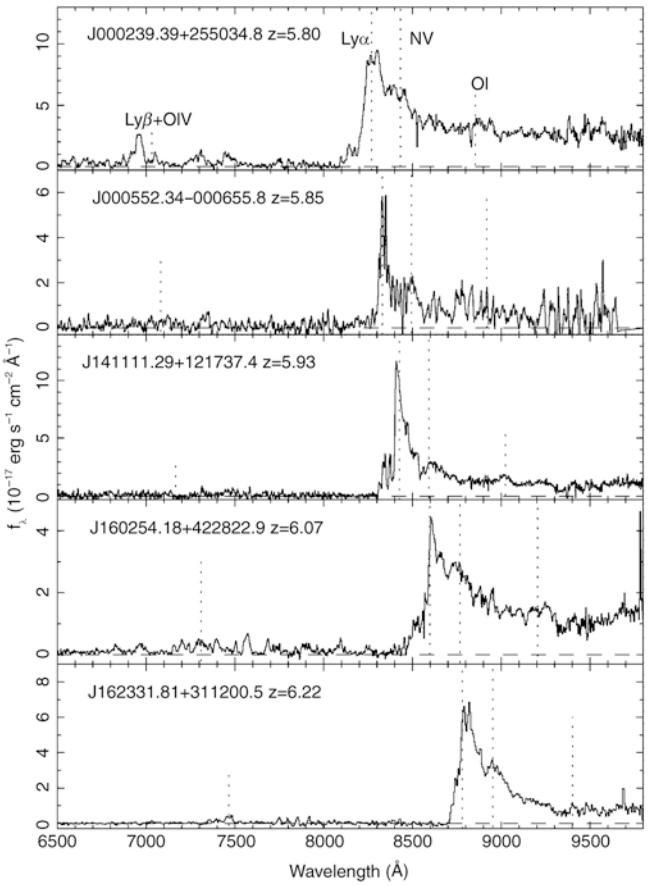


Fig. 10.5 Spectra of five QSOs at redshifts $z > 5.7$, discovered in multi-color data from the Sloan Digital Sky Survey. The positions of the most important emission lines are marked. Particularly remarkable is the almost complete lack of flux bluewards of the Lyα emission line in some of the QSOs, indicating a strong Gunn–Peterson effect. However, this absorption is not complete in all QSOs, which points at strong variations in the density of neutral hydrogen in the intergalactic medium at these high redshifts. Either the hydrogen density varies strongly for different lines-of-sight, or the degree of ionization is very inhomogeneous. Source: X. Fan et al. 2004, *A Survey of $z > 5.7$ Quasars in the Sloan Digital Sky Survey. III. Discovery of Five Additional Quasars*, AJ 128, 515, p. 517, Fig. 1. ©AAS. Reproduced with permission

excluded for $z \lesssim 5$, from the tight upper bounds on the strength of the Gunn–Peterson effect (Sect. 8.5.1), the Universe must have been reionized between the recombination epoch and the redshift $z \sim 7$ of the most distant known QSOs. As we have seen in Sect. 8.6.6, the anisotropies of the CMB led us to conclude that reionization occurred at $z \sim 10$.

This raises the question of how this reionization proceeded, in particular which process was responsible for it. The latter question is easy to answer—reionization must have happened by photoionization. Collisional ionization can be ruled out because for it to be efficient the intergalactic medium (IGM) would need to be very hot, a scenario which can be excluded due to the perfect Planck spectrum of the CMB—the argument here is the same as in Sect. 9.5.3, where we excluded the idea of a hot IGM as the source of the cosmic X-ray background. Hence, the next question is what produced the energetic photons that caused the photoionization of the IGM.

Two kinds of sources may in principle account for them—hot stars or AGNs. Currently, it is not unambiguously clear which of these is the predominant source of energetic photons causing reionization since our current understanding of the formation of supermassive black holes is still insufficient. However, it is currently thought that the main source of photoionization photons is the first generation of hot stars.

10.3.1 The first stars

Following on from the above arguments, understanding reionization is thus directly linked to studying the first generation of stars. In the present Universe star formation occurs in galaxies; thus, one needs to examine when the first galaxies could have formed. From the theory of structure formation, the mass spectrum of dark matter halos at a given redshift can be computed by means of, e.g., the Press–Schechter model (see Sect. 7.5.2). Two conditions need to be fulfilled for stars to form in these halos. First, gas needs to be able to fall into the dark halos. Since the gas has a finite temperature, pressure forces may impede the infall into the potential well. Second, this gas also needs to be able to cool, condensing into clouds in which stars can then be formed, a process that we considered in the preceding section.

The Jeans mass. By means of a simple argument, we can estimate under which conditions pressure forces are unable to prevent the infall of gas into a potential well. To do this, we consider a slightly overdense spherical region of radius R whose density is only a little larger than the mean cosmic matter density $\bar{\rho}$. If this sphere is homogeneously filled with baryons, the gravitational binding energy of the gas is about

$$|E_{\rm grav}| \sim \frac{GMM_{\rm g}}{R} \,,$$

where M and $M_{\rm g}$ denote the total mass and the gas mass of the sphere, respectively. The thermal energy of the gas can be computed from the kinetic energy per particle, multiplied by the number of particles in the gas, or

$$E_{\rm th} \sim c_{\rm s}^2 \, M_{\rm g} \,, \text{where} \quad c_{\rm s} \approx \sqrt{\frac{k_{\rm B} T_{\rm g}}{\mu m_{\rm p}}}$$

is the speed of sound in the gas, which is about the average velocity of the gas particles, and $\mu m_{\rm p}$ denotes, as before, the average particle mass in the gas. For the gas to be bound in

the gravitational field, its gravitational binding energy needs to be larger than its thermal energy, $|E_{\rm grav}| > E_{\rm th}$, which yields the condition $GM > c_{\rm s}^2 R$. Since we have assumed an only slightly overdense region, the relation $M \sim \bar{\rho} R^3$ between mass and radius of the sphere applies. From the two latter equations, the radius can be eliminated, yielding the condition

$$M > M_{\rm J} \equiv \frac{\pi^{5/2}}{6} \left(\frac{c_{\rm s}^2}{G}\right)^{3/2} \frac{1}{\sqrt{\bar{\rho}}} , \qquad (10.6)$$

where the numerical coefficient is obtained from a more accurate treatment. Thus, as a result of our simple argument we find that the mass of the halo needs to exceed a certain threshold for gas to be able to fall in. The expression on the right-hand side of (10.6) defines the *Jeans mass* $M_{\rm J}$, which describes the minimum mass of a halo required for the gravitational infall of gas. The Jeans mass depends on the temperature of the gas, expressed through the sound speed $c_{\rm s}$, and on the mean cosmic matter density $\bar{\rho}$. The latter can easily be expressed as a function of redshift, $\bar{\rho}(z) = \bar{\rho}_0 (1+z)^3$.

The baryon temperature $T_{\rm b}$ has a more complicated dependence on redshift. For sufficiently high redshifts, the small fraction of free electrons that remains after recombination provides a thermal coupling of the baryons to the cosmic background radiation, by means of Compton scattering. This is the case for redshifts $z \gtrsim z_{\rm t}$, where

$$z_{\rm t} \approx 140 \left(\frac{\Omega_{\rm b} h^2}{0.022}\right)^{2/5} ;$$

hence, $T_{\rm b}(z) \approx T(z) = T_0(1+z)$ for $z \gtrsim z_{\rm t}$. For smaller redshifts, the density of photons gets too small to maintain this coupling, and baryons start to adiabatically cool down by the expansion, so that for $z \lesssim z_{\rm t}$ we obtain approximately $T_{\rm b} \propto \rho_{\rm b}^{2/3} \propto (1+z)^2$ (see problem 4.9).

From these temperature dependences, the Jeans mass can then be calculated as a function of redshift. For $z_{\rm t} \lesssim z \lesssim 1000$, $M_{\rm J}$ is independent of z because $c_{\rm s} \propto T^{1/2} \propto (1+z)^{1/2}$ and $\bar{\rho} \propto (1+z)^3$, and its value is

$$M_{\rm J} = 1.35 \times 10^5 \left(\frac{\Omega_{\rm m} h^2}{0.15}\right)^{-1/2} M_\odot , \qquad (10.7)$$

whereas for $z \lesssim z_{\rm t}$ we obtain, with $T_{\rm b} \simeq 1.7 \times 10^{-2} (1+z)^2$ K,

$$M_{\rm J} = 5.7 \times 10^3 \left(\frac{\Omega_{\rm m} h^2}{0.15}\right)^{-1/2} \times \left(\frac{\Omega_{\rm b} h^2}{0.022}\right)^{-3/5} \left(\frac{1+z}{10}\right)^{3/2} M_\odot . \qquad (10.8)$$

Hence, gas can not fall into halos with mass lower than these values.

Cooling of the gas. The Jeans criterion is a necessary condition for the formation of proto-galaxies, i.e., dark matter halos which contain baryons. In order to form stars, the gas in the halos needs to be able to cool further. Here, we are dealing with the particular situation of the first galaxies, whose gas is metal-free, so metal lines cannot contribute to the cooling. As we have seen in Fig. 10.3, the cooling function of primordial gas is much smaller than that of enriched material; in particular, the absence of metals means that even slow cooling though excitation of fine-structure lines cannot occur, as there are no atoms with such transitions present. Thus, cooling by the primordial gas is efficient only above $T \gtrsim 2 \times 10^4$ K. However, the halos formed at high redshift have low mass. We have seen in Sect. 7.5.2 that the abundance of dark matter halos depends on the parameter ν in (7.51), given by the product of the density fluctuations on a given mass scale and the growth factor. At high redshift, the growth factor $D_+(a)$ is small, and thus to have a noticeable abundance of halos of mass M, $\sigma(M)$ must be correspondingly large. At redshift $z \sim 10$, the parameter ν is about unity for halos of mass $\sim 10^3 M_\odot$. Hence, at that time, substantially more massive halos than that were (exponentially) rare—i.e., only low-mass halos were around, and their virial temperature

$$T_{\rm vir} \approx 2 \times 10^2 \left(\frac{M}{10^5 h^{-1} M_\odot}\right)^{2/3} \left(\frac{1+z}{10}\right) {\rm K} \qquad (10.9)$$

is considerably below the energy scale where atomic hydrogen can efficiently cool. To derive (10.9), we have replaced $V_{\rm c}$ in (10.2) in favor of halo mass and radius, and used the fact that the mean matter density of a halo inside its virial radius is ~ 200 times the critical density at a given redshift. Therefore, atomic hydrogen is a very inefficient coolant for these first halos, insufficient to initiate the formation of stars. Furthermore, helium is of no help in this context, since its excitation temperature is even higher than that of hydrogen.

The importance of molecular hydrogen. Besides atomic hydrogen and helium, the primordial gas contains a small fraction of molecular hydrogen which represents an extremely important component in cooling processes. Whereas in enriched gas, molecular hydrogen is formed on dust particles, the primordial gas had no dust, and so H_2 must form in the gas phase itself, rendering its abundance very small. However, despite its very small density and transition probability, H_2 dominates the cooling rate of primordial gas at temperatures below $T \sim 10^4$ K—see Fig. 10.6—where the precise value of this temperature depends on the abundance of H_2.

By means of H_2, the gas can cool in halos with a temperature exceeding about $T_{\rm vir} \gtrsim 1000\,{\rm K}$, corresponding to a halo mass of $M \gtrsim 5 \times 10^4 M_\odot$ at $z \sim 20$. In these halos, stars may then be able to form. These stars will certainly be different from those known to us, because they do not contain any metals. Therefore, the opacity of the stellar plasma is much lower. Such stars, which at the same mass presumably have a much higher temperature and luminosity (and thus a shorter lifetime), are called *population III stars*. Due to their high temperature they are much more efficient sources of ionizing photons than stars with 'normal' metallicity.

10.3.2 The reionization process

Dissociation of molecular hydrogen. The energetic photons from these population III stars are now capable of ionizing hydrogen in their vicinity. More important still is another effect: photons with energy above 11.26 eV can destroy H_2. Since the Universe is transparent for photons with energies below 13.6 eV, photons with $11.26\,{\rm eV} \le E_\gamma \le 13.6\,{\rm eV}$ can propagate very long distances and dissociate molecular hydrogen. This means that as soon as the first stars have formed in a region of the Universe, molecular hydrogen in their vicinities will be destroyed and further gas cooling and star formation will then be prevented.[2] At this point, the Universe contains a low number density of isolated bubbles of ionized hydrogen, centered on those halos in which population III stars were able to form early, but this constitutes only a tiny fraction of the volume; most of the baryons remain neutral.

Metal enrichment of the intergalactic medium. Soon after population III stars have formed, they will explode as supernovae. Through this process, the metals produced by them are ejected into the intergalactic medium, by which the initial metal enrichment of the IGM occurs. The kinetic energy transferred by SNe to the gas within the halo can exceed its binding energy, so that the baryons of the halo can be blown away and further star formation is prevented. Whether this effect may indeed lead to gas-free halos, or whether the released energy can instead be radiated away, depends on the geometry of the star-formation regions. In any case, it can be assumed that in those halos where the first generation of stars was born, further star formation was considerably suppressed, particularly since all molecular hydrogen was destroyed.

We can assume that the metals produced in these first SN explosions are, at least partially, ejected from the halos into the intergalactic medium, thus enriching the latter. The existence of metal formation in the very early Universe is concluded from the fact that even sources at very high redshift (like QSOs at $z \sim 6$) have a metallicity of about one tenth the Solar value. Furthermore, the Lyα forest also contains gas with non-vanishing metallicity. Since the Lyα forest is produced by the intergalactic medium, this therefore must have been enriched.

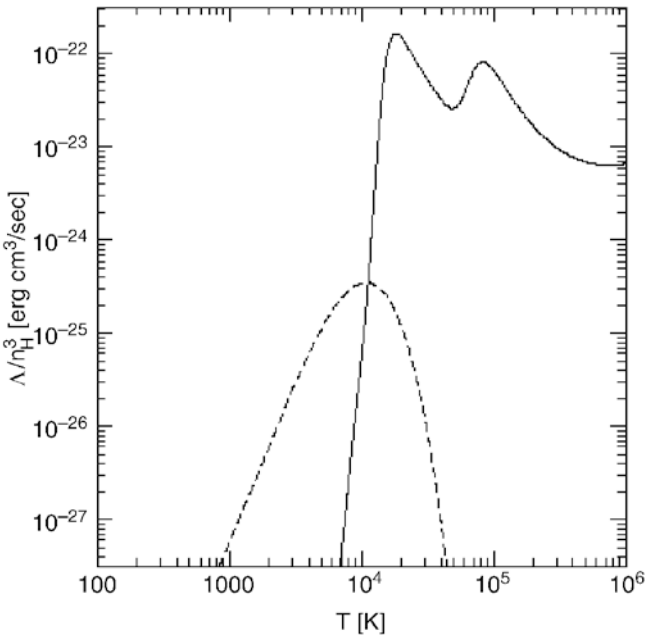


Fig. 10.6 Cooling rate as a function of the temperature for a gas consisting of atomic and molecular hydrogen (with 0.1 % abundance) and of helium. The *solid curve* describes the cooling by atomic gas, the *dashed curve* that by molecular hydrogen; thus, the latter is extremely important at temperatures below $\sim 10^4$ K. At considerably lower temperatures the gas cannot cool, hence no star formation can take place. Source: R. Barkana & A. Loeb 2000, *In the Beginning: The First Sources of Light and the Reionization of the Universe*, astro-ph/0010468, Fig. 12. Reproduced by permission of the author

The final step to reionization. For gas to cool in halos without molecular hydrogen, their virial temperature needs to exceed about 10^4 K (see Fig. 10.6). Halos of this virial temperature form with appreciable abundance at redshifts of $z \sim 10$, corresponding to a halo mass of $\sim 10^7 M_\odot$, as can be estimated from the Press–Schechter model (see Sect. 7.5.2). In these halos, efficient star formation can then take place and the first proto-galaxies form. These then ionize the surrounding IGM in the form of HII-regions, as sketched in Fig. 10.7. The corresponding HII-regions expand because increasingly more photons are produced. If the halo density is sufficiently high, these HII-regions start to overlap and soon after, to fill the whole volume. Once this occurs, the IGM is ionized, and reionization is completed.

We therefore conclude that reionization is a two-stage process. In a first phase, population III stars form through

[2] To destroy all the H_2 in the Universe one needs less than 1 % of the photon flux that is required for the reionization.

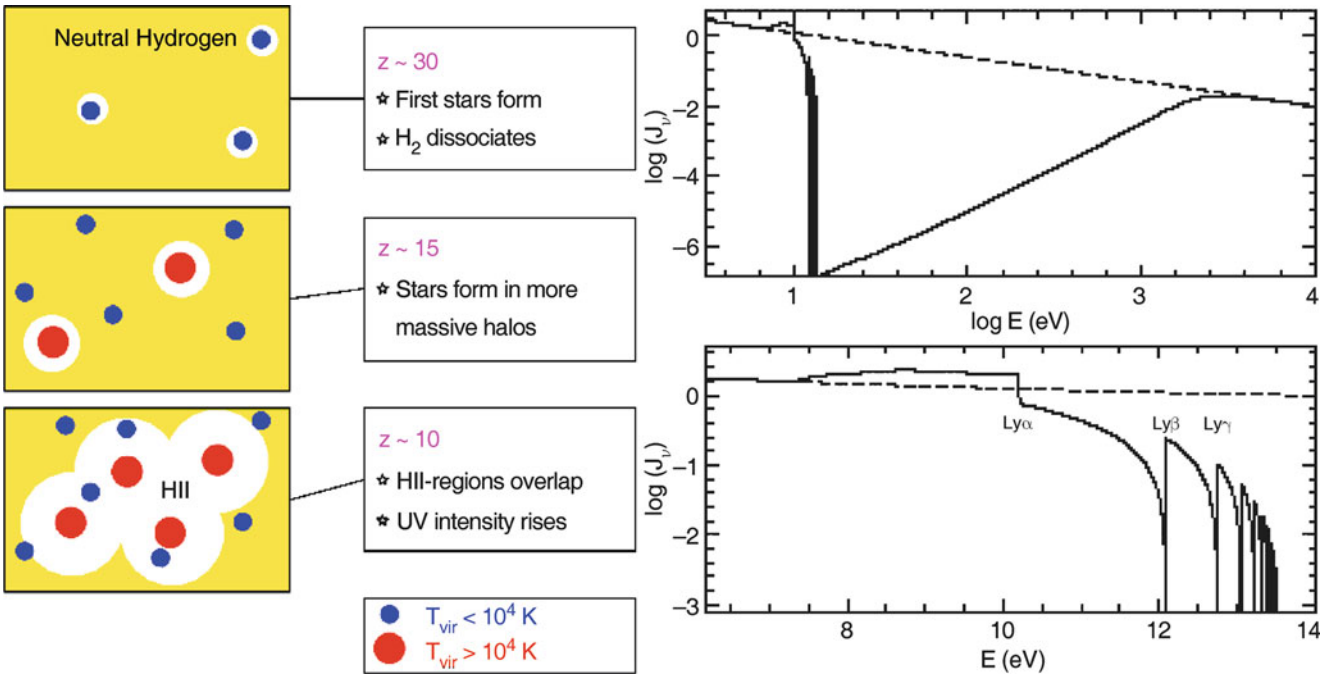


Fig. 10.7 *On the left*, a sketch of the geometry of reionization is shown: initially, relatively low-mass halos collapse, a first generation of stars ionizes and heats the gas in and around these halos. By heating, the temperature increases so strongly (to about $T \sim 10^4$ K) that gas can escape from the potential wells; these halos may never again form stars efficiently. Only when more massive halos have collapsed will continuous star formation set in. Ionizing photons from this first generation of hot stars produce HII-regions around their halos, which is the onset of reionization. The regions in which hydrogen is ionized will grow until they start to overlap; at that time, the flux of ionizing photons will strongly increase. *On the right*, the average spectrum of photons at the beginning of the reionization epoch is shown; here, it has been assumed that the flux from the radiation source follows a power law (*dashed curve*). Photons with an energy higher than that of the Lyα transition are strongly suppressed because they are efficiently absorbed. The spectrum near the Lyman limit shows features which are produced by the combination of breaks corresponding to the various Lyman lines, and the redshifting of the photons. Source: R. Barkana & A. Loeb 2000, *In the Beginning: The First Sources of Light and the Reionization of the Universe*, astro-ph/0010468, Figs. 4, 11. Reproduced by permission of the author

cooling of gas by molecular hydrogen, which is then destroyed by these very stars. Only in a later epoch and in more massive halos cooling is provided by atomic hydrogen, leading to reionization.

Escape fraction of ionizing photons. We note that only a small fraction of the baryons needs to undergo nuclear fusion in hot stars to ionize all hydrogen, as we can easily estimate: by fusing four H-nuclei (protons) to He, an energy of about 7 MeV per nucleon is released. However, only 13.6 eV per hydrogen atom is required for ionization. Hence, from a purely energetic point of view, reionization is not particularly demanding.

The number density of hot stars required to reionize the Universe is uncertain due to the unknown escape fraction $f_{\rm esc}$ of ionizing photons from the first galaxies, i.e., the ratio of the number of ionizing photons which can propagate out of the galaxy to the total number of ionizing photons produced by hot stars. Photons with energy $E_\gamma \geq 13.6$ eV are easily absorbed by the neutral fraction of the gas. For local star-forming galaxies, the escape fraction can be estimated to be between a few percent up to ~ 0.5. However, the first galaxies that formed were denser. Furthermore, the escape fraction depends on the geometrical arrangement of the hot stars relative to the interstellar medium, as well as the clumpiness of the latter. If the stars are located in the inner part of the galaxy, surrounded by a smooth interstellar medium, the escape fraction will be very small. If, however, the ISM is clumpy such that it only occupies a small fraction of the volume, photons can escape 'between the clumps', and the escape fraction can be appreciable. There is also the possibility that the star formation and subsequent supernovae drive much of the gas out of the galaxy halos, increasing the escape fraction for later stellar generations.

Clumpiness of the intergalactic medium. A further uncertainty in the quantitative understanding of reionization lies in the clumpiness of the intergalactic medium. An ionized hydrogen atom may become neutral again due to recombination. Hence, one may need more than one ionizing photon per atom for complete reionization. Since recombination is a two-body process (i.e., its rate depends quadratically

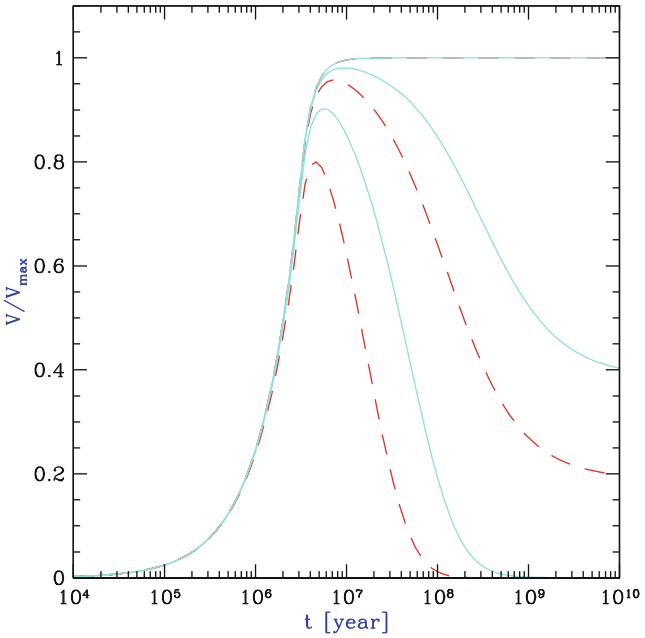


Fig. 10.8 The volume of an expanding HII region from an instantaneous starburst, normalized to the maximally possible volume $V_{\max}$ (which is given by equating the number of hydrogen atoms $N_{\rm H} = \bar{n}_{\rm H} V_{\max}$ with the total number of ionizing photons generated by the starburst). The *upper solid curve* assumes that no recombination takes place. The two other *solid curves* assume that the starburst occurs at redshift $z = 10$, and that the intergalactic medium is uniform (*middle solid curve*) or strongly clumped (*lower solid curve*). The two *dashed curves* show the same, except that for them, $z = 15$ is assumed; since the density is higher at larger redshift, the recombination rate is accordingly higher. Source: R. Barkana & A. Loeb 2000, *In the Beginning: The First Sources of Light and the Reionization of the Universe*, astro-ph/0010468, Fig. 21. Reproduced by permission of the author

on the gas density), its relative importance depends on the redshift of reionization and the clumpiness of the gas distribution: The higher the redshift, the larger the density of the intergalactic medium, and the higher the recombination rate. Clumpiness also increases the mean of the squared gas density, yielding a higher mean recombination rate (see Fig. 10.8).

Once reionization is completed, the intergalactic medium has a temperature of about 10^4 K, due to the heating of the gas by photoionization: the typical energy of a photon which ionizes a hydrogen atom is somewhat larger than 13.6 eV, and the energy difference is transferred to the electron, which is tightly coupled by Coulomb interactions with the other gas particles. Thus, this surplus energy causes a heating of the gas. The resulting temperature depends on the spectrum of the ionizing radiation; the harder the spectrum, the higher the temperature.

Suppression of low-mass galaxies. The increase of temperature causes an increase of the Jeans mass (10.6), due to its dependence on the sound velocity. Once the intergalactic medium is heated to $\sim 10^4$ K by intergalactic UV radiation, the gas pressure prevents gas inflow into low-mass halos, corresponding[3] to circular velocities $\lesssim 30$ km/s. For this reason, one expects that halos of lower mass have a lower baryon fraction than that of the cosmic mixture, $f_{\rm b} = \Omega_{\rm b}/\Omega_{\rm m}$. The actual value of the baryon fraction depends on the details of the merger history of a halo. Quantitative studies yield an average baryon mass of

$$\bar{M}_{\rm b} = \frac{f_{\rm b}\, M}{\left[1 + (2^{\alpha/3} - 1)(M_{\rm C}/M)^{\alpha}\right]^{(3/\alpha)}} , \qquad (10.10)$$

where $M_{\rm C} \sim 10^9 M_\odot$ is a characteristic mass, defined such that for a halo with mass $M_{\rm C}$, $\bar{M}_{\rm b}/M = f_{\rm b}/2$. For halos of mass smaller than $M_{\rm C}$, the baryon fraction is suppressed, decreasing as $(M/M_{\rm C})^3$ for small masses, whereas for halo masses $\gg M_{\rm C}$, the baryon fraction corresponds to the cosmic average. The index $\alpha \sim 2$ determines the sharpness of the transition between these two cases. The characteristic mass $M_{\rm C}$ depends on redshift, being much smaller at high z due to the stronger ionizing background.

The ionizing flux has two additional effects on the gas that resides in halos: it provides a source of heating, due to photoionization, and it leads to a higher degree of ionization in the gas, reducing the number density of atoms which can be excited by collisions and cool through de-excitation. Both effects act in the same direction, by impeding an efficient cooling of the gas and hence the formation of stars. For halos of larger mass, intergalactic radiation is of fairly little importance because the corresponding heating rate is substantially smaller than that occurring by the dissipation of the gas which is needed to concentrate the baryons towards the halo center. For low-mass halos, however, this effect is important. Together, these two effects reduce the cooling rate of the gas, which is a dominant effect for low-mass halos. Thus, the gas in low-mass halos cannot cool efficiently, suppressing star formation—unless star formation occurred before the reionization was completed. We hence found one of the elements for the second part of the answer to the question about the different mass-to-light ratios in halos, illustrated in Fig. 10.2: star formation in low mass halos is strongly suppressed due to the ionizing background radiation. As already discussed in Sect. 7.8, this also provides an explanation of the 'missing satellite problem'.

Helium reionization. Our discussion was confined to the ionization of hydrogen and we ignored helium. To singly ionize helium, photons of energy ≥ 24.6 eV are required, and the ionization energy of He II is four times that of hydrogen.

[3] We remind the reader about the connection between halo masses and circular velocities; cf. Sect. 7.6.1; see also (10.2).

In addition, the recombination rate of fully ionized helium is about five times higher than that of hydrogen. Therefore, the reionization of helium is expected to be completed at a later epoch when the density of photons with $\lambda < 304$ Å was high enough. Since even massive stars do not generate photons exceeding this energy in large quantities, the photons leading to helium reionization presumably are emitted by quasars; therefore, the ionization of helium has to wait for the 'quasar epoch' of the Universe, at $z \lesssim 4$. From the statistical analysis of the Lyα forest and from the analysis of helium absorption lines and the helium Gunn–Peterson effect in high-redshift QSOs, a reionization redshift of $z \sim 3$ for helium is obtained.

10.3.3 Observational probes of reionization

One of the challenges of current observational cosmology is to link the history of reionization, as outlined above, to the observation of the highest redshift sources, i.e., to see whether we can observe the sources which are responsible for cosmic reionization. Are the galaxy populations that we can find at very high redshifts sufficient to understand the reionization process? Here we shall mention some of the major obstacles for a direct observation probe of these ionizing sources.

The stellar mass at high redshifts. If reionization was caused by the energetic photons emitted during star formation, the remnants of this first generation of stars must be present in the post-reionization Universe, and thus be observable. As we discussed in some detail in Chap. 9, galaxies at redshift > 6 are observed, either as Lyman-break galaxies (LBGs), Lyman-alpha emitters (LAEs) or as sub-millimeter galaxies (SMGs). Their stellar masses can be estimated from observed light. However, most of the LBGs are observed only in the near-IR, which means that we see their restframe UV-emission. Converting the UV-light into a stellar mass is highly uncertain, since it depends strongly on the instantaneous star-formation rate. For LAEs, it is even more challenging to determine a stellar mass, since they are typically fainter in their broad-band (continuum) emission, which renders the determination of the stellar mass even more challenging.

Nevertheless, galaxies at very high redshift were found which appear to have high stellar masses, including a LAE at $z = 6.60$ with an estimated stellar mass $M_* \gtrsim 10^{11} M_\odot$. The high-redshift QSOs require a SMBH with $M_\bullet \gtrsim 10^9 M_\odot$ to power their energy output, and these must be hosted is galaxies with very large stellar mass. Therefore, massive galaxies have formed very early on, delivering ionizing photons.

However, these highest mass objects are very rare and, by themselves, by far not able to explain reionization. This fact can be clearly seen by considering the spectral shape of the

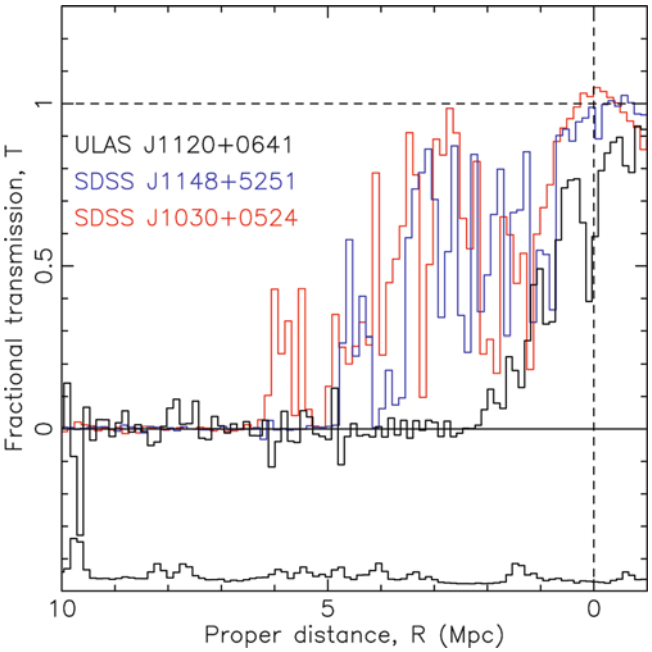


Fig. 10.9 The spectra of three high-redshift QSOs (SDSS J1148+5251 at $z = 6.42$, SDSS J1030+0524 at $z = 6.31$, and the $z = 7.085$ QSO ULAS J1120+0641) at the Lyman-α emission line. For this figure, the wavelength difference to the Lyman-α transition is expressed in proper distance away from the QSOs. The spectra are normalized, dividing them by the extrapolation of the continuum on the red side of the emission line, yielding the transmission. Source: D.J. Mortlock et al. 2011, *A luminous quasar at a redshift of* $z = 7.085$, Nature 474, 616, Fig. 3. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2011

Lyα emission line of high-redshift QSOs. Figure 10.9 shows the spectrum of three very high redshift QSOs near to the Lyman-α emission line. Whereas all three QSO show essentially no flux shortward of Lyman-α, once the wavelength difference exceeds $\sim$ 20 Å in the restframe, there is some transmitted flux very close to the Lyman-α transition. This near-zone transmission is understood as a region around the QSO where the intergalactic gas is fully ionized by the QSO, so it becomes transparent. The figure shows a clear trend that the size of this near zone decreases for higher redshifts, as would be expected due to the higher gas density and probably larger mean neutral fraction in the Universe. Thus, these very luminous objects are able to reionize the intergalactic medium in their immediate surroundings, but their effect is constrained to a rather limited volume. Most of the ionizing photons must come from the far more numerous lower-mass galaxies, i.e., far less luminous sources.

The UV-luminosity function at high redshifts. The large number of LBG candidates at redshifts $z \gtrsim 7$ recently obtained yields constraints on the luminosity function of galaxies in the rest-frame ultraviolet regime of the spectrum. As pointed out in Sect. 9.2.4, for most of them no spectroscopic confirmation is available, so that each individ-

ual case is burdened with uncertainty. We have an idea of what the UV-luminosity function looks like for $z \lesssim 8$, as shown in Fig. 9.41, but the star-formation rate density beyond $z \sim 8$ is still very uncertain, as shown in Fig. 9.57.

Since at such high redshifts, high-mass dark matter halos were extremely rare, we actually expect that most star formation at $z \sim 10$ occurs in very low-mass systems which will be very difficult to detect. Thus, in order to translate the observed luminosity function into a star-formation rate, large extrapolations towards very low-luminosity sources are required, burdened with substantial uncertainties.

The UV-slope. The radiation we observe from high-redshift galaxies corresponds to wavelengths longward of the Lyα transition, i.e., at wavelengths considerably larger than that of ionizing photons. Therefore, to relate the observed properties to the ionizing power, the spectral shape needs to be extrapolated to shorter wavelengths.

This extrapolation is done using a power law for the UV-continuum which is conventionally parametrized as $S_\lambda \propto \lambda^\beta$. A source with slope $\beta = -2$ corresponds to a flat spectrum in S_ν, for which the AB-magnitudes (see Sect. A.4) would be independent of the chosen filter. Hence, in order to relate the observed flux of sources to their emission of ionizing photons, the slope β must be known. In principle, a very young, low-metallicity stellar population can have a hard spectrum with $\beta \sim -3$, but as soon as the metallicity increases above $\sim 10^{-2} Z_\odot$ or the age of the stellar population is larger than $\sim 10^7$ yr, the spectrum will get flatter; of course, any extinction (and related reddening) leads to an increase of β as well.

In principle, the slope β can be obtained from observing galaxies in at least two wavebands. For the highest-redshift sources, that corresponds to bands in the observed near-IR regime. Unfortunately, even relatively small photometric uncertainties translate into rather large error bars on β. At present, observations seem to indicate that the mean value of β is between -2 and -2.5 for $z \sim 7$ galaxies.

The escape fraction. Even if the extrapolation from the observed rest-frame UV at $\lambda \sim 1500$ Å to the ionizing region of $\lambda < 912$ Å were accurate, we still would not know the emission of ionizing photons from these galaxies. The interstellar medium in these objects is expected to absorb many of the ionizing photons, before they can escape the galaxy. The escape fraction $f_{\rm esc}$ is very uncertain, and any theoretical estimate of it is highly model dependent.

We thus conclude that, using reasonable guesses (within the current observational constraints) regarding the UV-luminosity function at high-z, the UV-slope β, and the escape fraction ($f_{\rm esc} \sim 0.2$, as is suggested from the properties of $z \sim 3$ LBGs), the number density of ionizing photons emitted from the early galaxies may be sufficient to explain the reionization of the Universe at $z \sim 10$, as suggested by the results from the CMB anisotropies.

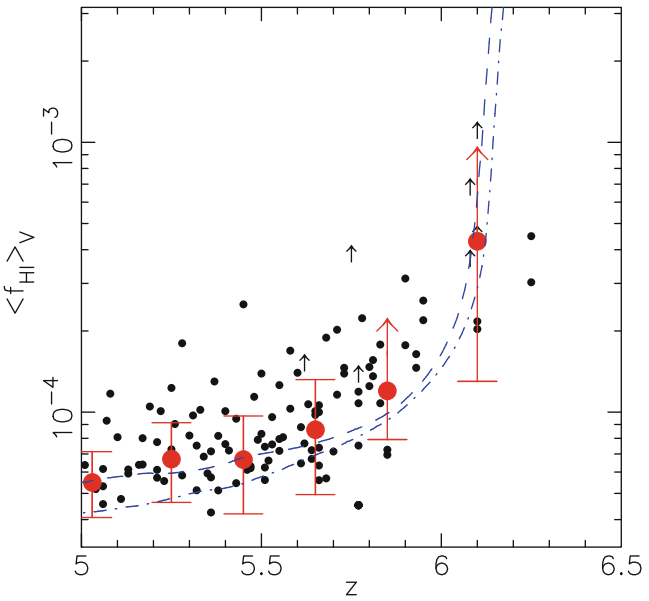


Fig. 10.10 Redshift evolution of the mean neutral fraction of hydrogen in the intergalactic medium, as obtained from the absorption of ionizing radiation from high-redshift QSOs (Gunn–Peterson effect). Individual measurements are shown as *small dots*, whereas the *large circles* with error bars represent averages over redshift bins. The two *curves* show results from numerical simulations. Source: X. Fan et al. 2006, *Constraining the Evolution of the Ionizing Background and the Epoch of Reionization with $z \sim 6$ Quasars. II. A Sample of 19 Quasars*, AJ 132, 117, p. 126, Fig. 7. ©AAS. Reproduced with permission

Towards a larger neutral hydrogen fraction. The observed spectrum of high-redshift QSOs shortwards of the Lyα emission line shows that an increasing fraction of the radiation is absorbed by neutral hydrogen on the line-of-sight. We have seen that the density of the Lyα forest increases with redshift (cf. Fig. 10.5) in such a way that only a tiny fraction of ionizing photons manage to escape absorption. This observation may be seen as an indication that we approach the epoch of reionization as the QSO redshift increases beyond $z \sim 6$. However, as shown in Fig. 10.10, the mean neutral fraction of intergalactic hydrogen needed to cause this strong absorption of ionizing photons is still very small—a neutral fraction of much less than 1 % is sufficient to entirely block the light of QSOs shortward of the Lyα emission. Hence, the strong absorption implied by QSO spectra cannot be taken as evidence for $z \sim 6$ signalling the end of the reionization epoch. Nevertheless, the trend of the data shown in Fig. 10.10 may suggest that beyond $z \sim 6$, we may approach a phase where the neutral hydrogen fraction indeed starts to increase significantly.

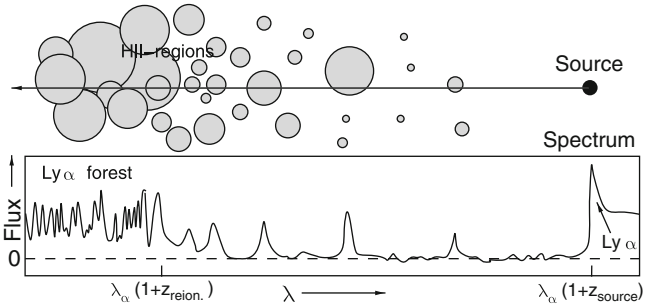


Fig. 10.11 Sketch of a potential observation of reionization: light from a very distant QSO propagates through a partially ionized Universe; at locations where it passes through HII-regions, radiation will get through—flux will be visible at the corresponding wavelengths. When the HII-regions start to overlap, the normal Lyα forest will be produced. Adapted from: R. Barkana & A. Loeb 2000, *In the Beginning: The First Sources of Light and the Reionization of the Universe*, astro-ph/0010468

Observing reionization directly may in principle be possible if a very high-redshift QSO could be identified whose absorption spectrum could reveal a tomographic view through the ionized 'bubbles' of the intergalactic medium, as sketched in Fig. 10.11. But we point out again that the very dense Lyα forest seen towards QSOs at high redshift, is no unambiguous sign for approaching the redshift of reionization, because a very small fraction of neutral atoms (about 1 %) is already sufficient to produce a large optical depth for Lyα photons.

With the upcoming Next Generation Space Telescope, the James Webb Space Telescope (JWST), one hopes to observe the epoch of reionization directly and to discover the first light sources in the Universe; this space telescope, with a diameter of 6.5 m, will be optimized for operation at wavelengths between 1 and 5 μm.

10.4 The formation of disk galaxies

We now turn to describe in somewhat more detail the fate of the cooling gas inside halos. The most important aspect in addition to the cooling processes described before is the fact that dark matter halos, and the gas inside of them, contain angular momentum. As we shall see, this naturally leads to the formation of galactic disks.

10.4.1 The contraction of gas in halos

We described in Sect. 7.6.2 that a non-spherical overdensity can attain an angular momentum, due to a torque caused by the tidal gravitational field in which the overdensity is located. Therefore, dark matter halos are born with a finite angular momentum, which we quantified by the spin parameter λ [see (7.65)]. Analytical estimates and numerical simulations show that the typical value is $\lambda \sim 0.05$, however with a rather broad distribution.

In the initial stages of the evolution of the overdensity, we expect that baryons and dark matter have the same spatial distribution, thus the specific angular momentum of the baryons and dark matter are the same. When the halo collapses, the gas distribution may become different from that of the dark matter, but the torque on the halo is strongest at maximum radius (i.e., at turnaround), and thus during collapse, little angular momentum is obtained.

When the gas in a halo cools, it collapses toward the center, thereby conserving its angular momentum. The gas can therefore not collapse to an arbitrarily small region; the angular momentum barrier prevents this. Frictional forces in the gas drive the gas onto approximately circular orbits, depending on the symmetries of the halo, in a plane perpendicular to the angular momentum vector—it forms a flat disk. The gas in the disk is much denser than it would be if the gas retained on almost spherical distribution; hence, gas in the disk finds it easier to cool and form stars—in accordance with observations: most of the quiescent star formation in the current Universe occurs in galactic disks.

The necessity for dark matter. Understanding the formation of disk galaxies requires the presence of dark matter, as we shall see now. Let us assume the contrary, namely that the density concentration which formed through gravitational instability consists solely of baryons. In this case, the baryons are also the only source of gravity. The characteristic spin parameter of the forming halo is about 0.05. The spin parameter of a self-gravitating, thin exponential disk can be calculated to be $\lambda_\mathrm{d} \approx 0.425$. As the gas cloud collapses into a disk, it conserves its mass and its angular momentum, whereas it can get rid of energy by radiation emitted in the cooling processes. The binding energy scales like r^{-1}. Therefore, the spin parameter scales like $r^{-1/2}$, as follows from (7.65). The final spin parameter is thus related to the initial spin parameter λ_i by

$$\lambda = \lambda_\mathrm{i} \sqrt{\frac{r_\mathrm{i}}{r}} \,, \tag{10.11}$$

where r_i is the radius of the virialized gas cloud before cooling. In order to get a spin parameter of ~ 0.42 for the exponential disk from a spin parameter λ_i obtained from tidal torques, the gas must collapse by a factor $\sim (0.42/0.05)^2 \approx 70$.

We can take our Milky Way as an example for this process. The radius of the visible disk is of the order of 10 kpc, which, according to the previous assumptions, would have collapsed from an initial radius of $\sim$ 700 kpc. With a baryonic mass of $\sim 5 \times 10^{10} M_\odot$ for the Milky Way, the free-

fall time from radius 700 kpc is $\sim 4\times10^{10}$ yr, i.e., about three times the current age of the Universe. Therefore, the Milky Way disk could not have formed until today if it consisted only of baryonic matter. In fact, given that the Milky Way contains old stars, we have good reasons to assume that it has formed quite a bit before today, so that the discrepancy of time scales becomes even stronger.

Gas collapse in a dark matter halo. If, however, the gas contracts in a dark matter halo, the situation is quite different. Assume, for simplicity, that the density profile of the dark matter halo behaves like $\rho \propto r^{-2}$, up to the virial radius; this corresponds to the isothermal sphere which we discussed in Sect. 3.11.2 yielding a radius-independent rotational velocity V_c. If the halo has a spin parameter of $\lambda = 0.05$, then the rotational velocity of the halo, and the gas inside of it, is about $V_c/7$. If the gas sinks to the center, thereby conserving its specific angular momentum $\propto r\,v$, it needs to reduce its radius by a mere factor of 7 to form a rotationally supported disk—an order of magnitude less than in the hypothetical case of baryon-only halos.[4] The time-scale for the formation of a Milky Way-like disk is then reduced to $\sim 10^9$ yr, thus such disks can form sufficiently early in the cosmic evolution.

10.4.2 The formation of galactic disks

Empirically, it is found that the light distribution of disk galaxies follows an exponential law. Assuming a fixed mass-to-light ratio, this implies that the surface mass density $\Sigma(R)$ behaves like

$$\Sigma(R) = \Sigma_0 \exp\left(-\frac{R}{R_d}\right) , \tag{10.12}$$

where Σ_0 is the central surface mass density, and R_d the scale-length of the disk. For the considerations that follow, we shall assume that the dark matter in the halo follows an isothermal density profile, and that the self-gravity of the disk is negligible. The former assumption is motivated by the observed flat rotation curves of disk galaxies; we point out that the rotational velocity predicted by NFW density profiles (see Sect. 7.6.1) is fairly constant over a broad range of radius.

Estimating the disk scale length. Starting from a dark matter halo, its virial mass M, virial radius r_{200} and virial velocity V_{200} are related through (7.58). The assumption of an isothermal profile then implies that the rotational velocity $V_{\rm rot}(r) = V_{200}$, independent of radius, and that the density profile is $\rho(r) = V_{200}^2/(4\pi G r^2)$. If we assume that a fraction m_d of the halo mass is contained in the disk, we find for the disk mass

$$\begin{aligned} M_d &= m_d\, M = m_d\, \frac{V_{200}^3}{10\, G\, H(z)} \\ &\approx 9\times10^{10} h^{-1} M_\odot \left(\frac{m_d}{0.05}\right) \left(\frac{V_{200}}{200\,\mathrm{km/s}}\right)^3 \frac{1}{E(z)} , \end{aligned} \tag{10.13}$$

where $E(z) = H(z)/H_0$ is the scaled Hubble function. On the other hand, the disk mass follows from (10.12),

$$\begin{aligned} M_d &= 2\pi \int_0^\infty \mathrm{d}R\; R\, \Sigma(R) = 2\pi R_d^2 \Sigma_0 \int_0^\infty \mathrm{d}x\; x\, \mathrm{e}^{-x} \\ &= 2\pi R_d^2 \Sigma_0 \end{aligned} \tag{10.14}$$

where we set $x = R/R_d$ in the last step. In the isothermal density profile of the dark matter halo, the rotational velocity of the disk is constant, and so its angular momentum is

$$\begin{aligned} J_d &= 2\pi \int_0^\infty \mathrm{d}R\, R^2\, V_{200}\, \Sigma(R) = 2\pi V_{200} \Sigma_0 R_d^3 \int_0^\infty \mathrm{d}x\, x^2\, \mathrm{e}^{-x} \\ &= 4\pi V_{200} \Sigma_0 R_d^3 = 2 M_d R_d V_{200} , \end{aligned} \tag{10.15}$$

where in the last step we used (10.14). We assume that the angular momentum of the disk is a fraction j_d of the total angular momentum of the halo, $J_d = j_d J_h$. The latter can be related to the spin parameter λ in (7.65), which in addition contains the total energy of the halo and the halo mass. The total energy follows from the virial theorem and the simple properties of an isothermal sphere, $|E| = M V_{200}^2/2$. We then find

$$\lambda = \frac{J_h |E|^{1/2}}{G M^{5/2}} = \left(\frac{m_d}{j_d}\right) \frac{2 R_d V_{200} |E|^{1/2}}{G M^{3/2}} , \tag{10.16}$$

where we used (10.15) in the last step. Solving for R_d, this yields

$$R_d = \frac{\lambda G M}{\sqrt{2} V_{200}^2} \left(\frac{j_d}{m_d}\right) , \tag{10.17}$$

where we inserted the expression for the binding energy. Finally, using (7.58) again, this can be written in the form

$$\begin{aligned} R_d &= \frac{1}{\sqrt{2}} \left(\frac{j_d}{m_d}\right) \lambda\, r_{200} \\ &= \frac{1}{\sqrt{2}} \left(\frac{j_d}{m_d}\right) \lambda \left(\frac{V_{200}}{10 H(z)}\right) \\ &\approx 7 h^{-1}\,\mathrm{kpc} \left(\frac{j_d}{m_d}\right) \left(\frac{\lambda}{0.05}\right) \left(\frac{V_{200}}{200\,\mathrm{km/s}}\right) \frac{1}{E(z)} . \end{aligned} \tag{10.18}$$

[4] Note that in this case, the baryons are embedded in a dark matter halo, so the consideration of the spin parameter, which applies for the total energy and angular momentum, no longer applies to the baryons only. Therefore, in this case (10.11) does not hold for the baryons alone.

Interpretation. This equation contains a number of interesting aspects. The first expression relates the virial radius of the halo to the scale-length of the disk. If we assume that the average specific angular momentum of the gas in the disk is the same as the average specific angular momentum of the halo, then $j_\mathrm{d} = m_\mathrm{d}$, and we simply get $R_\mathrm{d} = r_{200}\lambda/\sqrt{2}$; using the characteristic value of $\lambda \sim 0.05$, we obtain $r_{200} \sim 30 R_\mathrm{d}$. For the Milky Way, $R_\mathrm{d} \approx 3.5\,\mathrm{kpc}$, so that its virial radius is predicted by this consideration to be about 100 kpc.

The final expression in (10.18) relates the virial velocity—for the assumed isothermal distribution, this is the same as the rotational velocity—to the scale-length of the disk. Again using the Milky Way as an example, for which $V_\mathrm{rot} \approx 220\,\mathrm{km/s}$, we see that the predicted scale length is about a factor of two larger than the observed one, for the same parameters. Thus, although this simple model provides a result which is within a factor ~ 2 of the observed properties of the Milky Way disk, it fails to yield an accurate quantitative agreement. Of course it is possible that our Galaxy formed inside a halo where the spin parameter has a rather low value, or that the disk fraction of angular momentum is different from its mass fraction. For example, if we keep the assumption $j_\mathrm{d} = m_\mathrm{d}$, a spin parameter of $\lambda \sim 0.02$ would predict roughly the correct scale length, but such low values of λ have a rather small probability to occur. Furthermore, it would also lead to a large virial radius of $\sim 250\,\mathrm{kpc}$, predicting a very massive halo for the Milky Way.

However, there is another issue of (10.18) which does not really fit the observations. The function $E(z)$ at redshift $z = 1$ is $E(1) = \sqrt{8\Omega_\mathrm{m} + \Omega_\Lambda} \sim 1.7$, implying that galactic disks at that epoch are considerably smaller than those today. Such a strong size evolution of disks is not consistent with the observations.

A third issue with this simple consideration is the mass fraction of baryons that end up in the disk. With a disk mass of $M_\mathrm{d} \sim 5 \times 10^{10} M_\odot$ for the Milky Way, (10.13) predicts about $m_\mathrm{d} \sim 0.02$. If we assume that the halo contained the same baryon fraction as the cosmic mean at halo formation, then only about 10 % of the baryons end up in the disk. As we shall see later, there are processes which prevent gas from settling down is a disk, but it is difficult to find such processes efficient enough to hold back 90 % of the baryons.

Refinements of the model. The simplified model made a number of assumption which we know can not be correct in detail: Real dark matter halos do not have an isothermal profile, but follow approximately an NFW profile in which the rotational velocity is a slow function of galactocentric radius. In addition, the contraction of gas changes the overall gravitational potential of the halo, which also affects the dark matter distribution; the dark matter also gets somewhat more concentrated towards the halo center. This halo contraction will change the rotational velocity further.

The rotation curves of spiral galaxies show that the neglect of the disk self-gravity is an oversimplification. Within the optical radius of a disk, the baryons in the disk contribute substantially to the gravitational field. This is in accord with what we have learned from gravitational lensing studies of galaxies which show that within the Einstein radius, about half the mass is contributed by the baryonic component. Numerical simulations of disk galaxy formation which take the gas cooling and halo contraction into account indicate that the rotational velocity of disks is closely approximated by the maximum rotational velocity of an NFW profile (see Fig. 7.19) instead of the virial circular velocity.

Both, inclusion of self-gravity and the halo contraction lead to larger rotational velocities in the inner part of the halo compared to the simple model. As a consequence, the size and mass of the halo is smaller than obtained from the simple model, so that the corresponding estimate of m_d is increased. The proper inclusion of these two effects also yields a much smaller redshift-dependence of the scale-length than predicted by (10.18), i.e., considerably closer to the observational situation.

We thus conclude that the model described here, once accounting for the effects of disk self-gravity and halo contraction, provides a good quantitative model for understanding the formation of disk galaxies.

10.4.3 Dynamical effects in disks

Once the disk has formed, the gas is sufficiently dense so that star formation can proceed; we have seen in Sect. 3.3.3 before that the Schmidt-Kennicutt law describes the star-formation rate (per unit disk area) as a function of surface mass density. Hence, after some time a thin stellar disk is formed, with some fraction of the baryons left over in the form of gas.

Such a thin disk is subject to dynamical instabilities. Whereas in an axi-symmetric gravitational potential, stars move on circular orbits, perturbations of the gravitational field can perturb these orbits, which in turn can amplify the deviation from axial symmetry. The formation of spiral arms is one example of such perturbations. Another important aspect is the formation of bars in the center of a large fraction of spiral galaxies. The asymmetry of the bars can yield significant perturbations of the potential with corresponding changes of orbits, leading to a redistribution of mass and angular momentum. In particular, bars can cause stars and gas to migrate inwards, towards the center.

Pseudo-bulges. The corresponding accumulation of gas can trigger increased star formation in the center of galaxies. These stars then form a concentration at the galactic center. It is generally believed that this is the mechanism for the

formation of pseudo-bulges in spiral galaxies—we recall that bulges are divided into classical bulges and pseudo-bulges, the latter being characterized by a Sérsic-index close to unity and fast rotation, whereas the former ones have a Sérsic-index close to that of ellipticals and considerably slower rotation. The formation of classical bulges is thus suspected to be related to the formation of elliptical galaxies, which will be discussed below.

Heating of the stellar distribution. We have seen in Sect. 2.3.1 that the velocity dispersion of stars in the Milky Way disk depends on their age—the older the stars, the higher their random velocities. Stars are formed by the molecular gas which is observed to have the thinnest distribution. Over their lifetime, the stars can gain a random velocity component, by scattering on the perturbations of the gravitational potential, such as caused by giant molecular clouds, spiral arms, or the subhalo population that we discussed in Sect. 7.8. Whatever the main source of heating, the trend with stellar age is expected in all these cases.

10.4.4 Feedback processes

Although the story as told above naturally leads to the formation of disk galaxies, early studies have shown that some ingredients are missing. In fact, hydrodynamical simulations of disk formation show that star formation in the gas disks is far too efficient, consuming the available gas in too short a time, so that most of the stars would be formed at high redshift, with little current star formation left. Furthermore, the resulting disks are too concentrated and too small, leading to rotation curves which are declining outwards beyond the (small) half-light radius of the disk, in marked contrast with observed rotation curves. This together is known as the overcooling problem in galaxy evolution. Real disk galaxies have a slower conversion of gas into stars and their disks remain larger. And finally, the efficient conversion of gas into stars in our simple model would predict that the stellar mass density in the Universe is much higher than observed—whereas $\Omega_{\rm b} \sim 0.04$, the density parameter in stars is less than 1 %. Hence, most baryons in the Universe habe not been converted to stars.

Feedback by supernovae. In order to balance the efficient gas cooling, heating sources need to be considered. An unavoidable source of heating is the energy injected into the interstellar medium by supernovae. Very shortly after star formation sets in, the most massive stars of the stellar population undergo a core-collapse supernova. The mechanical energy of the explosion is partly transferred to the gas surrounding the exploding star. Thereby the gas is heated, causing it to expand, thus to decrease its density, which in turn reduces its cooling efficiency. Note that this is a feedback process—the higher the star formation rate, the more energy is injected into the interstellar gas to prevent, or at least delay, further star formation. Depending on the efficiency of this feedback, the local gas of the disk may be blown out of the disk into the halo (and produce a hot gas corona outside the disk—see Sect. 3.3.7), or, in particular for low-mass halos, be removed from the halo through outflowing gas.

In fact, there is direct observational evidence of the occurrence of outflows from star-forming galaxies. For example, we have seen in Sect. 9.1.1 that the spectra of Lyman-break galaxies reveal substantial mass outflows, at a similar rate as their star-formation rate and with velocities of several hundreds of km/s.

The details of this feedback process are somewhat uncertain—how much of the supernova energy is converted into heat, and how much is transferred to the interstellar medium in form of bulk kinetic energy, is not well determined. Furthermore, the feedback by supernovae depends on the assumed initial mass function (IMF; see Sect. 3.5.1) of stars, which yields the fraction of newly formed stars which explode as core-collapse supernova. The flatter the IMF at the high-mass end, the more supernova energy per unit mass of newly formed stars is injected.

Assuming a universal IMF, the energy released by supernovae per unit mass of newly-formed stars is $\eta_{\rm SN} E_{\rm SN}$, where $\eta_{\rm SN}$ denotes the expected number of supernovae per unit mass of formed stars, and $E_{\rm SN}$ is the energy released per supernova. If we assume that this energy reheats some of the cold gas back to virial temperature of the halo, the amount of gas that is reheated after formation of a group of stars with mass Δm_* is

$$\Delta m_{\rm reheat} \sim \epsilon \, \frac{\eta_{\rm SN} \, E_{\rm SN}}{V_{200}^2} \Delta m_* \,, \tag{10.19}$$

where ϵ parametrizes the efficiency of the reheating process. The reheated gas may be transferred back to the hot gaseous halo, whereas other models assume that the reheated gas is first ejected from the halo, and only later reincorporated into the hot halo on the dynamical time-scale of the halo. This ejection scenario effectively delays the time at which the reheated gas can cool and becomes available for star formation again.

As can be seen from (10.19), supernova feedback is more efficient at suppressing star formation in low-mass galaxies—which is due to the fact that the binding energy per unit mass is an increasing function of halo mass. This simply expresses the fact that for low-mass halos it is easier to drive the gas outwards.

AGN feedback. Whereas supernova feedback explains a decreasing conversion of gas into stars with decreasing halo mass, and thus can account for the difference of the slopes between the galaxy luminosity function and the halo mass function at the low mass/luminosity end (see Fig. 10.2), it is less efficient for higher-mass halos, due to the larger V_{200} in (10.19). The increase of the cooling time for higher-mass halos (see Fig. 10.4) by itself cannot account for the abrupt exponential decrease of the galaxy luminosity function beyond L^*. One requires another process which delays the cooling of gas in high-mass halos.

For very massive halos, we have already encountered such a process: The suppression of cooling flows in galaxy clusters is due to AGN activity of the central galaxy in the cluster. Since (almost) all massive galaxies contain a supermassive black hole (see Sect. 3.8), this kind of feedback may be operational not only in groups and clusters, but actually in individual massive galaxies as well. In particular, there is a great deal of evidence for a relation between nuclear starbursts in galaxies and AGN activity. The gas needed for a starburst in the center of a galaxy is also potential fuel for the central black hole. Again, the details of this process are quite uncertain, but with plausible prescriptions, the cut-off of the luminosity function at $L \gtrsim L^*$ can be successfully modeled.

Feedback by an AGN can occur in several ways. In the case of galaxy clusters, the major effect of the AGN is the insertion of hot bubbles into the intracluster medium through radio jets. The AGNs in most central cluster galaxies are not very luminous, and seem to be in the 'radio mode' (see Sect. 5.5.5) of low accretion rate. Thus, for low accretion rates, the main channel of feedback is the injection of mechanical energy into the surrounding gas. At high accretion rates, in the 'quasar mode', the main source of feedback is presumably heating of the gas. Furthermore, the strong radiation field from quasars changes the ionization structure of the surrounding gas, which affects its cooling curve compared to the one shown in Fig. 10.3 and at low temperatures actually leads to radiative heating. These various effects should be included in realistic models of the evolution of galaxies, at least in an approximate way; we shall come back to this below.

10.4.5 The formation and evolution of supermassive black holes

Black holes grow in mass by accreting material, a process we witness through the radiation from accreting black holes in AGNs (Chap. 5). Hence, once a population of supermassive black holes (SMBHs) is present, their evolution can be studied observationally, as well as through modeling. But how did the first generation of SMBH form? There is no firm conclusion on this question, but three plausible formation processes have been studied in detail. What we do know, however, is that the first SMBHs must have formed very early in the Universe, as indicated by the presence of very luminous QSOs at $z > 6$.

Remnants of population III stars. The first stars in the Universe form out of primordial gas, i.e., gas with zero metallicity. The cooling properties of this gas are quite different from those of enriched material, since no metal lines are available for radiating energy away. From simulations of star formation in primordial gas, it is suggested that many stars can form with very high masses, well above $100 M_\odot$. These stars burn their nuclear fuel very quickly, in a few million years, before they end their lives explosively. If the mass of a star is above $\sim 250 M_\odot$, its supernova will leave a black hole behind with a mass of $\gtrsim 100 M_\odot$. Since the first stars are expected to form at $z \gtrsim 20$, this formation mechanism would yield a very early population of seed black holes. However, it is still unknown whether such very massive population III stars indeed formed.

Gas-dynamical processes. Another route for the formation of supermassive black holes arises if the primordial gas in a high-redshift dark matter halo manages to concentrate in its center, through global dynamical instabilities (e.g., related to the formation of bar-like structures) that are able to transport angular momentum outwards. This angular momentum transport is needed since otherwise, the central concentration of gas would be prevented by the angular momentum barrier. Subsequent cooling by molecular hydrogen may then lead to the formation of a rapidly rotating supermassive star with up to $10^6 M_\odot$, provided the accumulation of the gas occurs rapidly enough. Once the inner core of this supermassive star has burned its hydrogen, the core will collapse and form a black hole with a few tens of $M_\odot$, where this mass depends on the initial angular velocity of the star. This black hole subsequently accretes material from the outer layers of the star, and this quasi-spherical accretion has a very low radiative efficiency ϵ. Therefore, the black hole can grow in mass quickly, until finally its exceeds the Eddington luminosity and the remaining gas is expelled, leaving behind a SMBH with $\sim 10^5 M_\odot$.

Stellar-dynamical processes. In the inner part of a forming galaxy, dense nuclear star clusters may form. Because of the high density, star-star collisions can occur which can lead to the formation of very massive stars with mass exceeding $10^3 M_\odot$. This has to happen very quickly, before the first stars explode as supernovae, since otherwise the massive star would be polluted with metals, its opacity increased, and it would no longer be stable. The fate of this supermassive star is then similar to the scenario described above, resulting in a black hole remnant of several hundred Solar masses.

These three possibilities are not mutually exclusive. At present, our theoretical understanding of these processes is not sufficient to establish their likelihood of occurrence. Whereas one may be able to distinguish between these scenarios, e.g., from the statistics of black hole masses in present day low-mass galaxies, the current observational situation does not conclusively support or reject any of these three routes.

Mass growth. Once the seed black holes have formed, they can grow in mass by accreting material. We saw in Sect. 5.3.5 that the characteristic time-scale for mass growth, i.e., the time on which the black hole mass can double, is $\epsilon t_{\rm gr} = \epsilon M_\bullet c^2/L_{\rm edd} \approx 5\epsilon \times 10^8$yr. With $\epsilon \sim 0.1$, a $10^4 M_\odot$ seed black hole formed at $z \sim 20$ could grow to a few $\times\, 10^8 M_\odot$ by redshift 7 if it accreted continuously at the Eddington rate. The situation is more difficult for seed black holes formed from population III stars; they probably require super-Eddington accretion rates to be able to power the luminous QSOs at $z > 6$. As mentioned in Sect. 5.3.5, the accretion rate may exceed the Eddington rate though probably not by a large factor.

10.4.6 Cosmic downsizing

The hierarchical model of structure formation predicts that smaller-mass objects are formed first, with more massive systems forming later in the cosmic evolution. As discussed before, there is ample evidence for this to be the case; e.g., galaxies are in place early in the cosmic history, whereas clusters are abundant only at redshifts $z \lesssim 1$. However, looking more closely into the issue, apparent puzzles are discovered. For example, the most massive galaxies in the local Universe, the massive ellipticals, contain the oldest population of stars, although at first sight, their formation should have occurred later than those of less massive galaxies. In turn, most of the star formation in the local Universe seems to be associated with low- or intermediate-mass galaxies, whereas the most massive ones are passively evolving. Now turning to high redshift: for $z \sim 3$, the bulk of star formation seems to occur in LBGs and SMGs, which, according to their clustering properties (see Sect. 9.1.1), are associated with high-mass halos. The study of passively evolving EROs indicates that massive old galaxies were in place as early as $z \sim 2$, hence they must have formed very early in the cosmic history. The phenomenon that massive galaxies form their stars in the high-redshift Universe, whereas most of the current star formation occurs in galaxies of lower mass, has been termed 'downsizing'. We saw in Sect. 5.6.2 that a similar phenomenon also is observed for AGNs.

This downsizing can be studied in more detail using redshift surveys of galaxies. The observed profile of the absorption lines in the spectra of galaxies yields a measure of the characteristic velocity and thus the mass of the galaxies (and their halos). Studies carried out in the local Universe showed that local galaxies have a bimodal distribution in color (see Sect. 3.1.3), which in turn is related to a bimodal distribution in the specific star-formation rate. Extending such studies to higher redshifts, by spectroscopic surveys at fainter magnitudes, we can study whether this bimodal distribution changes over time. In fact, such studies reveal that the characteristic mass separating the star-forming galaxies from the passive ones evolves with redshift, such that this dividing mass increases with z. For example, this characteristic mass decreased by a factor of ~ 5 between $z = 1.4$ and $z = 0.4$. Hence, the mass scale above which most galaxies are passively evolving decreases over time, restricting star formation to increasingly lower-mass galaxies.

Studies of the fundamental plane for field ellipticals at higher redshift also point to a similar conclusion. Whereas the massive ellipticals at $z \sim 0.7$ lie on the fundamental plane of local galaxies when passive evolution of their stellar population is taken into account, normal ellipticals of lower mass at these redshifts have a smaller mass-to-light ratio, indicating a younger stellar population. Also here, the more massive galaxies seem to be older than less massive ones. To reproduce these evolutionary effect requires to account for AGN feedback in models of galaxy evolution.

10.5 Formation of elliptical galaxies

Properties of ellipticals. Whereas the formation of disk galaxies can be explained qualitatively in a relatively straightforward way, the question of the formation of ellipticals is considerably more difficult to answer. Stars in ellipticals feature a high velocity dispersion, indicating that they were not formed inside a cool gas disk, or that the stellar distribution was subsequently heated very strongly. On the other hand, it is hard to comprehend how star formation may proceed without gas compression induced by dissipation and cooling.

In Sect. 3.4.3 we saw that the properties of ellipticals are very well described by the fundamental plane. It is also found that the evolution of the fundamental plane with redshift can almost completely be explained by passive evolution of the stellar population in ellipticals. In the same way, we stated in Sect. 6.8 that the ellipticals in a cluster follow a very well-defined color-magnitude relation (the red cluster sequence), which suggests that the stellar populations of ellipticals at a given redshift all have a similar age. By comparing the colors of stellar populations in ellipticals with models of population synthesis, an old age for the stars in ellipticals is obtained, as shown in Fig. 3.35

Monolithic collapse. A simple model is capable of coherently describing these observational facts, namely the monolithic collapse. According to this description, the gas in a halo is nearly instantaneously transformed into stars. In this process, most of the gas is consumed, so that no further generations of stars can form later. For all ellipticals with the same redshift to have nearly identical colors, this formation must have taken place at relatively high redshift, say $z \gtrsim 2$, so that the current ellipticals are all of essentially the same age. This scenario thus requires the formation of stars to happen quickly enough, before the gas can accumulate in a disk. The process of star formation remains unexplained in this picture, however, and most likely this model does not describe the processes that are responsible for the formation of ellipticals.

Instead, we have very good reasons to believe that elliptical galaxies form as a consequence of galaxy transformations. For example, we have seen that most ellipticals are found in dense environments, like groups and clusters, and within these high-mass structures, they are concentrated towards their center. In other words, elliptical galaxies are located in regions where, due to the enhanced density, interactions of galaxies happen preferentially. Furthermore, elliptical galaxies have rather complicated kinematics, often exhibiting small disks (sometimes counter-rotating) around their center, shells and ripples, which indicate a lively history of these objects. From a theoretical view, hierarchical structure formation predicts that high-mass halos are formed by merging of smaller ones, and so the collision of halos and their embedded galaxies must play a role in the distribution of galaxy properties. We shall therefore take a closer look at such halo mergers.

10.5.1 Merging of halos and their galaxies.

When two halos merge to form one with larger mass, their baryonic components will be affected as well. We have seen spectacular examples of this process in the form of colliding galaxies (e.g., Fig. 1.16). Clearly, after the two spiral galaxies collided, the resulting stellar distribution does not resemble that of a spiral anymore. Mergers of halos, and associated collision of galaxies, lead to morphological transformation of galaxies. Furthermore, such galaxy collisions are generally accompanied by massive star bursts. Hence, also the stellar population of the resulting object is affected by collisions.

In the Antennae (see Fig. 9.25), the mass of the two galaxies which collide is about equal. However, one expects that the collision of galaxies with very different masses is more frequent, and such mergers will have different consequences for the respective galaxies. One thus distinguishes between *minor mergers*, where the mass ratios of halos is large (typically in excess of 3:1), and major mergers where the two masses are similar.

Conditions for merging. Not every (near) collision of two halos leads to a merger. For example, we have seen in the bullet cluster (Sect. 6.6.2) that the two clusters simply move through each other, since their dark matter and stellar components are collisionless. Only the (collisional) gas components of the two clusters are strongly affected by this collision, but no merging will take place. The reason is that the relative velocity of these two clusters at collision is much larger than their internal velocity dispersion, or expressed differently, that the collision speed is much higher than the escape velocity of each cluster component.[5] In order for a merger to happen, the collisional speed has to be of the same order, or smaller, than the intrinsic velocity dispersion. This implies that effective mergers of galaxies do not occur in massive clusters, where the velocity dispersion of the galaxies of the cluster—which is also the characteristic collision velocity—is considerably higher than the stellar velocity dispersion of the individual galaxies. In contrast, groups of galaxies have both, a high density of galaxies making collisions probable, and a sufficiently low velocity dispersion to enable the merging of galaxies. Hence we expect that the most efficient merging of galaxies happens in groups.

Minor mergers. Consider what may happen in the merging of two halos with their embedded galaxies. The outcome of a merger depends on several parameters, like the relative velocity, the impact parameter, the angular momenta, the orientation of their rotation, and particularly the mass ratio of the two merging halos. If a smaller galaxy merges with a massive one, the properties of the dominating galaxy are expected to change only marginally: the small galaxy will be embedded into the bigger halo, and survive as a satellite galaxy for a long while. Examples of this are the Magellanic Clouds, which orbit around the center of the Milky Way in its dark matter halo. Depending on the orbit of the satellite galaxy, it will not survive forever. Tidal forces strips matter from the outer parts of the satellite's dark halo, which is thus expected to lose mass—the closer it orbits near the center, the stronger the tidal forces, and thus the higher the mass-loss rate.

Dynamical friction (see Sect. 6.3.3) acts on the satellite, causing it to lose orbital energy and angular momentum, which is transferred (mostly) to the dark matter halo of the massive collision partner. The satellite slowly migrates towards the center, and gets disrupted due to the stronger

[5] In this case of high collision velocity, the time it takes a galaxy from one of the two clusters to cross the gravitational potential of the other cluster is shorter than the time it takes the matter of the second cluster to react to the changing conditions caused by the merger; therefore, the gravitational potential of the second cluster can be considered almost stationary during the collision process. Thus, the galaxy leaves the potential of the second cluster with almost the same velocity it had on entering, i.e., it is not gravitationally bound to the second cluster.

tidal forces there. The stars of the satellite galaxy are simply added to the stellar population of the massive galaxy, since the stars of the satellite have a small velocity dispersion, they are added as coherent 'streams' to the main galaxy (see Fig. 3.17). Such a 'minor merger' is currently taking place in the Milky Way, where the Sagittarius dwarf galaxy is being torn apart by the tidal field of the Galaxy, and its stars are being incorporated into the Milky Way as an additional population. This population has, by itself, a relatively small velocity dispersion, forming a cold stream of stars that can also be identified as such by its kinematic properties. However, the large-scale structure of the Galaxy is nearly unaffected by a minor merger like this.

The thick disk and the stellar halo. Spiral galaxies have, beside the thin stellar and gas disk, also a thick disk with distinct properties: it has a substantially larger scale-height (by a factor of $\sim$ 3) and a stellar population with lower metallicity and old age. Thick disks have been explained by a number of different models. For example, they could consist of stars formed in the thin disk, and being heated so strongly that their vertical velocity dispersion causes this population to thicken substantially. However, the clearly different age distribution of thick-disk stars provides an obstacle for this explanation which rather predicts a continuous transition from thin to thick-disk stars. Nevertheless, the satellite galaxies and their associated subhalos may well be a substantial source of heating.

Minor mergers provide an alternative explanation for the origin of thick disks. Due to dynamical friction, satellite galaxies are dragged into the plane of the disk of the parent galaxy, and their subsequent disruption leaves their stars in the plane of the disk. As the minor merger partner is of low mass, the age of the thick disk is expected to be old—we have seen that low-mass halos preferentially form their stars very early in cosmic history, before heating by an ionizing background radiation prevents efficient star formation. It is thus conceivable that the stars of the thick disk, and also those of the stellar halo, are relics of earlier minor mergers. The fact that an increasing number of stellar streams are found in the Milky Way and other neighboring galaxies, as well as numerical simulations, support this picture.

Thus, in summary, minor mergers do not alter the properties of the major collision partner strongly. The dark matter halo increases its mass, in the form of subhalos (which later on may be disrupted), the stellar population of the low-mass galaxy first forms a satellite galaxy, which later can be disrupted and added to the stellar population of the parent galaxy, probably with somewhat different kinematical properties.

Major mergers and morphological transformations of galaxies. The situation is different in a merger process where both partners have a comparable mass. In such 'major mergers' the galaxies will change completely. The disks will be destroyed, i.e., the disk population attains a high velocity dispersion and can transform into a spheroidal component. Furthermore, the gas orbits are perturbed, which may trigger massive starbursts like, e.g., in the Antenna galaxies. By means of this perturbation of gas orbits, the SMBH in the centers of the galaxies can be fed, initiating AGN activity, as it is presumably seen in the galaxy Centaurus A shown in Fig. 10.12. Due to the violence of the interaction, part of the matter is ejected from the galaxies. These stars and the respective gas are observable as tidal tails in optical images or by the 21 cm emission of neutral hydrogen. From these arguments, which are also confirmed by numerical simulations, one expects that in a 'major merger' an elliptical galaxy may form. In the violent interaction, the gas is either ejected, or heated so strongly that any further star formation is suppressed.

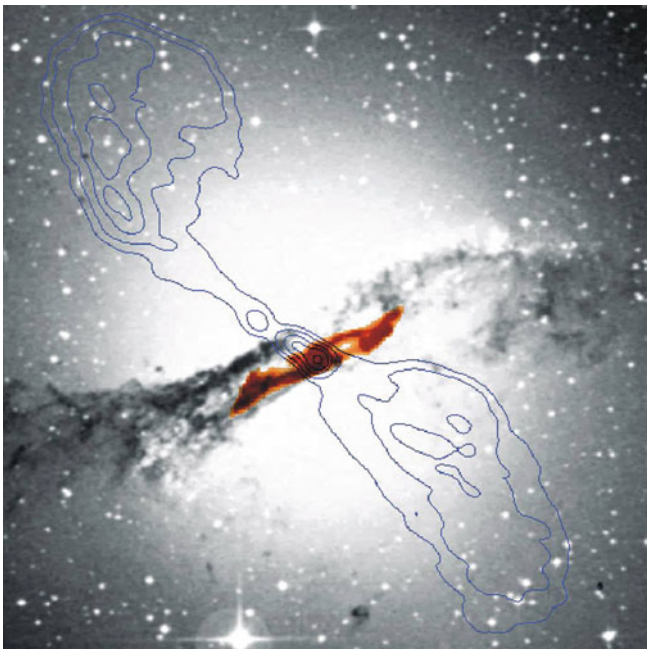


Fig. 10.12 The galaxy Centaurus A. The optical image is displayed in grayscales, the contours show the radio emission, and in *red*, an infrared image is presented, taken by the ISO satellite. The ISO map indicates the distribution of dust, which is apparently that of a barred spiral. It seems that this elliptical galaxy features a spiral that is stabilized by the gravitational field of the elliptical. Presumably, this galaxy was formed in a merger process; this may also be the reason for the AGN activity. Credit: ESA/ISO, ISOCAM Team, I.F. Mirabel and O. Laurent (CEA/DSM/DAPNIA), et al. 1998, astro-ph/9810419

Dry vs. wet mergers. However, the situation is slightly more complicated than this. The violent starbursts, associated with the collision of gas-rich galaxies, generate a population of newly-born stars. If such mergers happen at redshifts $z \lesssim 2$, the stellar population of the resulting galaxy may not resemble the 'dead and red' properties of observed ellipticals. Therefore, if ellipticals are formed through major

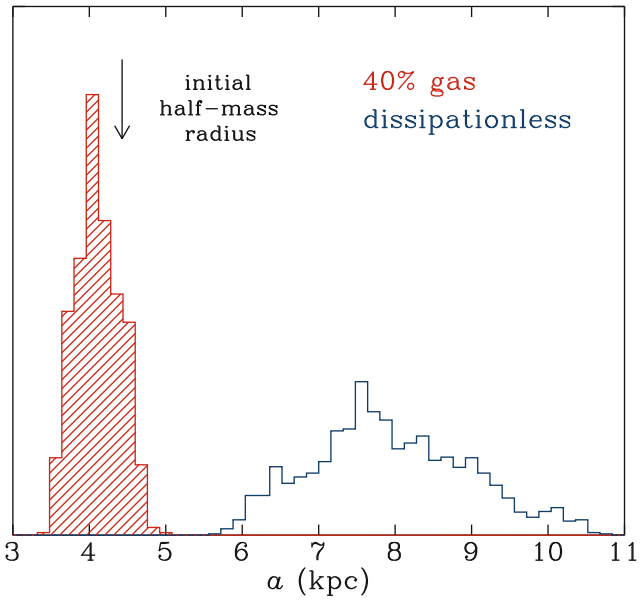


Fig. 10.13 Resulting distribution of the (half-light) semi-major axis a of merger remnants. The merging of identical disk galaxies was simulated, using a distribution of initial conditions, concerning orbital parameter and orientation of the disks. The merger remnants have fairly elliptical isophotes. Two families of simulations were considered: in the first one, the stellar disk of the progenitor galaxies consisted only of stars (dissipationless), whereas for the second family, a gas fraction of 40 % was assumed. The figure shows that the stellar distributions from dissipationless merger remnants have a rather large size, considerably larger than elliptical galaxies (the *arrow* indicates the half-mass radius of the progenitor disks). The inclusion of gas and corresponding cooling and star formation drastically changes this distribution towards considerably smaller sizes, in agreement with observations. Source: T.J. Cox et al. 2006, *The Kinematic Structure of Merger Remnants*, ApJ 650, 791, p. 795, Fig. 3. ©AAS. Reproduced with permission

mergers of gas-rich galaxies, that had to happen at an early epoch. One often calls the mergers where the two progenitor galaxies are gas-rich 'wet' mergers, and contrasts them to 'dry' mergers where gas plays only a small role.

Besides the issue of star formation, wet mergers are characterized by the dissipational properties of the gas. The associated friction can lead to higher spatial densities than it is possible for collisionless matter only. The gas can be driven towards the center of the merger remnant, condense there and form new stars. This process increases the matter density relative to the case of dry mergers.

Early numerical simulations of galaxy mergers considered just the collisionless matter. Although the merger remnants resembled elliptical galaxies in many respects, in detail they differed from real ellipticals. For example, the resulting sizes were considerably larger than those of ellipticals (see Fig. 10.13). However, when merger simulations including gas physics became possible, the situation changed drastically. As we can see from Fig. 10.13, the inclusion of gas leads to considerably more concentrated merger remnants, in accord with observed properties of ellipticals. This is because the gas condenses in the central region of the merger remnant and forms stars there, yielding a higher mass (and stellar) concentration. Furthermore, as illustrated in Fig. 10.14, the distribution of the ellipticities of the stellar distribution in the remnant is changed significantly and much better resembles that found in observations. Wet mergers lead to considerably larger rotational velocities and central velocity dispersions than dry mergers, again in agreement with observations.

A further strong difference between dry and wet mergers is the distribution of merger remnants with regards to their ratio of rotational velocity and velocity dispersion, and the projected ellipticity of the stellar light. We infer from Fig. 10.15 that dry mergers of disk galaxies predict far too small rotation of ellipticals when compared to observations, whereas wet mergers astonishingly well reproduce the observed distribution. Simulations like these therefore yield strong support for the merger hypothesis as the origin of elliptical galaxies. The required high gas fraction of the disk is a natural consequence of the requirement that these wet mergers have to happen early in cosmic history, to reproduce the old stellar population of current ellipticals. At high redshift, a smaller fraction of the gas has yet been converted into stars; thus, high-redshift disks are expected to be more gas rich than current spiral galaxies. Indeed, we saw in Sect. 9.4.4 that the gas-mass fraction of high-redshift galaxies is considerably higher than that of local ones.

Still, this is not the full story. Whereas the properties of 'normal' elliptical galaxies are well reproduced by the aforementioned gas-rich merger simulations, they fail to account for some of the characteristics of massive ellipticals, namely that these are slowly rotating and have boxy isophotes. Such objects, on the other hand, *are* produced by (dry) mergers of ellipticals.

The resulting scenario for the formation of ellipticals. Therefore, the following picture emerges: lower-mass normal ellipticals (i.e., not including dwarfs) are formed by wet major mergers of gas-rich (disk) galaxies at high redshift. Such mergers preferentially occur in overdense regions, i.e., in galaxy groups, which explains why ellipticals are preferentially found in groups and galaxy clusters (clusters are mainly formed by merging and accretion of groups, together with the galaxies they contain). In these dense environments, some of the ellipticals merge with other ellipticals, and these dry mergers lead to the formation of more massive galaxies with the characteristics of observed massive ellipticals.[6]

[6]The fact that spectacular images of merging galaxies show mainly gas-rich mergers (such as in Fig. 9.25 or 1.16) can be attributed to selection effects. On the one hand, gas-rich mergers lead to massive star formation, yielding a statistically increased luminosity of the systems, whereas dry mergers basically preserve the luminosity. On the other hand, gas-rich mergers can be recognized as such for a longer period

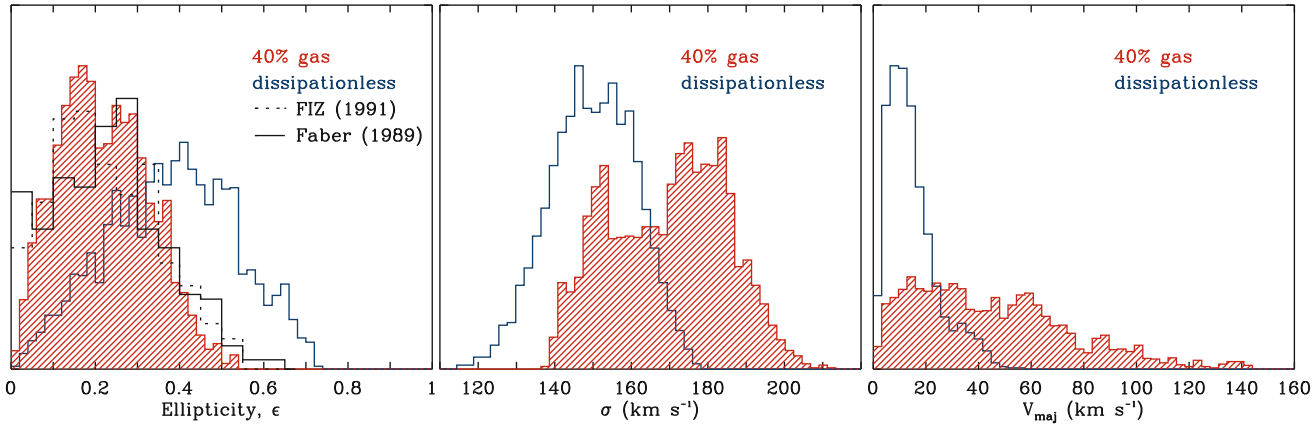


Fig. 10.14 From the same simulations as those described in Fig. 10.13, the distribution of ellipticity (*left panel*), central velocity dispersion (*middle*) and maximum velocity along the major axis (*right*) are shown. In each panel, the *blue curves* are from the dissipationless simulations, whereas for the *red hatched histograms*, gas physics was taken into account. The *black curve* in the left panel depicts the observed distribution of galaxy ellipticities. Source: T.J. Cox et al. 2006, *The Kinematic Structure of Merger Remnants*, ApJ 650, 791, p. 795, Fig. 3. ©AAS. Reproduced with permission

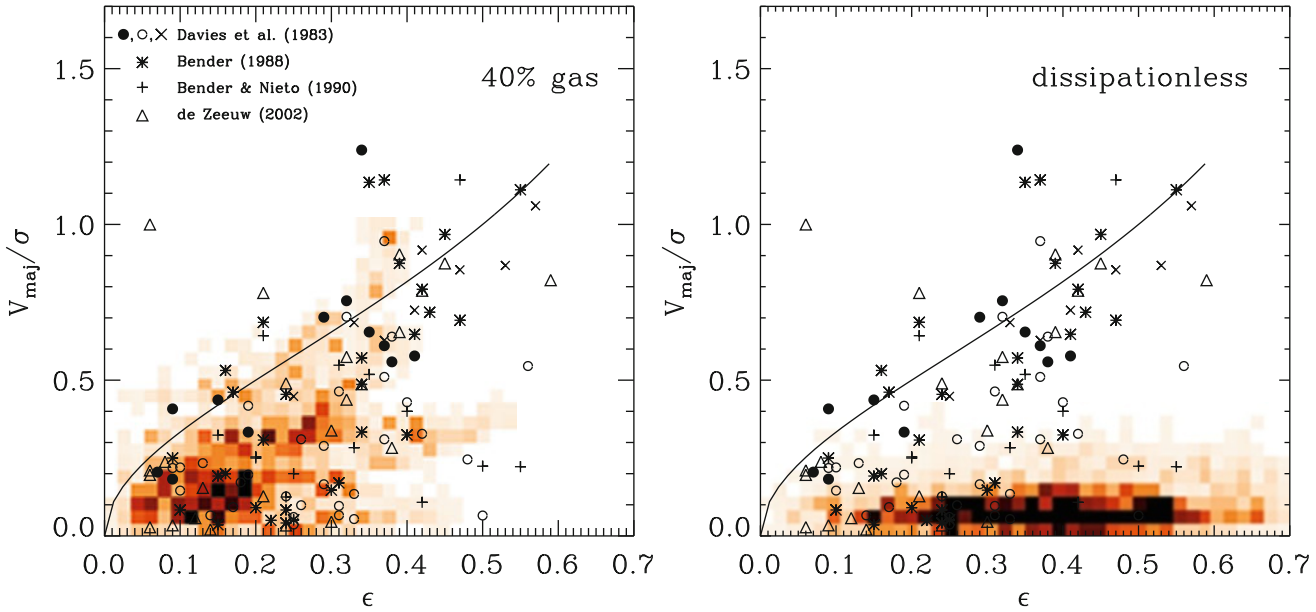


Fig. 10.15 Based on the same simulations as in Fig. 10.13, the distribution of the merger remnants in the parameter plane spanned by the ratio of the maximum rotational velocity along the major axis and the mean velocity dispersion within the half-light radius, and the ellipticity of the half-light ellipse of the stellar distribution of the merger remnant is shown as *shaded areas*. The *left panel* includes gas physics, whereas the *right panel* shows the dissipationless mergers. The *curves* in each panel shows the velocity ratio that would be needed to cause the flattening of ellipticals due to rotational support. Overplotted are the corresponding quantities of several samples of elliptical galaxies. Source: T.J. Cox et al. 2006, *The Kinematic Structure of Merger Remnants*, ApJ 650, 791, p. 797, Fig. 5. ©AAS. Reproduced with permission

Numerical simulations have shown that gas-free mergers preserve the fundamental plane, in the sense that the merging of two ellipticals that live on the fundamental plane will lead to a merger remnant that lies on the plane as well.

Brightness profiles of merger remnants. Support for this picture comes from the brightness profiles of elliptical galaxies. The left panel in Fig. 10.16 shows the surface density profile of stars in the merger remnants. At large radii, they seem to be well described by a de Vaucouleurs profile (or, more generally, by a Sérsic profile), but there are significant differences closer to the center. The profiles of the dissipationless merger remnants near the center lie significantly

of time than dry ones, owing to the clearly visible tidal tails traced by luminous newly formed stars.

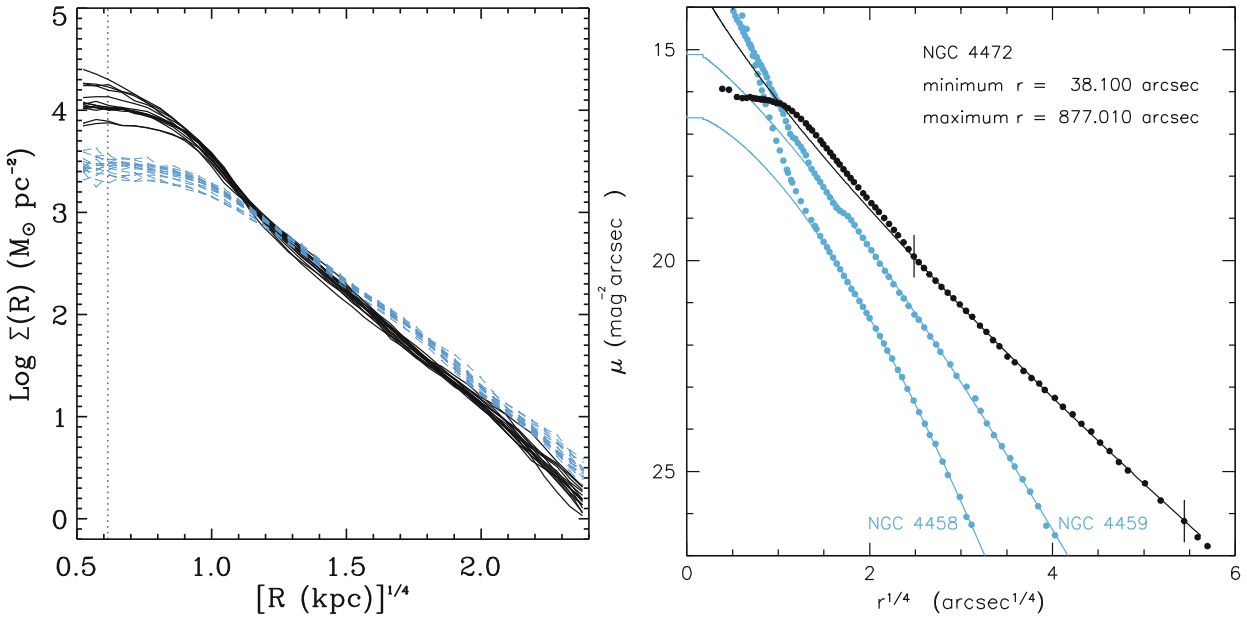


Fig. 10.16 *The left panel* shows the radial density profile of merger remnants, obtained from the gas-rich (*black*) and dissipationless (*blue*) merger simulations that were also considered in the previous figures. *The right panel* shows the corresponding radial surface brightness distribution of three elliptical galaxies in the Virgo cluster. The *black points* correspond to NGC 4472, a core elliptical, with the best fitting Sérsic profile shown as *black curve*. The angular region over which this fit was obtained is indicated by the *short vertical lines*. The two sets of *blue points* and *curves* show the brightness profiles of NGC 4458 and NGC 4459 and the best Sérsic profile fits at large radii, respectively. Source: *Left*: T.J. Cox et al. 2006, *The Kinematic Structure of Merger Remnants*, ApJ 650, 791, p. 796, Fig. 4. ©AAS. Reproduced with permission. *Right*: J. Kormendy et al. 2009, *Structure and Formation of Elliptical and Spheroidal Galaxies*, ApJS 182, 216, p. 274, Fig. 49. ©AAS. Reproduced with permission

below the extrapolation of the de Vaucouleurs profile from larger radii—these profiles have developed a finite core. On the other hand, the density of gas-rich merger remnants is higher in their center than the de Vaucouleurs extrapolation, which can be accounted for by the increased density through the star formation in wet mergers.

Interestingly enough, these two kinds of behavior are also found in elliptical galaxies. In a complete census of all known elliptical galaxies in the Virgo cluster, it was found that *all* the ten brightest galaxies have a core; one example is NGC 4472 shown in the right panel of Fig. 10.16. All of the 17 least luminous normal ellipticals have an excess of light above the extrapolation of the fitted Sérsic profile; two such examples are also shown in Fig. 10.16. The excess light can be explained by the gas dissipation and star formation in wet mergers, whereas dry mergers are not expected to develop such a light excess.

This picture is also supported further by the strong size evolution of elliptical galaxies with redshift (see Fig. 9.47). An elliptical which formed at high redshift by a wet merger is more compact than one which is the result of a dry merger at lower redshift (see Fig. 10.13). Additionally, the fact that the total stellar mass in massive elliptical galaxies is smaller by a factor ~ 3 at $z \sim 1$ than today implies that most of the current ellipticals have formed rather recently—however, not their stellar population which is required to be old. Such an evolution of the population of ellipticals can at least be qualitatively understood with the hypothesis of dry mergers.

Evidence for the importance of mergers for ellipticals is also provided by their small-scale brightness structure. We have seen in Sect. 3.2.5 that many ellipticals show signs of complex evolution which can be interpreted as the consequence of mergers. This is in accord with the picture where the formation of ellipticals in galaxy groups happens by violent merger processes, and that these then contribute to the cluster populations by the merging of groups into clusters.

The rate of mergers can be roughly estimated from the number of close pairs of galaxies with the same redshift. An example of this is found in Fig. 6.68, where several gravitationally bound pairs of early-type galaxies are seen in the outskirts of a cluster at $z = 0.83$. These pairs will merge on a time-scale of $\lesssim 1$ Gyr.

Whereas the impact of a major merger on the fate of a galaxy is dramatic, these events are not the primary process by which galaxies obtain their mass. Most of the mass growth of dark matter halos occurs through minor mergers and accretion of surrounding material, with major mergers contributing at the $\sim 20\,\%$ level. Indeed, from the large population of disk galaxies in the current Universe one

Fig. 10.17 An HST image of NGC 4650A, one out of about 100 known polar-ring galaxies. Spectroscopy shows that the inner disk-like part of the galaxy rotates around its minor axis. This part of the galaxy is surrounded by a rotating ring of stars and gas which is intersected by the polar axis of the disk. Hence, the inner disk and the polar ring have angular momentum vectors that are pretty much perpendicular to each other; such a configuration cannot form from the 'collapse' of the baryons in a dark matter halo. Instead, the most probable explanation for the formation of such special galaxies is a huge collision of two galaxies in the past. Originally the disk may have been the disk of the more massive of the two collision partners, whereas the less massive galaxy has been torn apart and its material has been forced into a polar orbit around the more massive galaxy. New stars have then formed in the disk, visible here in the bluish knots of bright emission. Since the polar ring is deep inside the halo of the other galaxy, the halo mass distribution can be mapped out to large radii using the kinematics of the ring. Credit: J. Gallagher & the Hubble Heritage Team (AURA/STScI/NASA)

concludes that at least for them, major mergers have played no role in the more recent cosmic history.

Polar ring galaxies. Another class of particular galaxies may provide the clearest indication of a merging process for their formation: polar ring galaxies (see Fig. 10.17). The kinematics of their stellar population cannot be explained by the collapse of gas in a halo, but must be due to an encounter of two galaxies.

The impact of AGN feedback in mergers. The black holes in the center of galaxies can be switched to an active mode if gas can be channeled into the center and subsequently accreted. Due to the angular momentum of gas, this is possible only if the gravitational field is substantially perturbed, either by internal processes in a galaxy (e.g., the presence of a bar), or external perturbations. Indeed, observations of low-redshift QSOs show that they are preferentially found in host galaxies which show signs of tidal interactions. It is therefore natural to expect that AGN activity is promoted by galaxy interactions, in particular by mergers.

The feedback from an AGN, triggered by a merger event, has a substantial impact on the nature of the merger remnant. It can heat and expel the gas from the galaxy, shutting off subsequent star formation, whereas without this feedback mechanism, the merger remnant could still keep a substantial fraction of its gas to support further star formation. This consideration is strongly supported by numerical simulations of such merging events (Fig. 10.18).

Bulge formation. Depending on the masses of the progenitors, the resulting ellipses can have a fairly low mass. If the merger occurred in a region where the galaxy number density is rather small, the resulting small elliptical galaxy can survive for a long time without an additional (major) merger. In that time, together with its dark matter halo, it can accrete additional matter whose baryonic part may be able to cool. In this case, the baryons will undergo the same evolution as we have discussed in the context of disk galaxies before—a gas disk is built up which can then form stars. In this way, a disk galaxy is formed in the center of which one finds a small elliptical 'galaxy': this is the preferred explanation for (classical) bulges in disk galaxies. The bulge-to-disk ratio of these galaxies then depends on the mass of the merger remnant, the time available for accreting mass onto the halo, and the cooling time-scale of the gas.

10.5.2 Black hole binaries

The fate of the central black holes. Elliptical galaxies, or more generally, the spheroidal component of galaxies (i.e., the ellipticals, and the bulge of spirals) are observed to have a central supermassive black hole whose mass scales with the velocity dispersion of the stellar population (see Sect. 3.8). When two such galaxies merge, the behavior of their corresponding black holes is of interest. At first, they will follow the orbit of the progenitor galaxies; later, when the merging is in a later stage, they will orbit around

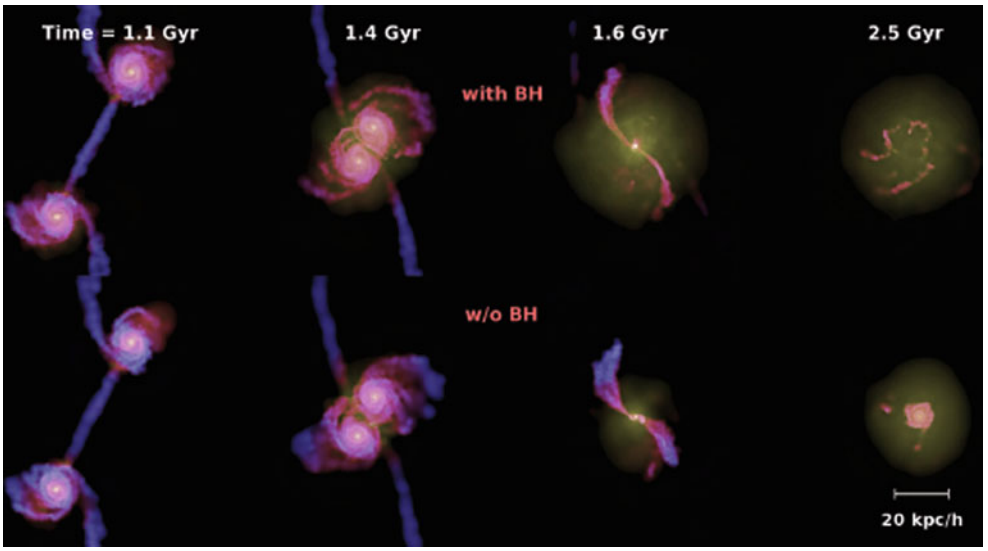


Fig. 10.18 Four different stages of a simulated merging event of two spiral galaxies whose dark matter halos have a virial velocity of 160 km/s. The galaxies also have a stellar bulge and a gas-mass fraction in the disk of 20 %. The *top panel* shows the case that both galaxies contain a supermassive black hole with an initial mass of $2 \times 10^5 M_\odot$, whereas no SMBHs are included in the *lower panel*. A radiative efficiency of 10 % is assumed for the black hole accretion, and that 5 % of this energy can be transferred to heating the gas. In both cases, the gas distribution is shown, and *color* indicates gas temperature, increasing from blue to red to yellow. After their first mutual passage (the first time step shown), the two galaxies show strong signs of interactions, as seen in the tidal tails they develop. Shortly before they merge (second time step), the gas is considerably hotter in the case where SMBHs are included—it is heated by the energy from the AGN. The difference becomes even larger in the later stages of the merging: in the simulation with SMBHs, the gas density is low and heated to a high temperature, so that further star formation in the merger remnant is strongly suppressed: it resembles an early-type galaxy. During the merger, the accretion of gas is very efficient, and the final black hole mass of the merger remnant is $4 \times 10^7 M_\odot$. In the simulation without SMBHs, a substantial amount of cool gas remains to enable ongoing star formation. Source: T. di Matteo et al. 2005, *Energy input from quasars regulates the growth and activity of black holes and their host galaxies*, Nature 433, 604, Fig. 1. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2005

the center of gravity in the merger remnant. One therefore expects to have a supermassive binary black hole orbiting within the newly formed galaxy.

The orbital radius of the binary black hole decreases in time. Owing to the high initial orbital angular momentum, the two SMBHs are, at the beginning of a merger, on an orbit with rather large mutual separation. By dynamical friction (see Sect. 6.3.3), caused by the matter distribution in the newly formed galaxy, the pair of SMBHs will lose orbital energy after the merger of the galaxies, and the two black holes will approach each other. Since this process takes a relatively long time, and since a massive galaxy will, besides a few major mergers, undergo numerous minor mergers, it is conceivable that many of the black holes that were originally the nuclei of low-mass satellite galaxies are today still on orbits at relatively large distances from the center of galaxies. This phase of orbit shrinking is estimated to bring the two black holes within a few parsecs of each other.

The subsequent evolution is less certain. The black hole binary orbit can further shrink through a number of processes. One of them is the interaction with stars. On average, due to the large mass ratio between stars and black holes, energy is transferred from the black holes to the stars, which can obtain enough energy to become gravitationally unbound to the galaxy (which may then lead to the occurrence of hypervelocity stars). This means that they carry away orbital energy which is thus lost from the black holes. In this way, the orbit of the binary black hole becomes tighter, at the expense of evaporating stars from the center of the galaxy. One can estimate that the total mass of ejected stars is of the same order as the black hole masses, and this estimate is further supported by numerical simulations. Hence, there is missing stellar light at the center of massive galaxies (i.e., a core) as found for all the most luminous ellipticals in the Virgo cluster (see the right panel of Fig. 10.16 for an example).

It is also possible that the black hole binary accretes matter from the merged host galaxy and forms a gas disk outside the orbital radius. Due to the strong tidal gravitational field, density waves are generated in this disk, at the expense of orbital momentum of the binary.

These processes can yield a hardening of the binary orbit down to $\sim 10^{-3}$ pc. When this separation is achieved, the black hole binary orbit will continue to shrink efficiently via the emission of gravitational waves (see Sect. 7.9), which then finally will lead to a black hole merger.

Black hole recoil. According to the theory of black holes, there is a closest binary separation at which an orbit still is stable. Once the separation has shrunk to that size, the merging occurs, accompanied by a burst of gravitational wave emission. If the two SMBHs have the same mass, each of them will emit the same amount of gravitational wave energy, but in opposite directions, so that the net amount of momentum carried away by the gravitational waves is zero. However, if the masses are not equal, this cancellation no longer occurs, and the waves carry away a net linear momentum. According to momentum conservation, this will yield a recoil to the merged SMBH, and it will therefore move out of the galactic nucleus. With numerical methods, the recoil velocity can be calculated.[7] It depends on the mass ratio of the two black holes, as well as on their angular momentum and the rotational directions relative to the orbital plane. For two non-rotating black holes, the maximum recoil velocity is $\sim$ 175 km/s, obtained for a mass ratio of $\sim$ 0.36. For rotating black holes, the recoil velocity can be much larger, and in extreme cases (when the black holes have maximum spin and they are anti-aligned) can exceed 4000 km/s.

The recoil will displace the merged black hole from the center of its host galaxy. Depending on the recoil velocity, it may return to the center in a few dynamical time-scales. However, if the recoil velocity is larger than the escape velocity from the galaxy, it may actually escape from the gravitational potential and become an intergalactic black hole. The likelihood of this effect is not quantitatively known, since we know too little about the spin of SMBHs, and these spins can be severely affected during the initial stages of the merging process. The black hole may carry away with it its accretion disk, and remain an active galactic nucleus for some time (say, $\sim 10^6$ yr).

Consequences. The merging of binary SMBHs has a number of consequences. First, it qualitatively predicts that the central supermassive black hole in galaxies should grow in proportion to the mass growth of galaxies due to mergers. This cannot be the full story, since at least some of the mass growth must occur due to accretion of gas in case of wet mergers; however, in wet mergers the galaxies are dominated by (gas rich) disks, and so the corresponding black hole masses are rather small if they follow the $M_\bullet$-σ relation. A second consequence is the existence of binary black holes in at least some galaxy merger remnants, when they are caught in the initial stages of binary hardening. If the two individual SMBHs can retain (or regain) a gas reservoir around them and accrete, they can become active. If only one of them accretes, one might expect an AGN off-center in the merger remnant. Similarly, if the recoil displaces the merged SMBH away from the center of the galaxy, an off-center AGN may become visible.

How frequent such situations occur in mergers is difficult to predict. The occurrence rate depends on the gas content and distribution in the center of the two merging galaxies, and on the time-scale the two black holes orbit in the merger remnant before final coalescence—the longer it takes, the higher the probability to detect a binary AGN.

Observational evidence for binary SMBHs. According to the above expectations, there are a number of possible observational probes for binary SMBHs and their remnants: (1) Two AGNs in the same galaxy. (2) One AGN which is not located in the center of its host galaxy, either because only one of the two black holes is accreting at the time of observations, or because the merged SMBH has been displaced by the recoil effect. (3) One AGN which shows signs of orbital motion, either through a periodicity (with the period being the orbital period), or through double-peaked broad emission lines, which could be formed if both black holes in a close (unresolved) pair are associated with their own broad line region.

Binary AGNs have indeed been found. The radio source 3C 75 shown in Fig. 6.30 has two radio nuclei, both of which launch a pair of jets. These jets are strongly bent, which is interpreted as being due to the motion of the host galaxy through the cluster Abell 400 in which it is embedded. The interaction of the jet plasma with the intracluster medium then deforms the jets in this wide angle tail source. The large projected separation between the two radio nuclei implies that the black hole merging process has not advanced very much in this system.

The galaxy NGC 6240 shown in Fig. 10.19 is a recent merger, as seen from the disturbed morphology. Its large infrared luminosity of $L_{\rm IR} \sim 7 \times 10^{11} L_\odot$ indicates that the merger induced a strong burst of star formation. In the center of the galaxy, two AGNs are seen, revealed by their X-ray emission. The projected separation is $\sim$ 1.4 kpc in this case.

Several more such binary AGNs have been found, with separation of $\sim$ 1 kpc or larger. In most of these systems, the host galaxy shows signs of a recent merger, such as strongly distorted morphology and/or intense star formation. However, one system was found where the separation is much smaller. In the radio galaxy 0402+379 ($z = 0.055$), there are two compact radio sources with a projected separation of 7.3 pc, suggesting that we are witnessing a more advanced merging stage.

Binary black holes candidates have also been claimed from spectral studies of AGNs, where a large velocity shift between the broad and the narrow emission lines was found. One interpretation of these observations is that the shift is due to the active SMBH orbiting in the host galaxy, carrying the

[7] Calculating the behavior of a binary black, using the equations of General Relativity, turns out to be very difficult endeavor. Only since 2005 it has become possible to find numerical solutions of this problem.

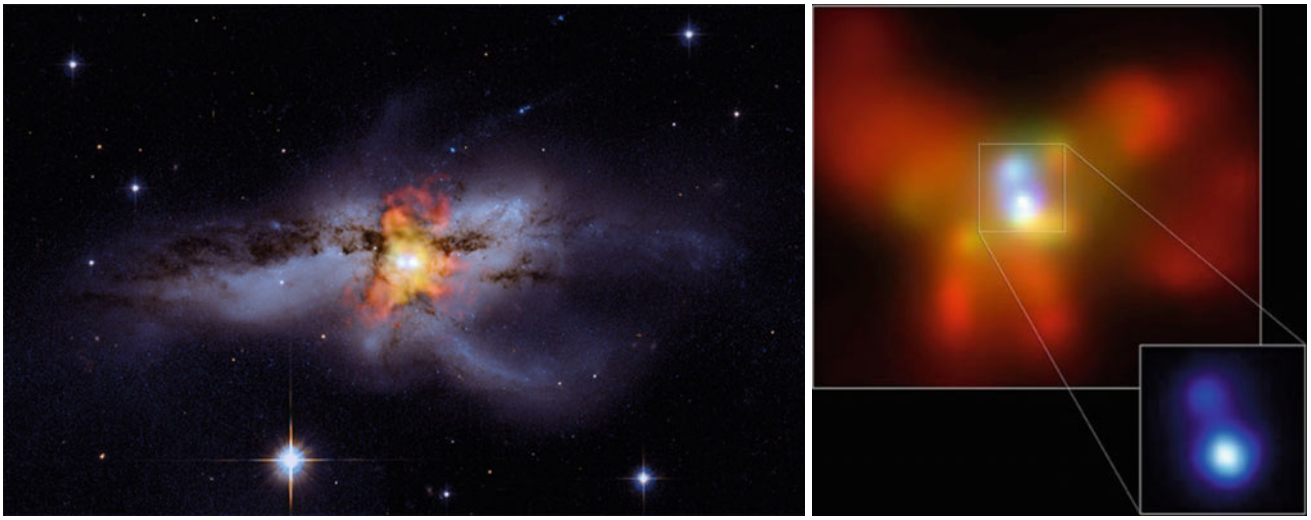


Fig. 10.19 *On the left*, a composite image of the galaxy NGC 6240 ($z = 0.0245$) is shown. The X-ray emission is shown in *red*, *orange* and *yellow*, superposed on an optical image of this galaxy. A pair of two compact X-ray sources in the center, zoomed in *on the right* (at different orientation), shows the presence of two AGNs in this galaxy; their projected separation is $\sim$ 1.4 kpc. With K-band integral field spectroscopy, the black hole mass of the more luminous of the two AGNs has been estimated to be $M_\bullet \sim 9 \times 10^8\, M_\odot$. Credit: *Left*: X-ray (NASA/CXC/MIT/C. Canizares, M. Nowak); Optical (NASA/STScI). *Right*: NASA/CXC/MPE/S. Komossa et al.

broad line region along its orbit, whereas the gas emitting the narrow emission line is at rest in the host galaxy. The active SMBH can either be a member of a binary black hole with the other one inactive, or the merged SMBH which obtained its velocity through recoil. AGNs with double peaked emission lines may also be interpreted as a pair of spatially unresolved active nuclei, where the two peaks of the emission lines reflect the line-of-sight velocity of the two SMBHs. However, there are alternative explanations for the nature of these sources, and the evidence for a binary black hole is not unchallenged.

The galaxy CID-42 (Fig. 10.20) has two bright optical nuclei; one of them is point-like and appears as an AGN, whereas the other is slightly extended and most likely is a nuclear star cluster. The AGN is also seen in X-rays, whereas the other compact optical source has no detectable X-ray emission. The AGN is off-center; in addition, it has broad emission lines which have a velocity offset from the narrow emission lines of $\sim 1300\,\mathrm{km/s}$; note that this velocity is much larger than the orbital velocity of a binary black hole at the separation between these two compact source components. Thus, in this system one has both kinematical as well as positional indications for a SMBH which has been ejected from the center of the galaxy through recoil; it is the best candidate observed so far for this effect.

Another class of sources may indicate the occurrence of binary SMBH mergers, the so-called X-shaped radio sources (see Fig. 10.21). These sources are characterized by their radio morphology, containing two pairs of jets in different directions. The more luminous, inner pair of jets is connected to the central source, whereas the outer jets with lower brightness appears to consist of plasma that was ejected from the core some time in the past. One likely possibility to explain these sources is a change of orientation of the accretion disk, which may be due to a change of the black hole angular momentum vector. During the hardening of the black hole binary, the interaction of the black holes with the gas inside the host galaxy may cause such spin flips.

An accreting SMBH in orbit around the center of the host galaxy can produce periodicity in its emission. The best example yet found is the blazar OJ 287 at $z = 0.306$. Variability of this object has been traced back to 1890, using archived photographic plates, and it shows a periodicity of 11.86 yr. Merger models of the source that explain the periodicity involve a second SMBH with about 10 times lower mass. If the period of variability is identified with the orbital period, then this binary will merge over the next $\sim 10^5$ yr. However, the periodicity can have a different origin, like a precessing accretion disk, in which case the variability could be due to changes of the jet direction.

Overall, there are quite a number of observational indications for binary black holes and merger candidates. However, the best way to track down the SMBH merger will be by observing the gravitational wave emission. The planned space-based gravitational wave observatory LISA is expected to detect virtually all such supermassive black hole mergers throughout the visible Universe, and thus provide exquisite demographics of the cosmic merger history.

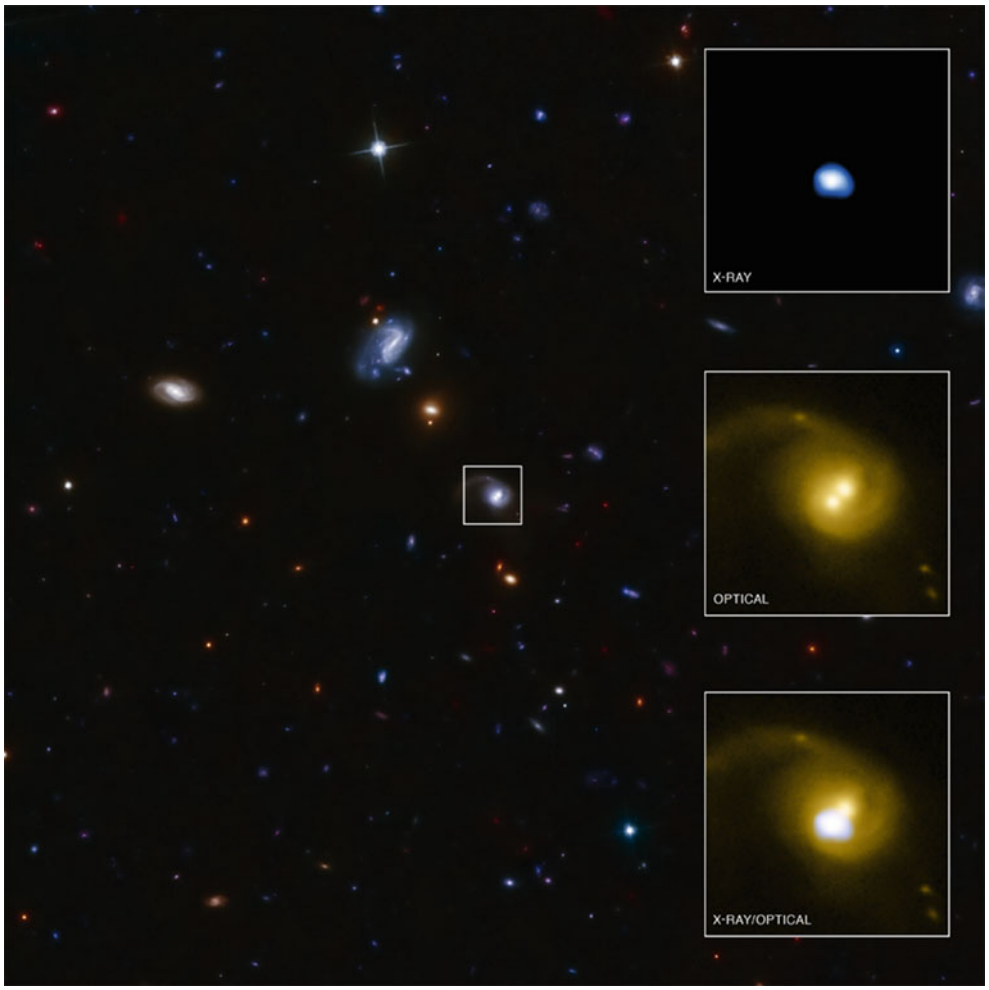


Fig. 10.20 The big multi-color optical image shows a $1' \times 1'$ part of the COSMOS field. The galaxy at the center is CID-42 (at $z = 0.36$), of which three zoomed images are shown on the *right*, with a side length of $3''.7$. The tidal tail seen in the optical suggests that this galaxy underwent a merger in the more recent past. The optical images displays two bright compact sources; only one of them has an X-ray counterpart. Credit: X-ray: NASA/CXC/SAO/F. Civano et al; Optical: NASA/STScI; Optical (wide field): CFHT, NASA/STScI

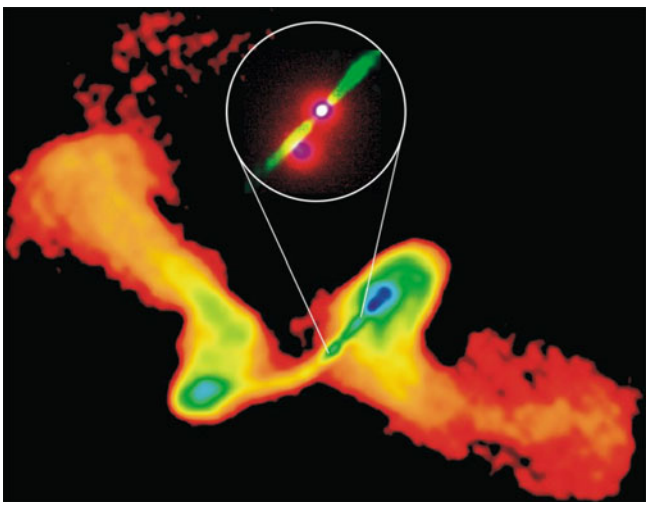


Fig. 10.21 The X-shaped radio galaxy 3C326, observed with the VLA. The *inset* shows an optical HST image of the innermost jets. The pair of radio jets at different directions may have its origin by a change of the black hole's spin direction, as may occur in the process of binary black hole mergers. Credit: Image courtesy of NRAO/AUI and Inset: STScI

10.5.3 Environmental effects on galaxy properties

Major merger events between galaxies change the morphology and physical properties of galaxies dramatically. We have seen that this is the probable road to the formation of normal elliptical galaxies. However, this is not the only process in which the properties of a galaxy can be altered; it is merely the strongest one. Merging preferentially occurs in groups, which combine an environment of a high number density of galaxies with a relatively small velocity dispersion needed for mergers.

Harassment. In clusters of galaxies, the characteristic collision speed between galaxies is considerably higher than their internal velocity dispersion; as we argued before, in such a case no merging can take place. However, a high-speed collision between galaxies affect their internal properties in a different way. If we consider such a collision in the rest frame of one of the galaxies, then its components experience a rapid

change of the gravitational potential as the other galaxies fly by. As a consequence, the matter in the galaxy increases its internal energy—it is impulsively heated. This causes the matter in the galaxy to expand, i.e., it is less gravitationally bound than before, and therefore is more easily affected by tidal gravitational forces. Furthermore, the heating of the stellar component changes the distribution function (i.e., the phase-space density f discussed in Sect. 2.3.1) of the stars—dynamically cold stellar disks are heated and may get destroyed, with the stars evolving into a spheroidal distribution. The cumulative effect of such high-speed collisions is often called galaxy harassment.

Ram-pressure stripping and strangulation. As a galaxy orbits in a cluster, it moves relative to the hot intracluster medium. In the rest frame of a galaxy, the ICM acts like a wind, with the wind speed equal to the orbital velocity of the galaxy. This wind causes a pressure force on the interstellar medium of the galaxy; it is proportional to the density of the ICM and the square of the velocity. If this force is stronger than the gravitational force of the galaxy which hosts the interstellar gas, this gas can be removed from the galaxy. This ram-pressure stripping can thus over time turn a gas-rich disk galaxy into a disk galaxy without gas, i.e., a spiral galaxy into an S0 galaxy. This effect may be the origin of the larger abundance of S0 galaxies in clusters than in the field population. It also provides a natural explanation for the Butcher–Oemler effect (see Sect. 6.8), which states that clusters of galaxies at higher redshift contain a larger fraction of blue galaxies. The blue (spiral) galaxies that have existed at higher redshift may have turned into S0 galaxies in the meantime. The fact that the fraction of ellipticals in a cluster remains rather constant as a function of redshift, whereas the abundance of S0 galaxies increases with decreasing z, indicates the importance of the latter process as an explanation of the Butcher–Oemler effect.

Gas which is removed from the galaxies is chemically enriched. The metallicity of the ICM is believed to be due to the mixing of this enriched gas with the intracluster medium. Hence, the metals of the ICM have been generated by earlier stellar populations in cluster galaxies.

The efficiency of this effect, as well as that of harassment, depends on the orbit of the galaxy. If the orbit comes close to the inner part of the cluster where the gas and galaxy number density are large, all the gas may be removed, whereas otherwise, only the outer, more loosely bound gas is lost. In this case, the galaxy retains its gas in the inner part and may continue to form stars for a while; only when this gas supply is exhausted, it then turns into a red galaxy, since no new gas can be gained from cooling or accretion. This effect is called strangulation.

Cannibalism. The orbit of a galaxy in a cluster is affected by dynamical friction (see Sect. 6.3.3); it loses energy and angular momentum, and so its orbit will shrink in time. The efficiency of this effect again depends on the galaxy orbit; the closer it comes to the inner parts of the cluster, the stronger are the gravitational friction forces. Furthermore, as seen from (6.30), it depends on the galaxy mass, with more massive galaxies being affected more strongly. Depending on the orbital parameters, a cluster galaxy can lose most of its angular momentum in a Hubble time, sink to the center, and there merge with the central galaxy. By this process, the central galaxy becomes more massive, as it 'cannibalized' other cluster members. The aforementioned mass dependence may lead to an increase of the mass and luminosity difference between the brightest cluster galaxy and the second-brightest one.

10.6 Evolution of the galaxy population: Numerical simulations

In the preceding sections, the formation of disk and elliptical galaxies were described; it is generally believed that the collapse of gas, together with its angular momentum, leads to the formation of disk galaxies, whereas mergers and interactions are the prime cause for the occurrence of early-type galaxies. Our understanding of these formation processes can now in principle be used to predict the evolution of the galaxy population in the cold dark matter universe. The cosmological model predicts the abundance of halos as a function of mass and redshift, the distribution of their spin parameter, as well as the frequency of major and minor mergers. One thus might expect that from these ingredients, the galaxy population can well be predicted.

However, there are some major difficulties which hamper easy progress in this direction. The evolution of galaxies (in contrast to their dark matter halos) is strongly governed by baryonic processes, many of which are not fully understood. For example, we have no quantitative understanding about star formation. The way how the explosion of a supernova feeds back energy into the interstellar medium is subject to considerable uncertainties; this is even more true for the feedback processes related to AGN activity in galaxies.

A further serious problem is related to the enormous dynamic range in length scales which are involved in galaxy evolution in the cosmological context. We have seen that galaxy evolution depends on the local environment; galaxies evolve differently in groups and clusters than in the field. Hence, one needs to consider a sufficiently large volume of the Universe such that it contains a representative population

of cluster-mass halos. As we argued in Sect. 7.5.3, the cosmological box should not be much smaller than $200h^{-1}$ Mpc on the side. On the other hand, the Galactic disk has a scale-height of $\sim$ 100 pc, and star formation occurs in molecular clouds with a typical size of $\sim$ 1 pc. Hence, an ab initio simulation of galaxy evolution would have to have a dynamic range of at least 10^8 in length—too ambitious to be carried out.

Nevertheless, enormous progress in our understanding of the galaxy population has been achieved in recent years. Essentially, two different methods are used to overcome the aforementioned problems: Historically the first was semi-analytic modelling of the evolution of galaxies; we will discuss these models in Sect. 10.7. But more recently, hydrodynamical cosmological simulations have been employed to study the formation and evolution of galaxies, which we describe in this section.

10.6.1 Numerical methods

The increase in computer power, as well as the evolution of efficient numerical codes have allowed cosmological simulations which include baryonic physics: heating and cooling of gas, hydrodynamical effects etc. Simulating these processes is much more difficult and time consuming than pure N-body simulations which solely contain gravity—correspondingly, either their box size and/or their spatial (and mass) resolution are smaller.

There are two widely spread methods for the numerical treatment of hydrodynamics. In the first case, a stationary grid is set up, and the differential equations of hydrodynamics (such as the continuity and Euler equations) are discretized on the grid.[8] In the second case, the fluid is represented by fluid particles, which are considered representative of a mass element of gas. The interaction between these fluid particles are prescribed such that the transport of mass, momentum, and energy follows the laws contained in the equations of hydrodynamics; this approach is termed *smooth particle hydrodynamics (SPH)*. Different variations of these two basic schemes have been developed. For example, in the grid-based approach, one wants to have a higher spatial resolution in regions of large gas density; for this purpose, one can generate sub-grids with a smaller mesh size which yield higher spatial resolution. Such a numerical scheme is called *adaptive mesh refinement* (AMR). Lately, a new method has been developed, which is also based on a grid; however, the grid is not stationary but moves with the fluid, and the density of grid points adapts to the fluid density. First tests indicate that this new scheme (called AREPO) overcomes some of the problems of the two former schemes and yields considerably more reliable results (see Fig. 10.22).

The necessity for sub-grid physics. In order to overcome the resolution issues, several small-scale physical effects can be treated only approximately. Since these simulations are several orders of magnitude away from being able to resolve the formation of molecular clouds, one needs a recipe for star formation. For example, if the gas density exceeds a threshold value in one region (or for one SPH particle), one assumes that a fraction of this gas is turned into stars. In the simulation one keeps track of this newly formed stellar population, i.e., its mass, formation time and metallicity.

Since massive stars very quickly after formation explode as core-collapse supernovae, for each massive star formed (given by the total newly formed stellar mass and the assumed initial mass function) an energy of $\sim 10^{51}$ erg is transferred back to the surrounding gas distribution. Also this feedback process occurs on scales below the numerical resolution, so it is assumed that this energy is used to heat the local gas. Also refinements have been successfully implemented, where each gas cell or particle is split into a hot, diffuse part and a cold and dense part. Gas can be exchanged between these two phases due to heating and cooling processes. It turns out that the outcome of simulations depend on the detailed prescription of this feedback mechanism. If it is assumed that the supernova energy is transferred mainly to the cold and dense gas, then it can be radiated away rather quickly without affecting the dynamics of the gas appreciably. On the other hand, if the feedback energy heats the diffuse gas, radiative cooling is much less efficient, the gas increases its pressure and expands, driving gas out of the dense region (or, in physical terms, out of the disk where star formation is located).

Similarly, the accretion of gas onto a central supermassive black hole and the corresponding feedback can be treated only approximately. The accretion disk (or more generally, the accretion region) can not be resolved, but the accretion rate, and the corresponding energy output, depends mainly on the rate at which gas can be driven into the central region of the galaxy. The accretion rate can then be estimated from the physical conditions on scales much larger than the accretion disk size, and is often approximated by the Bondy–Hoyle rate [see (5.16)], bounded above by the Eddington rate or a small multiple of it. The resulting luminosity of the supermassive black hole is assumed to be a fraction

[8] The equations of hydrodynamics describe the behavior—or transport—of the mass, momentum and energy in a fluid. Mass conservation is expressed by the continuity equation (7.2). The evolution of the fluid momentum is given, in the simplest case, by the Euler equation (7.3); however, since gas is dissipative, frictional terms need to be included (the resulting equation for the fluid velocity is then called Navier–Stokes equation). Finally, the transport of energy is described by an energy equation, which contains sources and sinks of energy, as they can be caused by absorption and emission of radiation and the local generation of heat by frictional forces.

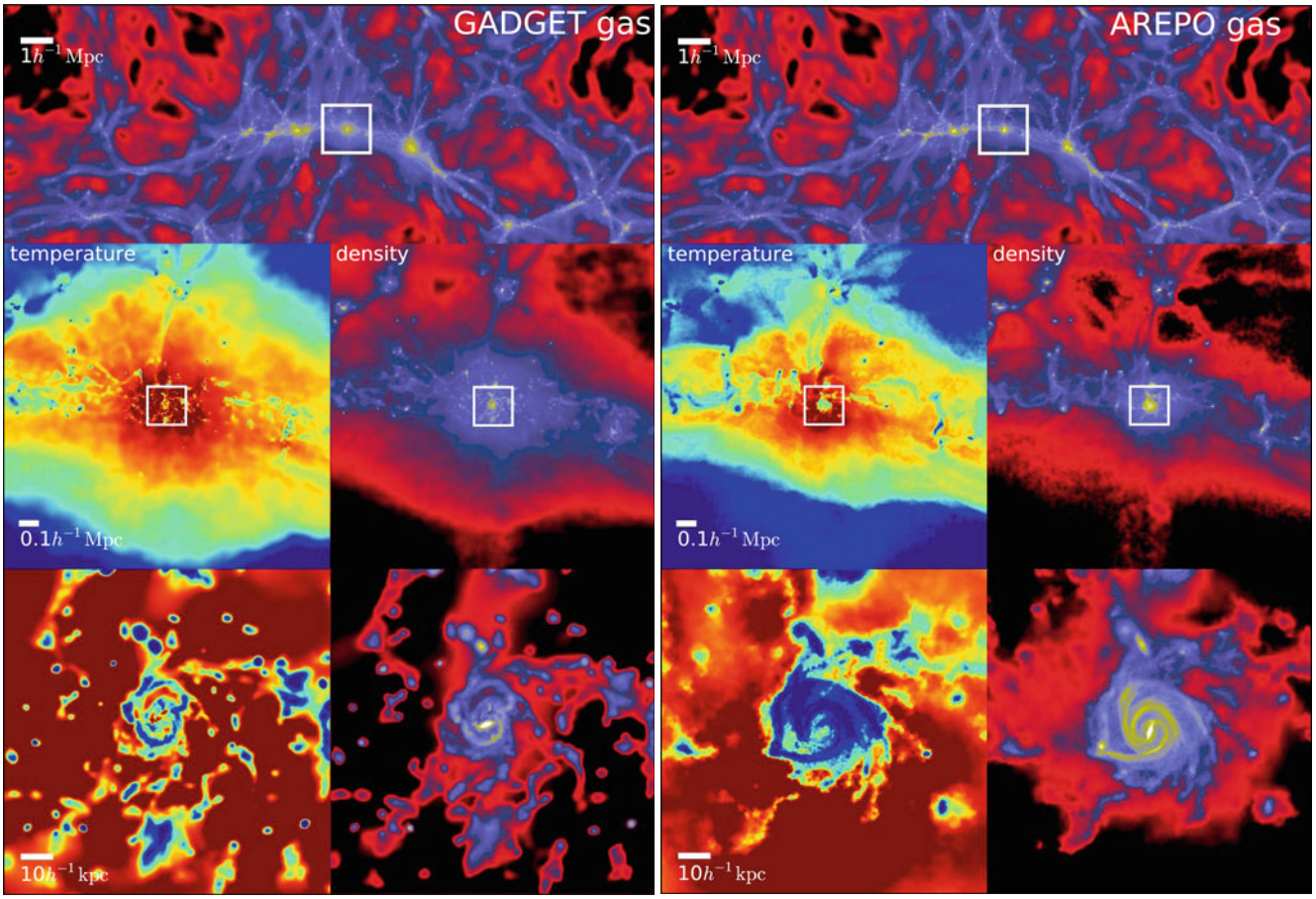


Fig. 10.22 The projected gas density from two hydrodynamical simulations of a galaxy. Both simulations use the same initial conditions, as well as the same prescriptions for star formation and feedback; the only difference is the numerical method with which the equations of hydrodynamics are treated. In the *left panel*, a smooth particle hydrodynamics (SPH) scheme was used, whereas in the *right panel* the new AREPO method was employed. The *upper panels* show the gas density at redshift $z = 2$ in a large fraction of the full numerical box, whereas the *smaller panels* show subsequent zooms (indicated by a *white square* in the previous step) of the gas temperature and density, centered on a disk galaxy. This galaxy has significantly different properties in both simulations. Whereas AREPO yields an extended disk with spiral arms and a bar, the corresponding galaxy is much smaller in the SPH simulation. Differences in the clumpiness of the medium are also visible. Source: M. Vogelsberger et al. 2012, *Moving mesh cosmology: numerical techniques and global statistics*, MNRAS 425, 3024, p. 3031, Fig. 1. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

ϵ of the rest-mass energy rate accreted, i.e., $L = \epsilon \dot{m} c^2$, with $\epsilon \sim 0.1$ (see Sect. 5.3.5). Some fraction of this energy output is assumed to be fed back into the surrounding gas. This gas, being suddenly heated, will greatly expand and drive a shock wave, by which it is blown out of the central region. This expanding blast wave can then be followed by the hydrodynamic solver.

The gravitational force is calculated in a very similar way as done for pure N-body simulations, except that the source of gravity is the sum of the densities of dark matter, gas, stars and the central black holes. Changes to the dark matter profile of halos due to the contraction of cooling gas is thus included in such simulations.

Comparison of numerical methods. We have reported some results of such simulations above, namely simulations of the merging of two disk galaxies (see Sect. 10.5.1). At present, the results from such simulations need to be analyzed with care; there are still considerable uncertainties regarding the small-scale physics (star formation and feedback), as well as the accuracy with which the hydrodynamical behavior of the gas can be followed. In the recent Aquila Comparison Project, a comparison of 13 different hydrodynamical simulations of one galaxy (where all simulations used the same initial conditions) was performed and significant differences were found. For example, the morphology of the galaxies shows strong variations between the different simulations. This can be traced back to the star-formation history: the earlier most of the stars are formed, the less pronounced is the disk today. Obviously, the amount of star formation in the early history of the galaxy depends on the amount of cooling gas and, in particular, the efficiency

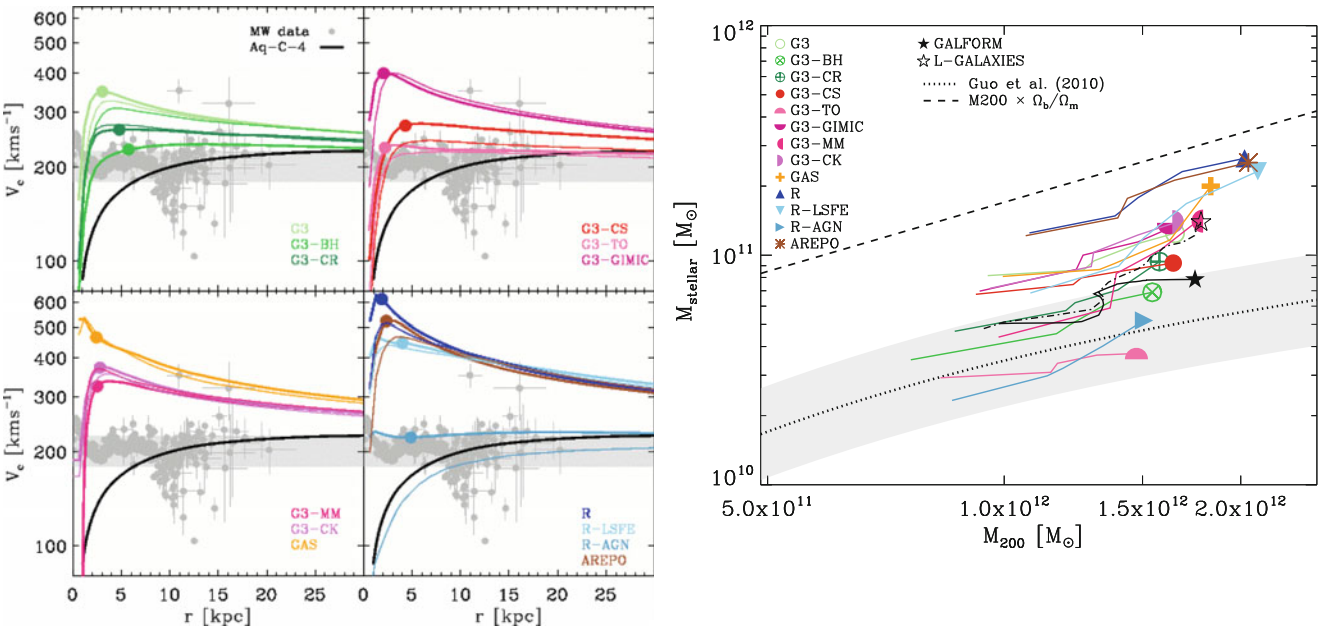


Fig. 10.23 Results from the Aquila Comparison Project, in which, starting from the same initial conditions, the evolution of a disk galaxy was followed with 13 different simulations. The halo mass of the galaxy is similar to the one of the Milky Way, $\sim 1.6 \times 10^{12} M_\odot$. *The left-hand panel* shows the rotation curves of the galaxy as obtained by the different simulations, with the rotation curve of the Milky Way shown in *light grey* for comparison. The *solid black curve* in each of the four subpanels is the rotation curve as obtained from a dark matter-only simulation of the same initial conditions. In most cases, the rotation curve has a peak at low radius, after which is strongly declines outwards—in contrast to observed rotation curves of spiral galaxies which are almost flat. The reason for this behavior is the too effective cooling of gas, yielding a far too concentrated baryonic distribution in the inner part of the galaxy. *The right panel* shows, for each of the 13 simulations, the total stellar mass as a function of the halo mass M_{200}, as the galaxy evolved from redshift $z = 2$ (*beginning of the curve*) to today (*symbol*). The predicted stellar mass varies by about a factor 10 between the simulations. The *dotted curve* shows the expected stellar-to-dark matter relation, as expected from matching the abundance of dark matter halos to that of the observed galaxy abundance [essentially by defining a function $M_*(M_{200})$ which brings the two curves in Fig. 10.2 into agreement], the *dashed curve* shows the maximally possible stellar mass, given by the halo mass times the mean cosmic ratio of baryons to total matter. The two *black curves* (and *stars*) show the model predictions of two semi-analytic models of the same halo. Source: C. Scannapieco et al. 2012, *The Aquila comparison project: the effects of feedback and numerical methods on simulations of galaxy formation*, MNRAS 423, 1726, p. 1734, 1735, Figs. 5, 6. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

of feedback processes. The tendency of turning too much gas into stars is often called the 'overcooling problem' in galaxy evolution. On the other hand, the feedback must not prevent the later accretion of gas, to build up the disk at lower redshift.

The left hand panel of Fig. 10.23 shows the predicted rotation curves of the galaxy, as obtained from these 13 simulations. Most of them exhibit a pronounced peak at small radii, after which they decline outwards. Again, this is due to the concentration of stars in the inner part of the galaxy, which not only acts as a source of gravity by itself, but the corresponding efficient cooling of baryons led to the contraction of the dark matter halo. Only simulations with a very strong feedback lead to approximately flat rotation curves; unfortunately, these models usually do not have a well-developed disk. The compactness of the baryonic distribution yields corresponding circular velocities which are well above the observed Tully–Fisher relation for spirals. The stellar mass as predicted by the models in shown in the right-hand panel of Fig. 10.23; also here the variations between the simulations are large, again mainly due to the different feedback prescriptions. In some simulations, nearly all available baryons inside the halo were turned into stars, whereas other simulations have $\sim$ 10 times lower stellar mass.

Lessons. We present this comparison here for a number of reasons. First, this exercise shows which of the various differences between codes matter most for the predictions and thus gives insight on how the assumptions must be modified in order to obtain results closer to the observed properties of

galaxies. That may seem like cheating at first sight: one turns the knobs in such a way that the results are in agreement with observations—what about the predictive power of such simulations then? However, one must keep in mind that some of the key processes (to repeat: star formation and feedback) cannot be followed from first principles, but are included in the form of recipes. In a sense, we parametrize our ignorance, and try to calibrate the set of parameters with a small number of key observational facts (such as the normalization of the Tully–Fisher relation, or the luminosity function of galaxies). The number of different predictions from such simulations is much larger than the number of parameters chosen; thus, once appropriate prescriptions for the 'sub-grid' physics are found, these models have predictive power.

A second, very important reason for this discussion here is to caution the reader about the reliability of some predictions of our cosmological model. We first note that a similar comparison was carried out for N-body simulations, with a very satisfactory overall agreement on scales larger than the resolution limit (of course, on scales below the numerical resolution, the results are even expected to be different). Thus, the predictions concerning dark matter-only are very robust. However, the inclusion of baryons, and the complex physical processes they are subject to, render predictions much less reliable. The smaller the scales and the denser the baryons, the more non-linear are the physical processes, and the harder it is to reliably trace them. In particular, processes on small scales have a strong effect on large scale—e.g., feedback.

This must be taken into account when arguments are made concerning the incompatibility of some observational results with ΛCDM. In most cases, these arguments concern the smallest scales or the least massive objects, for example properties of dwarf galaxies. From what was just stated, it is clear that currently we are not able to make detailed predictions about observational properties of small galaxies – when we cannot even predict the stellar mass of a massive galaxy halo to within a factor of 3! The fact that current simulations fail to reproduce spirals which fit the Tully–Fisher relation is most likely not a failure of the underlying cosmological model, but a lack of understanding of the complex small-scale physical processes involved.[9]

[9]The situation is rather similar in meteorology, where we believe to know all the essential physical processes that affect the Earth atmosphere; nevertheless, we all know that weather predictions can be terribly wrong, even on short time-scales. The reason is that, although the relevant physical laws are known, their consequences cannot be calculated with sufficient accuracy due to the complexity of the underlying equations. Also in this case, small-scale, highly non-linear processes (convection, turbulence) have an impact on the large-scale properties of the atmosphere.

10.6.2 Results

The challenge for models of galaxy evolution is to explain the observational results of the galaxy population at low and high redshift. In this section, we will illustrate the current status of gas-dynamical cosmological simulations and their ability to reproduce key observations.

Growth of black holes and galaxies. The tight correlation between central black hole mass and properties of the spheroidal stellar component in galaxies suggest a close connection of the evolution of both components. Furthermore, feedback processes from AGN activity are essential for understanding the evolution of galaxies. We next present some results of a gas-dynamical cosmological simulation which includes the evolution of the supermassive black holes in galactic centers. For this simulation, the Bondi–Hoyle accretion rate was assumed, as described above. All halos, once they exceed a mass threshold, were artificially provided with a seed black hole of $10^5 M_\odot$ at the location of the densest gas particle, thereby circumventing our lack of understanding on how the first massive black holes were formed. The mass of the seed black hole is rather unimportant as long as it is much smaller than the mass at later times. In particular, the total mass in these seed black holes is a minute fraction of the total mass of black holes at later epochs which is totally dominated by accretion processes.

Figure 10.24 shows the gas density and temperature in the simulation box, together with the location of black holes. As expected, these are located in the center of density maxima. The black hole distribution traces the overall density distribution, although with considerable scatter. The number density of black holes varies strongly between filaments of gas which apparently have very similar density. Already at redshift $z = 6.5$, quite a number of black holes have formed, some with masses close to $10^7 M_\odot$, although no SMBH has formed with masses needed to explain the luminous quasars seen at $z \gtrsim 6$ (i.e., $M_\bullet \gtrsim 10^8 M_\odot$). However, these quasars have a very low space density, and one cannot expect to find such massive black holes in a simulation box of the size considered here.

The total mass density of the SMBHs in the simulation as a function of redshift is shown in the upper panel of Fig. 10.25 where it is compared to the mean mass density of stars. We see that the SMBH density increases faster with cosmic time than the stellar mass density, which shows that the evolution of the stellar density precedes that of the SMBH. This is shown more clearly in the lower panel, where the growth rate of these densities are displayed (i.e., the time derivative of the curves in the upper left panel). Both growth rates exhibit a peak at intermediate redshifts; however, whereas the peak in the stellar mass density is fairly broad (as observed in the Madau diagram—see Fig. 9.55),

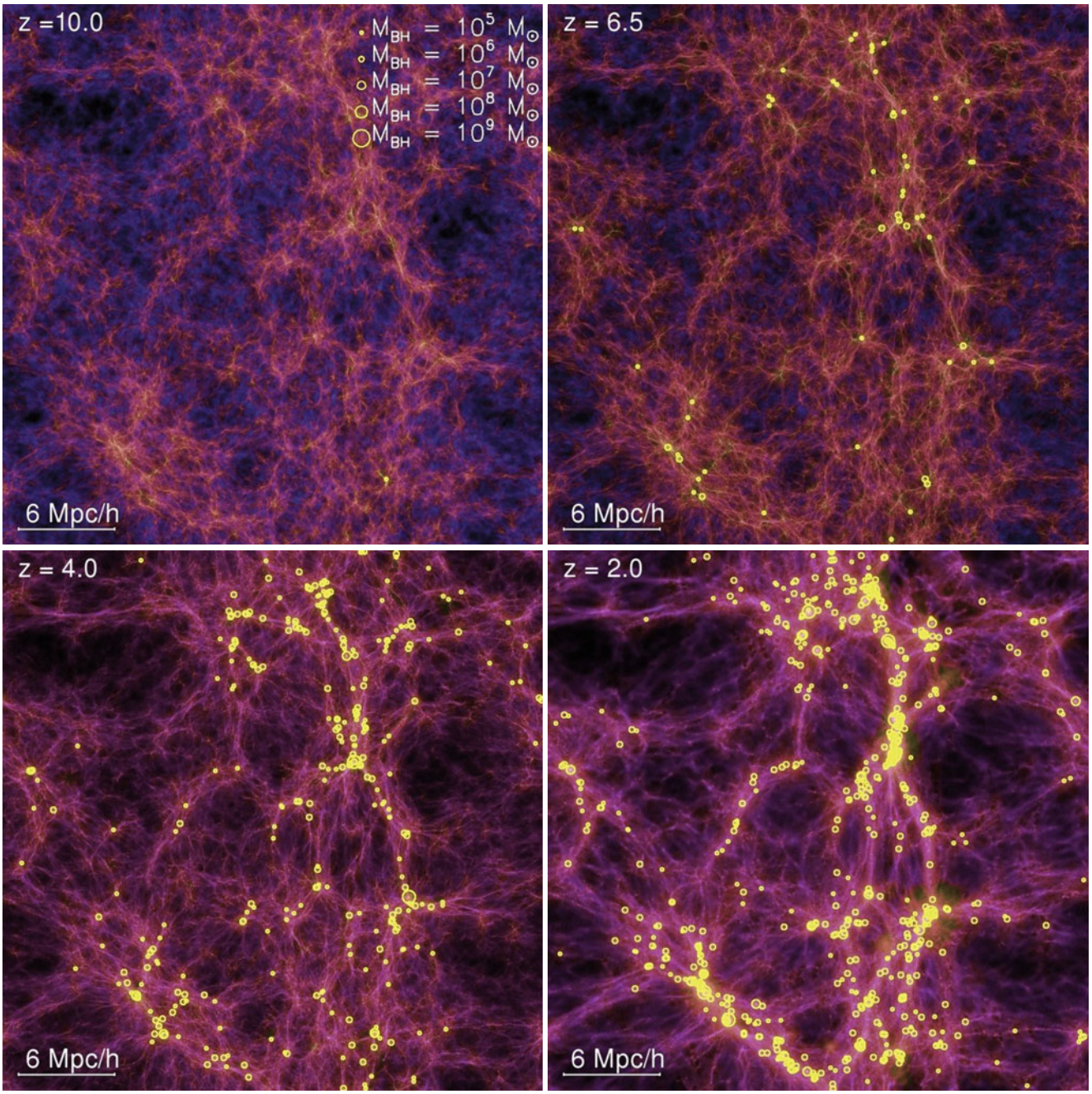


Fig. 10.24 The density of baryons from a hydrodynamical simulation, projected over a $5h^{-1}$ Mpc thick slice. *Intensity* and *color* indicate gas surface density and temperature, respectively. Four snapshots at different redshifts are shown. *Yellow circles* indicate the location of supermassive black holes, with the *symbol size* indicating the black hole mass. In this simulation, for which a box of size $L = 33.75h^{-1}$ Mpc was chosen, it was assumed that 5 % of the AGN luminosity is fed back to the interstellar medium; this choice was made in order to reproduce the observed relation between black hole mass and velocity dispersion of the spheroidal stellar component. Source: T. di Matteo et al. 2008, *Direct Cosmological Simulations of the Growth of Black Holes and Galaxies*, ApJ 676, 33, p. 38, Fig. 1. ©AAS. Reproduced with permission

it is much more peaked for the black holes. Comparing the mass density of the SMBH population with observational estimates, one finds broad agreement.

It is instructive to consider the mass history of individual black holes in the simulation, which is displayed in the upper panel of Fig. 10.26 for the six most massive (at $z = 1$) black holes and two less massive ones. The growth of the SMBH mass is quite rapid at the beginning, and apparently episodic. The mass accretion rate in units of the Eddington rate (5.27) for four of the SMBHs is plotted in the lower

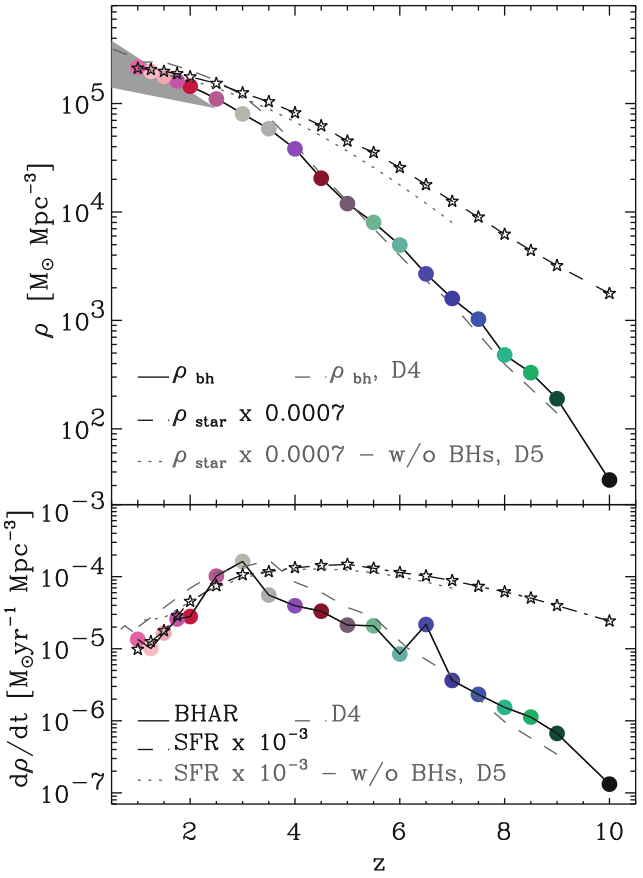


Fig. 10.25 The *upper panel* shows the mass density of black holes as a function of redshift (*colored points* and *solid curve*), for the simulation shown in Fig. 10.24. The *lower dashed curve* shows the same quantity for a simulation with the same initial conditions and physical assumptions, but lower mass resolution, indicating that this prediction of the model is not strongly affected by resolution effects. The *dashed curve* and the *star symbols* show the mean density of stars, scaled by a factor 7×10^{-4}, whereas the *dotted curve* shows the stellar mass density from the same simulation, but where the feedback from the accreting black holes was absent. The *shaded grey triangle* at low redshifts shows estimates of the black hole mass density from observations. The *lower panel* displays the growth rate of the black hole mass density and stellar mass density, with the same line styles as in the upper panel. Source: T. di Matteo et al. 2008, *Direct Cosmological Simulations of the Growth of Black Holes and Galaxies*, ApJ 676, 33, p. 41, Fig. 4. ©AAS. Reproduced with permission

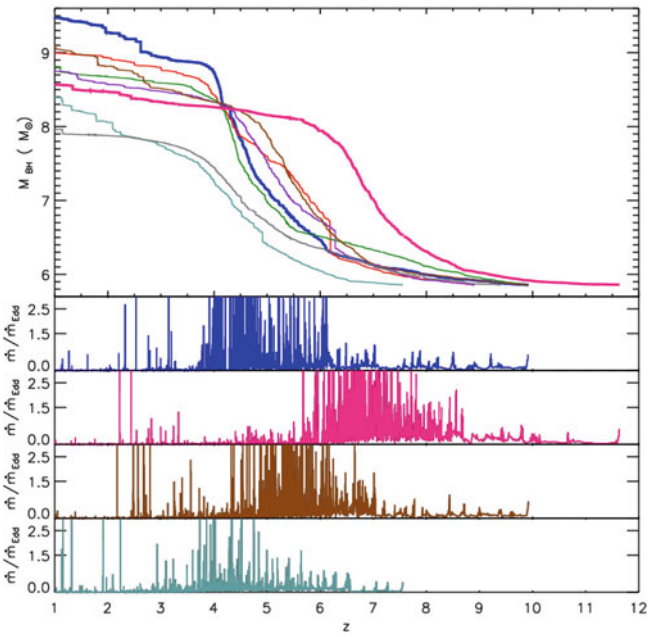


Fig. 10.26 The *upper panel* shows the mass of the six most massive black holes at $z = 1$, as well as that of two intermediate mass black holes, as a function of redshift. The two *thicker lines* highlight the most massive SMBHs in this simulation at $z = 6$ and at $z = 1$. The *lower panels* display the accretion rate in units of the Eddington rate as a function of redshift, with the same color coding as in the upper panel. Source: T. di Matteo et al. 2008, *Direct Cosmological Simulations of the Growth of Black Holes and Galaxies*, ApJ 676, 33, p. 48, Fig. 13. ©AAS. Reproduced with permission

panels (note that for this simulation, the maximum accretion rate was chosen to be three times the Eddington rate). The most massive black holes undergo extended periods where the accretion rate is very high, limited only by the Eddington ratio. Hence, these holes grow as fast as possible in these periods, since there is enough supply of fuel—presumably in the aftermath of a major merger. The most massive SMBH at $z = 6$ (pink curves) undergoes a very extended period of accretion between redshifts 5 and 7, after which it turns to become very inactive, with the exception of a few short episodes of accretion during which its mass is only slightly increased. The blue curve shows the most massive SMBH at $z = 1$, which started massive accretion only at redshift $z \lesssim 6$, but then rapidly grew in mass. Hence we infer from the figure that the fates of individual SMBHs, and their corresponding AGN activity, are quite diverse.

One of the most promising results of the simulation is the strong correlation between black hole mass and the velocity dispersion of the stellar component, as shown for six different redshifts in Fig. 10.27. There we see that beginning with $z \sim 4$, the best-fit relation from the simulation agrees with the locally observed one (see Fig. 3.45). The normalization of the power-law fits depends on the assumed fraction of AGN luminosity that is available for feedback, chosen to be 5 % here. However, more exciting than the precise normalization of this relation is the fact that hierarchical galaxy evolution is able to explain the observed tight correlation without additional ad-hoc assumptions. Also seen is that the tight relation is satisfied by black holes independent of their accretion state—i.e., active and inactive SMBH lie on the same relation.

However, it must be pointed out that the resolution of these simulations do not allow statements about the morphology of galaxies; therefore, the velocity dispersion plotted in Fig. 10.27 is that of the total stellar population, not that of

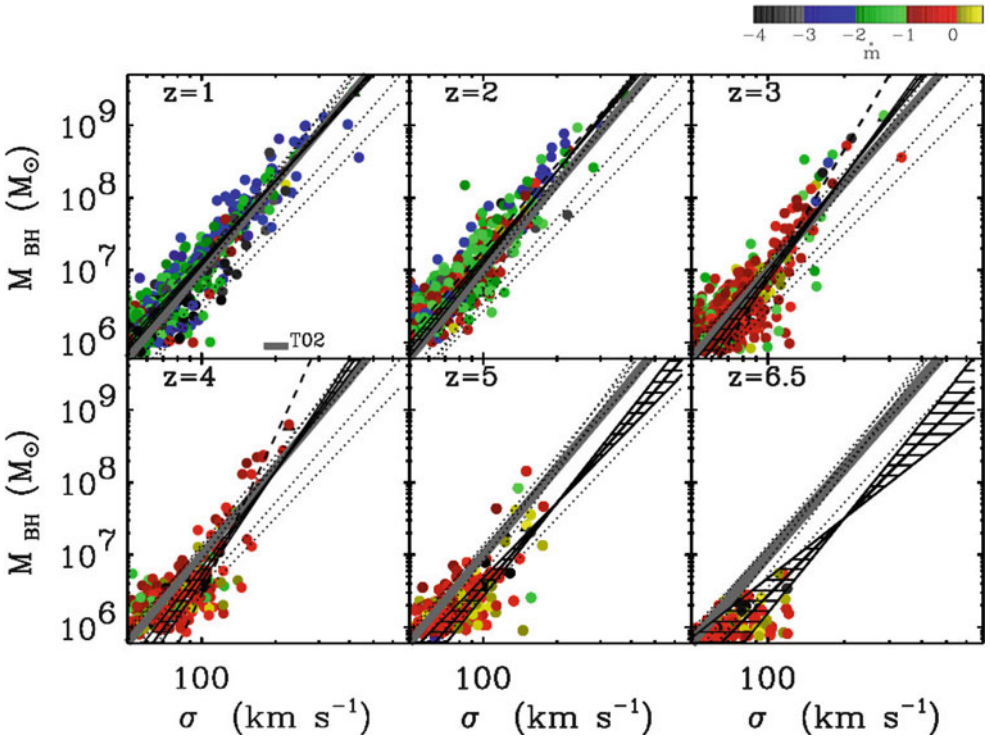


Fig. 10.27 From the same simulations as in Fig. 10.24, the black hole mass is plotted as a function of the velocity dispersion σ of the stellar mass particles within the half-mass radius, for six different redshifts as indicated. The *thick grey line* in all panels is the best fit to the local $M_\bullet$-σ relation. For each redshift, a power-law relation was fitted to the points, which is shown as *solid line*, together with its 1-σ uncertainty (as hatched region; for the low redshifts, this uncertainty is so small that the hatched region is essentially invisible). The *dotted lines* in each panel show the power-law fits obtained at the other redshifts, increasing from top to bottom. The *color* of each point codes the accretion rate of the SMBH at the snapshot, in units of the Eddington accretion rate, with the corresponding color bar at the top of the figure. Source: T. di Matteo et al. 2008, *Direct Cosmological Simulations of the Growth of Black Holes and Galaxies*, ApJ 676, 33, p. 44, Fig. 8. ©AAS. Reproduced with permission

the spheroidal component only, for which the $M_\bullet$-σ relation is observed.

Impact of feedback on the gas. The foregoing discussion has shown the challenges of hydrodynamic cosmological simulations, in particular concerning numerical resolution and the implementation of sub-grid physics. We present next some recent results from simulations carried out with the AREPO code (see Fig. 10.22). Several runs were produced in which the parameters of the description for sub-grid physics were varied. The properties of the feedback by supernovae, which result in an outflow ('wind'), were varied, both concerning the mass rate of the outflow as well as its velocity. Furthermore, several feedback descriptions of AGNs were employed.

Figure 10.28 illustrates the importance of the feedback on the properties of the gas. As seen in the left-hand panels, the distribution of the gas density is more extended when feedback processes are included, as the outflows generated by supernovae and AGN feedback distributes the gas over a larger volume, whereas in a model with no feedback, the high-density gas is more confined to dark matter halos and the denser regions of the dark matter filaments. The impact of feedback is more dramatic on the distribution of gas temperature, as seen in the middle panels; without feedback, hot gas is confined to the densest regions, whereas the action of strong radio-mode AGN activity distributes hot gas over large regions of space. The feedback-driven outflows also lead to a wide-spread enrichment of the intergalactic gas with metals (right-hand panels), which otherwise would stay close to their source of origin, i.e., the inner regions of halos in which stellar evolution takes place, in sharp contrast to observations of QSO absorption which show that the IGM is metal enriched.

The star-formation rate density. Every successful model of galaxy evolution must be able to reproduce the observed star-formation history in the Universe. We have seen in Sect. 9.6.2 that the star-formation rate density evolves strongly with redshift, showing a broad peak at redshifts between 2 and 4. The top left panel of Fig. 10.29 shows a recent version of the Madau-diagram, and predictions from the numerical simulations. Here and in the other panels of the figure, the blue curve corresponds to the fiducial set of

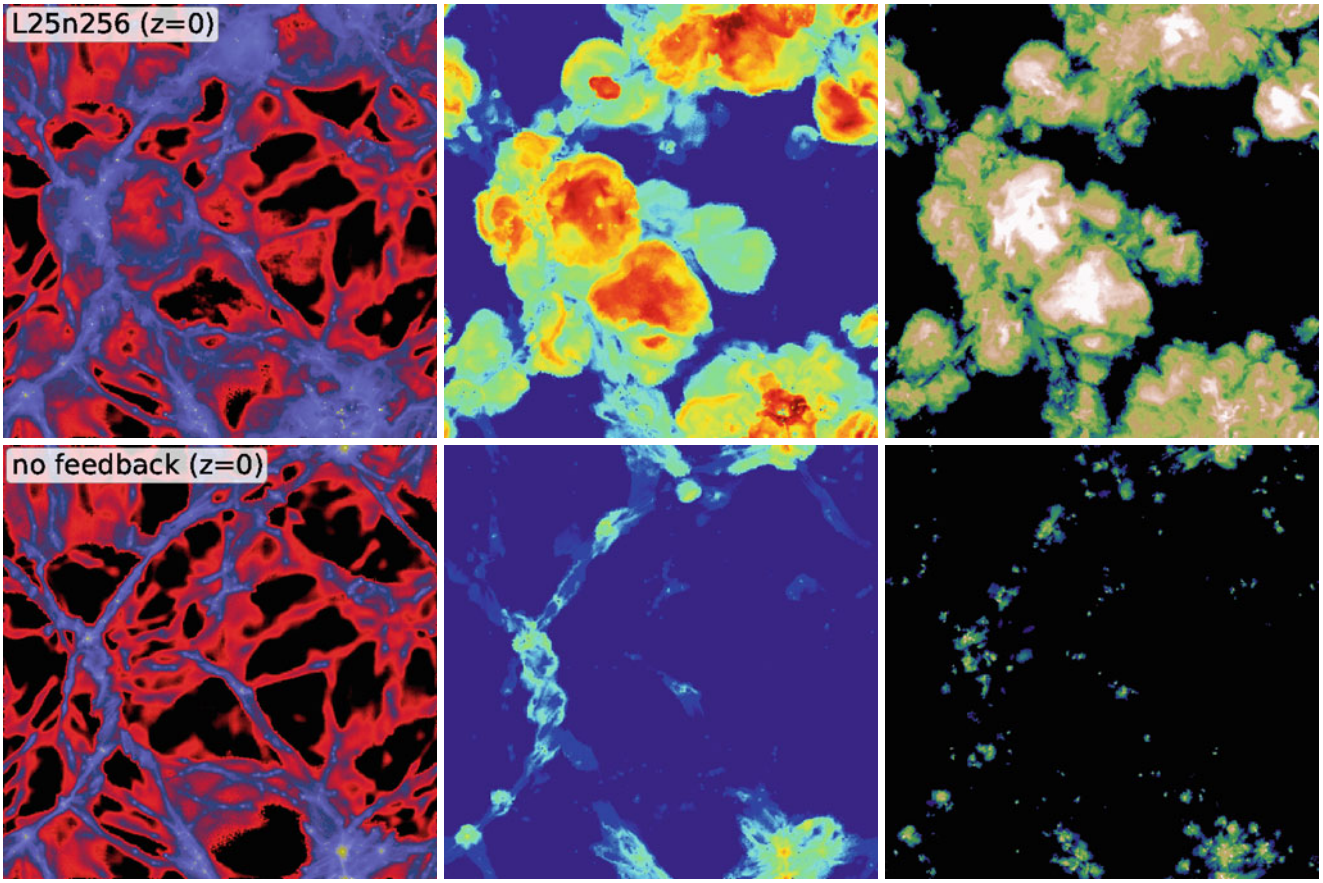


Fig. 10.28 Gas-dynamical simulations of structure formation, with (*top*) and without (*bottom*) the inclusion of feedback processes. Shown is the distribution of gas density (*left*), temperature (*middle*) and metallicity (*right*) at $z = 0$, over an area $25h^{-1}$Mpc on the side and projected over a slab of thickness $1h^{-1}$Mpc. Source: M. Vogelsberger et al. 2013, *A model for cosmological simulations of galaxy formation physics*, arXiv:1305.2913, Fig. 3. Reproduced by permission of the author

feedback parameters, and the other colored curves show variations of the model. Whereas the fiducial model provides a satisfactory fit to the observational results, some of the other models fail dramatically. Foremost, the model with no feedback overproduces stars by a large factor, and can safely be ruled out also for this reason. Changing the feedback from SNe can also alter the model prediction substantially; for example, the model termed 'fast winds' blows the gas out of halos and thus prevents the formation of stars at later epochs. Strong winds remove the gas from halos at early times, thus reducing the star-formation rate, but later gas is reaccreted and results in star formation rates at low redshifts which are larger than those estimated from observations. Other parameter variations are seen to have a smaller impact on the predictions.

The stellar mass-halo mass relation. We saw in Sect. 7.7.4 that the ratio of M_*/M_{200} varies substantially with M_{200}, which is the origin for the mismatch between the halo mass function and the stellar mass function, shown in Fig. 10.2. In particular, this ratio attains a maximum at a characteristic mass scale which corresponds to a massive galaxy in the current Universe. The top right panel of Fig. 10.29 shows the predictions of the $M_*(M_{200})$-relation from the simulations, compared to the observed relation (shown as black curves). The fiducial model appears to reproduce the observed relation quite well, though the turnover at $M_{200} \sim 10^{12.2}\,M_\odot$ is less pronounced than that obtained from observations. The 'fast wind' model fails in a similar way as for the star-formation rate density—too much gas is blown out of halos. In general, variations of AGN feedback affect the upper mass end of the relation more strongly than for lower masses, and is essential for the suppression of star formation in high-mass halos, as argued several times before. Conversely, the low-mass end of the relation is more sensitive to feedback from supernovae.

The stellar mass function of galaxies. Successful galaxy evolution models should be able to reproduce the observed luminosity function of galaxies, as a function of redshift.

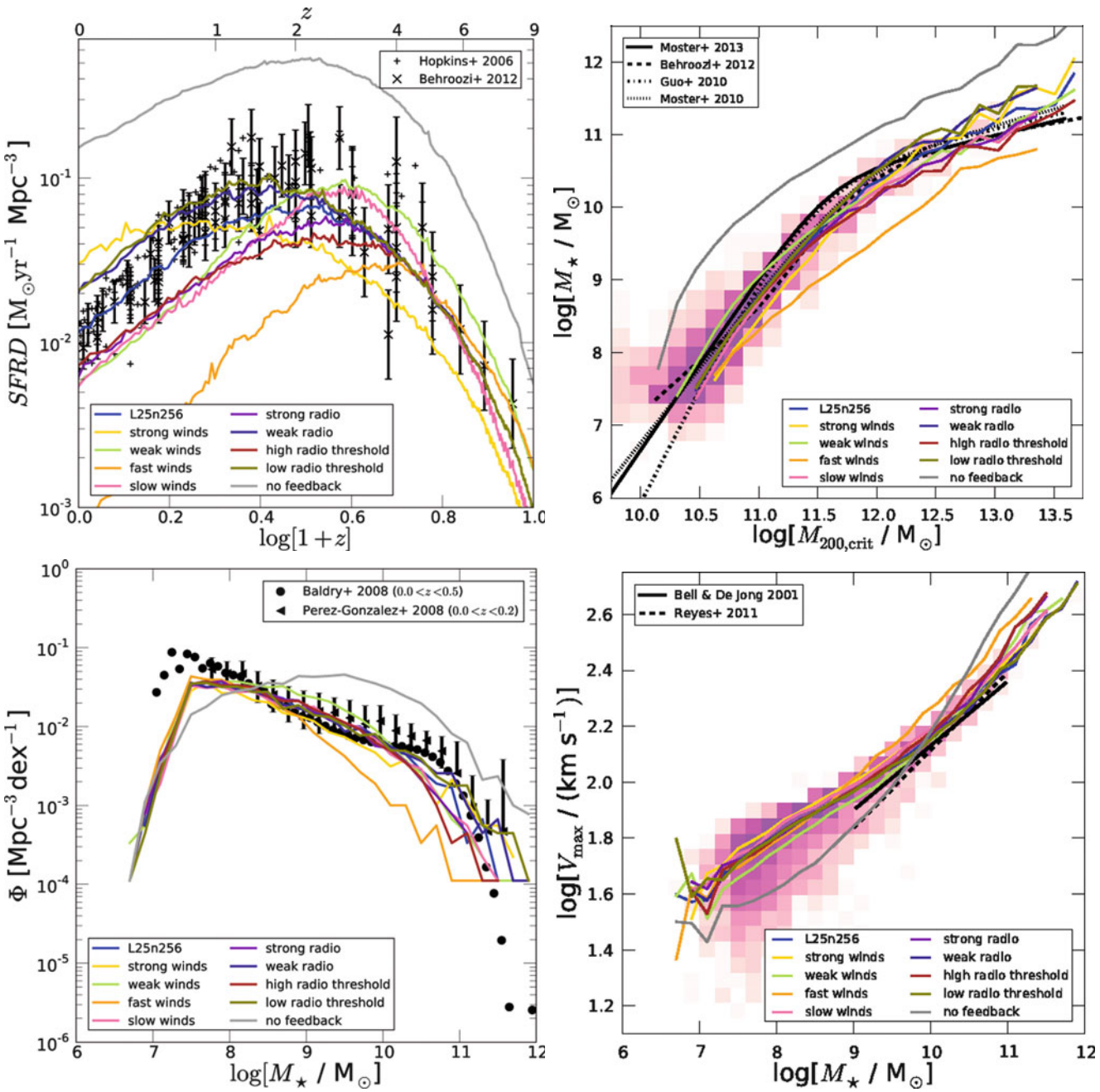


Fig. 10.29 Several results from the hydrodynamical simulations shown in Fig. 10.28 are displayed here. In all cases, the *blue curve* shows the simulation where all the free parameters of the model were set to their fiducial values. The other *curves* show variations of this model, which differ from the fiducial model by changing the prescription of various feedback processes by supernovae (these relate to the strong/weak/fast/slow wind models) and AGN. The *grey curve* corresponds to a model with no feedback. *Top left:* The star formation rate density as a function of redshift (i.e., the 'Madau plot'). *Black symbols* with error bars show estimates from observations, as described in Sect. 9.6.2, whereas the *curves* show the results from the simulation. *Top right:* The stellar mass vs. halo mass. The *shading* indicates the probability density for the fiducial model, with the *blue curve* showing the median of M_* at fixed halo mass. The *black curves* show estimates of the $M_*(M_{200})$-relation as obtained from abundance matching of galaxies with dark matter halos, the other *curves* correspond to variants of the numerical model. *Bottom left:* The stellar mass function at $z = 0$, compared to observational results (symbols with error bars). *Bottom right:* The maximum rotational velocity $V_{\rm max}$ of galaxies as a function of their stellar mass, i.e., the Tully–Fisher relation, for $z = 0$. The *shading* shows the probability density for the fiducial model, with the *blue curve* showing the median of $V_{\rm max}$ at fixed M_*. The two *black lines* show the observed Tully–Fisher relation, the other curves variations of the fiducial model. Source: M. Vogelsberger et al. 2013, *A model for cosmological simulations of galaxy formation physics*, arXiv:1305.2913, Figs. 6, 7, 8, 10. Reproduced by permission of the author

Since the prediction of the luminosity in a specific spectral band depends not only on the properties of the stellar population, but also on the dust content and distribution, the calculated luminosity function is affected by an additional uncertainty. For that reason, a comparison of the stellar mass function between simulations and observations is slightly more straightforward. This is shown in the bottom left panel of Fig. 10.29, where the stellar mass function from the simulations is compared to observational results at low redshifts. The cut-off below $M_* \sim 10^{7.5}$ is due to the finite resolution of the simulations which implies a minimum halo mass that can be resolved. Models with fast or strong winds from supernovae severely underpredict the mass function over a broad mass range. The impact of AGN feedback is most clearly seen at and beyond the mass scale where the mass function starts to bend over; in particular, reducing the strength of AGN feedback overpredicts the stellar mass function at the high-M_* end.

The Tully–Fisher relation. Finally, the lower right panel of Fig. 10.29 compares the observed Tully-Fisher relation with the model prediction. The fiducial model reproduces the observed relation fairly well, but the changes that occur by altering the feedback model parameters are modest in this case. However, the model without AGN feedback fails also this comparison, yielding a much steeper relation than observed.

Conclusion. The example just presented shows that modern hydrodynamic simulations of galaxy evolution can reproduce some key observables. By comparing the predictions from the model to observations, the various free parameters describing the sub-grid physics can be adjusted. Whereas the 'fiducial model' fares quite well in the comparison shown, there remain several shortcomings. For example, the observed mass-metallicity relation (see Fig. 3.40) is not well matched by the simulation, whereas the stellar mass-black hole mass relation can be reproduced fairly well. Without doubt, this field will see further strong developments in the future.

10.7 Evolution of the galaxy population: Semi-analytic models

Hydrodynamic simulations are difficult and computationally expensive. This means that one cannot carry out large numbers of such simulations, for example, to test a large number of different parameter sets for the sub-grid physics (like feedback efficiency). Furthermore, their spatial resolution and/or the total volume covered by these simulations are typically inferior to those of pure N-body simulations. Hence, it is a larger challenge to include both the large-scale density perturbations in the matter field (on scales larger than $\sim L/2$), and the high resolution necessary to resolve the smaller-mass galaxies.

Instead, one can follow a different approach, in which the behavior of the dark matter distribution is obtained from N-body simulations, and simplified descriptions of the behavior of baryons in this matter distribution are employed. The formation of galaxies happens in dark matter halos, and so each dark matter halo is a potential site for the formation of stars—i.e., a galaxy. At the moment a halo forms, one expects that it contains a baryon fraction equal to the cosmic mean, and that the baryons have approximately the same spatial distribution and the same specific angular momentum as the dark matter (which is obtained from the N-body simulation). The fate of the baryons then depends on various physical processes which we have already discussed above: cooling, star formation, supernova feedback, accretion of gas onto a central black hole, etc. Furthermore, the N-body simulation yield the merging history of all dark matter halos, and so the processes which occur in minor and major mergers can be treated as well.

Some of these processes are rather well understood, such as cooling, whereas for those physical processes which we are unable to describe with a quantitative physical model, a parametrized, approximate description is chosen. To give one example, the star-formation rate in a galactic disk is expected (and observed) to depend on the local surface mass density Σ_g of gas in the disk. Therefore, the star-formation rate is parametrized in the form $\dot{\Sigma}_{\rm SFR} = A\,\Sigma_g^{\beta}$ [see (3.16)], and the parameters A and β adjusted by comparison of the model predictions with observations. Such *semi-analytic models* of galaxy formation and evolution have contributed substantially to our understanding and interpretation of observations. We will discuss some of the properties and predictions of these models in the following.

10.7.1 Method for semi-analytic modeling

Merger trees. The distribution of particles resulting from an N-body simulation at a given output time can be used to identify dark matter halos. Several methods for that can be applied as described in Sect. 7.5.3, e.g., the friends-of-friends method, the spherical overdensity criterion, or a combination of these. Similarly, sub-halos within each halo can be identified as well. Comparing the lists of (sub-)halos and their particle contents at consecutive output times, one can identify whether a halo present at the earlier time has

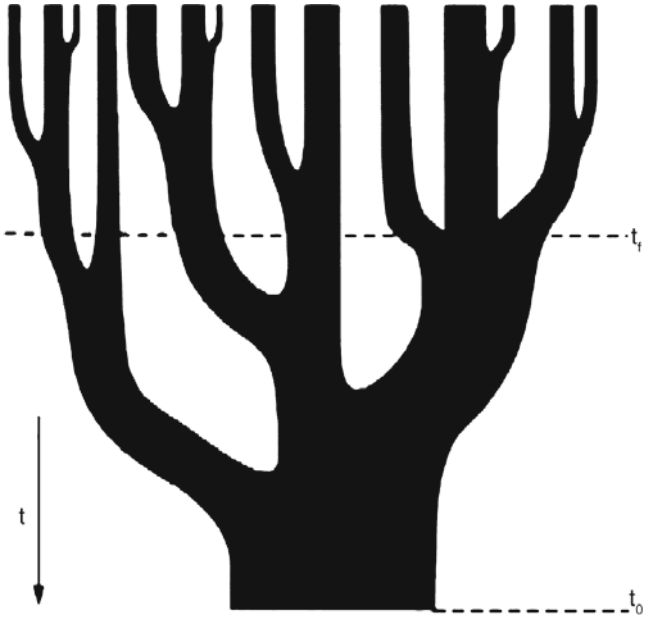


Fig. 10.30 A typical merger tree, as expected in a hierarchical CDM model of structure formation. The time axis runs from top to bottom. A massive halo at the present time t_0 has formed by mergers of numerous halos of lower mass, as indicated in the figure. One defines the time of halo formation as the time t_f at which one of the sub-halos had reached half the mass of the current halo. Source: C. Lacey & S. Cole 1993, *Merger rates in hierarchical models of galaxy formation*, MNRAS 262, 627, p. 636, Fig. 6. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

merged with another halo before the later output time. Over the course of time, more and more smaller halos have merged to more massive ones. Thus, for each halo at redshift $z = 0$, one can follow its complete merging history back in time, thereby obtaining its 'merger tree' (see Fig. 10.30).[10]

Gas cooling and star formation. In a halo where no merger process occurs at a given time, gas can cool. The cooling rate is determined by the chemical composition and the density of the gas as we described above. Besides the cooling processes, one can also account for the heating of the gas by the ionizing background radiation. Furthermore, one can account for the fact that low-mass halos are expected to have a smaller baryon fraction than the cosmic mean, if the gas is heated by the ionizing background to temperatures higher than the virial temperature of the halo, as described by (10.10).

[10]In fact, one can obtain a statistical ensemble of such merger trees also analytically from an extension of the Press–Schechter theory (see Sect. 7.5.2), but referring to N-body simulations also yields a prescription of the spatial distribution of the resulting galaxy distribution.

Cool gas is assumed to settle down in a rotationally-supported thin disk. If the density of the gas is sufficiently high, it can form stars, where the star-formation rate is assumed to follow the Schmidt–Kennicutt law (3.16), averaged over the disk. A simpler prescription for star formation is $\dot{M}_* \propto M_{\rm cool}/\tau$, where $M_{\rm cool}$ is the mass of cold gas in the halo, and τ a characteristic time, such as the dynamical time-scale of the disk. The newly formed stars are associated with a 'disk component'.

Supernova feedback. Shortly after the formation of stars, the more massive of them will explode in the form of supernovae. This will re-heat the gas, since the radiation from the SN explosions and, in particular, the kinetic energy of the expanding shell, transfers energy to the gas. By this heating process, some of the cool gas can be heated again and be driven out into the halo, i.e., the hot gas mass of the halo is increased in this way. Furthermore, if the energy input by supernova feedback is large enough, the heated gas can actually be expelled from the halo altogether (and at some later time reaccreted onto the halo). The suppression of the formation of low-mass galaxies by the effects mentioned here is a possible explanation for the apparent problem of CDM substructure in halos of galaxies discussed in Sect. 7.8. In this model, CDM sub-halos would be present, but they would be unable to have experienced an efficient star-formation history—hence, they would be dark.

In any case, feedback reduces the amount of cold gas available for star formation. This leads to a self-regulation of star formation, which prevents all the gas in a halo from being transformed into stars. This kind of self-regulation by the feedback from supernovae (and, to some extent, also by the winds from the most massive stars) is also the reason why the star formation in our Milky Way is moderate, i.e., not all the gas in the disk is involved in the formation of stars.

The left panel of Fig. 10.31 shows the importance of supernova feedback. Plotted here is the stellar mass fraction of baryons as a function of halo mass. A semi-analytic model without the inclusion of feedback yields the result that for halo masses below $\sim 10^{12} M_\odot$, more than half of the baryons are contained in stars. This is in sharp contradiction to observations which show that star formation is a rather inefficient process. This is just one of several arguments—in the absence of feedback, a Milky Way-like galaxy would have consumed all its gas early in its history, leaving no gas reservoir for current star formation. Including supernova feedback, the stellar mass fraction of the model can be made to agree with observations, for galaxy-mass halos. For more massive halos, feedback by supernovae is no longer efficient, and a different feedback mechanism is required (see below).

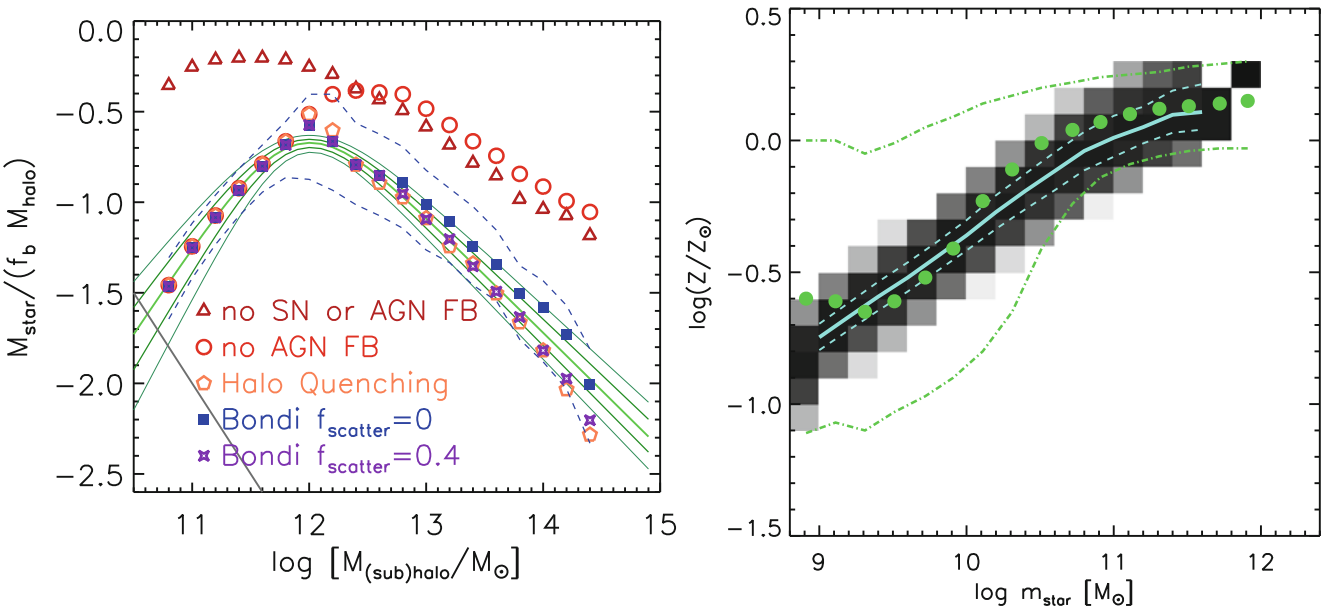


Fig. 10.31 *Left panel:* The fraction of baryons in the form of stars, as a function of halo mass, as predicted by a semi-analytic model. The *brown triangles* show the stellar mass fraction for a model run where no feedback was included. In this case, for galaxy-mass halos ($\sim 10^{12} M_\odot$) most of the baryons have been converted into stars. For larger halos masses, the fraction decreases, since cooling becomes less efficient in these halos. The *green curves* show the range of stellar mass fractions that is obtained from observations. Obviously, the no-feedback assumption violates observational constraints on all mass scales. The *red circles* show results from a model in which supernova feedback was included, but no feedback from AGN. Here, the stellar mass fraction is very substantially reduced at the low-mass end, bringing it into the observed range; however, supernovae are inefficient at high halo masses. The other three types of *symbols* correspond to different assumptions about AGN feedback; clearly, AGN feedback is needed to account for the small star-formation efficiency in high-mass halos, such as groups and clusters. *Right panel:* The metallicity as a function of stellar mass. *Grey shades* indicate the probability distribution that a galaxy of stellar mass $m_{\rm star}$ has a metallicity Z (in Solar units), with the *solid curve* showing the median and the *dashed curve* the 1-σ range, as obtained from a semi-analytic model. The *green points* show the observed metallicity of galaxies. The median of the two distributions agree very well, though the spread is considerably larger in the observed galaxies. Source: R. Somerville et al. 2008, *A semi-analytic model for the co-evolution of galaxies, black holes and active galactic nuclei*, MNRAS 391, 481, p. 492, 494, Figs. 3, 6. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

Minor mergers. The merger trees obtained from the N-body simulations describe for each halo at which time it merges with another one. As we discussed above, the outcome of a merger will depend to a large degree on the mass ratio of the two merging halos (and galaxies): If the mass ratio is substantially different from unity (e.g., smaller than 1:3; minor merger), the merger will cause little damage to the galaxy of the more massive component, whereas for almost equal mass mergers, one expects that both galaxies will be destroyed and the stellar distribution be changed drastically.

If the masses of the two components in a merger are very different, the merging process of the two components does not occur instantaneously, but since the smaller galaxy will have, in general, a finite orbital angular momentum, it will first enter into an orbit around the more massive component. The smaller mass halo and galaxy can survive as a satellite galaxy. This satellite galaxy is subject to several processes, though. By moving through the hot gas of the larger halo, ram-pressure stripping can remove gas, at a rate depending on the gas density in the halo, the orbit of the satellite (as determined by the N-body simulation), and the gas density of the satellite (which has been recorded by the earlier evolution of that galaxy before the merger event). The stripped gas is added to the gas distribution of the main halo. The stripping of the gas reduced the reservoir from which the satellite can form stars, a process which explains that satellite galaxies in groups and clusters are usually redder than their central galaxy.

Furthermore, dynamical friction (see Sect. 6.3.3) changes the orbit of the satellite in time, bringing it closer to the halo center. Once that happens, the cold gas and the stars of the satellite galaxy are added to the disk component of the central galaxy of the halo.

It may also be that the orbit of a satellite galaxy comes close to the center of the main halo where the tidal forces are

strong. In such a case, the galaxy may be tidally disrupted. Since the satellite in this case has a large velocity relative to the central galaxy, its stars are then assumed to be dispersed in the halo, contributing to the intracluster stellar population which has been found in individual clusters, as well as in the cluster population as a whole (see Sect. 6.3.4). Since the cold gas and the stars are more concentrated than the dark matter subhalo of the satellite, the galaxy (i.e., stars + gas) may survive tidal effects, even after the dark matter subhalo has been tidally disrupted. Hence, there may be orphan galaxies—satellite galaxies without a corresponding dark matter subhalo.

Major mergers. If the two merging galaxies have a mass ratio close to unity (i.e., larger than $\sim 1:3$), it is assumed that their disks are completely destroyed and their stars being rearranged into a spheroidal distribution. Furthermore, a fraction of the sum of the cold gas in both components is assumed to undergo a starburst. The newly formed stars are added to the spheroidal stellar component. For minor mergers, a corresponding collisional starburst can be added as well, where the newly formed stars are added to the disk component. The resulting strong supernova feedback can then expel most of the remaining gas from the remnant of a major merger, leaving a (gas-poor) elliptical galaxy.

After the formation, an elliptical can attain new cold gas from the cooling of hot gas in the halo, accretion of surrounding material, or subsequent minor merger events. By these processes, a new disk population may form. In this model, a spiral galaxy is created by forming a bulge in a 'major merger' at early times, with the disk of stars and gas being formed later in minor mergers and by accretion of gas. Hence the bulge of a spiral is, in this picture, nothing but a small elliptical galaxy, which is also suggested by the very similar characteristics of bulges and ellipticals, including the fact that both types of objects seem to follow the same relation between the black hole mass and the stellar velocity dispersion, as explained in Sect. 3.8.3.

Black hole growth, and feedback from AGN. When they form, galaxies are implanted a central black hole of small seed mass, as described above for the hydrodynamical simulations. The mass of the black holes then grows as a result of mergers and accretion of gas. The former process drives gas into the center of the galaxies, where a star-formation episode sets in; this process also feeds gas onto the supermassive black hole. The two SMBHs in a merger event are assumed to also merge. In this mode of accretion, star formation and AGN activity happen in parallel, and so do the corresponding feedback processes. Hence, only their sum is relevant.

However, we have seen in clusters that feedback must be highly efficient in suppressing cooling flows, and found clear direct evidence for the AGN feedback on the intracluster medium, in the form of extended radio emission (e.g., jets), and the corresponding cavities in the X-ray emitting gas. The corresponding AGN activity is rather moderate in terms of overall luminosity—the center of cool-core clusters usually do not contain a bright QSO, despite the large mass of the central galaxy and the corresponding large mass of the SMBH. Hence, these AGNs must accrete at a rate substantially lower than the Eddington rate. In this mode, a large fraction of the energy is released in form of radio jets, i.e., kinetic energy of a relativistic plasma. This 'radio-mode' accretion is highly inefficient in generating optical and UV-radiation. It is assumed that this low-rate accretion is related to a cooling flow from the intracluster medium. A simple picture would be that of a self-regulating feedback which quenches the cooling once it becomes too effective, thus leading to a large accretion rate, and subsequently a larger energy output from the central SMBH. Suppressing the cooling then reduces the accretion flow, leading to a decreased accretion rate, less feedback, and consequently,

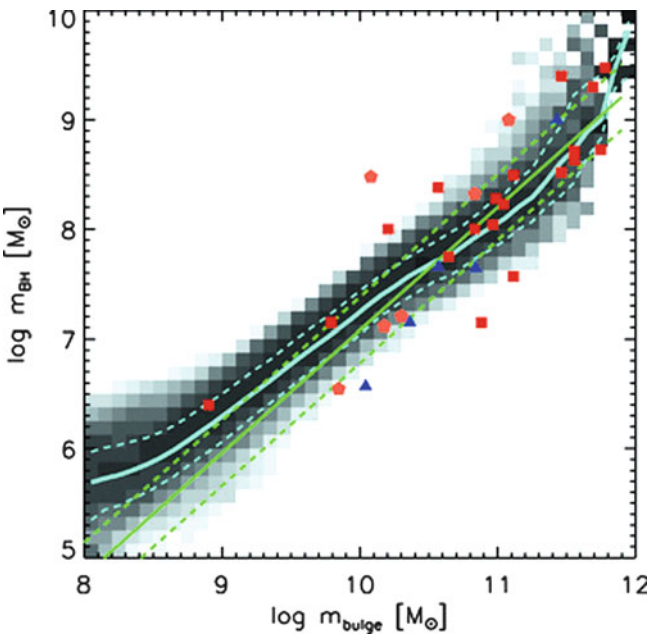


Fig. 10.32 The black hole mass vs. bulge mass relation, as predicted from a semi-analytic model. *Grey shading* indicates the probability distribution of the black hole mass for a given bulge mass, the *blue solid* and *dashed curves* yield the median of the black hole mass and its 1-σ range. The *green lines* show the corresponding results from observations, whereas *symbols* show individual observed galaxies. Source: R. Somerville et al. 2008, *A semi-analytic model for the co-evolution of galaxies, black holes and active galactic nuclei*, MNRAS 391, 481, p. 495, Figs. 7. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

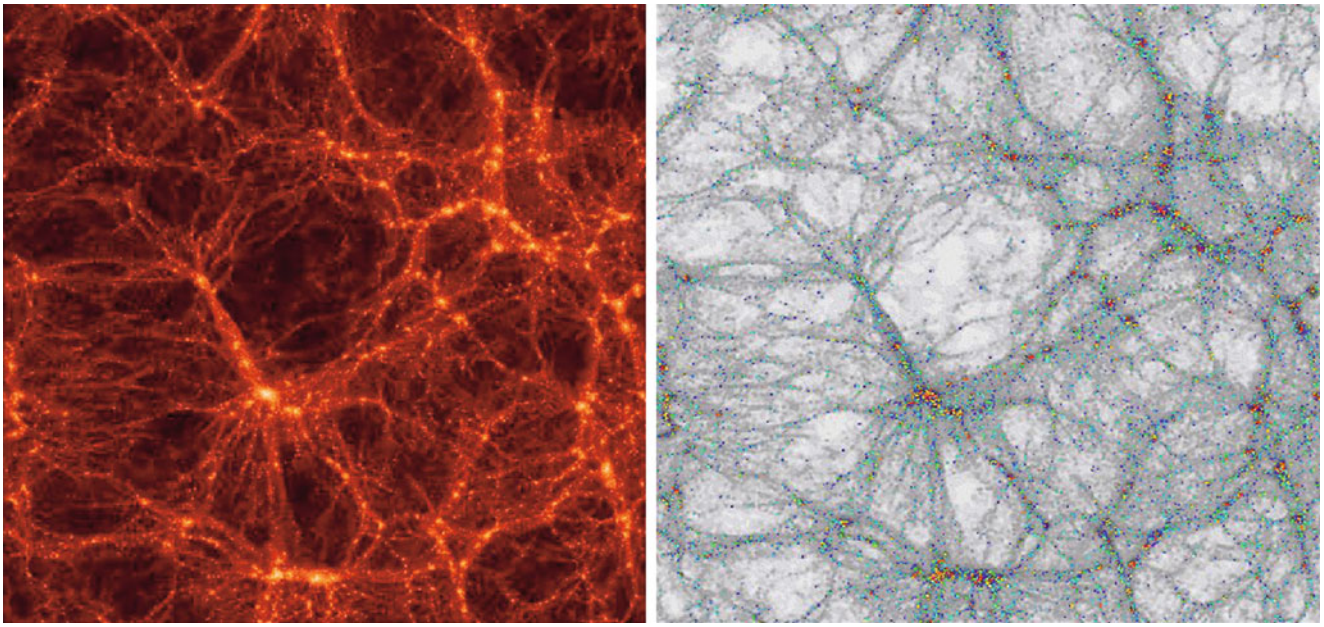


Fig. 10.33 *On the left,* the distribution of dark matter resulting from an N-body simulation is shown. The dark matter halos identified in this mass distribution were then modeled as the location of galaxy formation—the formation of halos and their merger history can be followed explicitly in the simulations. Semi-analytic models describe the processes which are most important for the gas and the formation of stars in halos, from which a model for the distribution of galaxies is built. In the *panel on the right*, the resulting distribution of model galaxies is represented by *colored dots*, where the color indicates the spectral energy distribution of the respective galaxy: galaxies with active star formation are shown in *blue*, while galaxies which are presently not forming any new stars are marked in *red*. The latter are particularly abundant in clusters of galaxies—in agreement with observations. Credit: G. Kauffmann, J. Colberg, A. Diaferio & S.D.M. White, and the GIF-Collaboration

higher cooling rate after some time.[11] In semi-analytic models, the accretion rate can then be calculated from the cooling rate of the hot gas in the halo, and a certain fraction of the resulting energy release is assumed to be used for heating the gas in the halo.

The importance of this AGN feedback can be seen in the left panel of Fig. 10.31, where the stellar mass fraction of baryons is shown as a function of halo mass. AGN feedback is essential to suppress star formation in high-mass halos, i.e., to explain the small ratio of stellar-to-hot gas mass in galaxy clusters. Supernova feedback by itself is not efficient in high-mass halos. Furthermore, these models are successful in reproducing the relation between the SMBH mass and the properties of the stellar population, such as the bulge mass (see Fig. 10.32), luminosity, or velocity dispersion of the spheroidal component.

Stellar populations and chemical evolution. For each galaxy formed, the models keep track of their star-formation history. Hence, one can assign to each galaxy the stellar populations formed in time, once an initial mass function is selected. Using stellar population synthesis models, one can then obtain the stellar luminosity and spectral energy distribution for each galaxy [using (3.37)], and turn these parameters into 'observables', like magnitude and colors. In order to compare these predictions to observations, the effects of dust need to be accounted for. The amount of dust depends on the amount of gas and the metallicity of the gas which in turn is determined by the history of chemical enrichment. This is followed for each galaxy by the amount of metals ejected into the gas by supernovae and stellar winds. These metals are then mixed with the other gas, the newly forming stars are assigned the corresponding metallicity of the cool gas. In this way, the models can make predictions of observable properties of galaxies and their statistical distribution.

10.7.2 Results from semi-analytic models

The free parameters in semi-analytic models—such as the star-formation efficiency or the fraction of energy from SNe that is transferred into the gas—are fixed by comparison with some key observational results. For example, one requires that the models reproduce the correct normalization of the

[11] Of course, this simple picture ignores all the difficulties in understanding the transport of gas from large distances to the immediate vicinity of the black hole where it can be accreted.

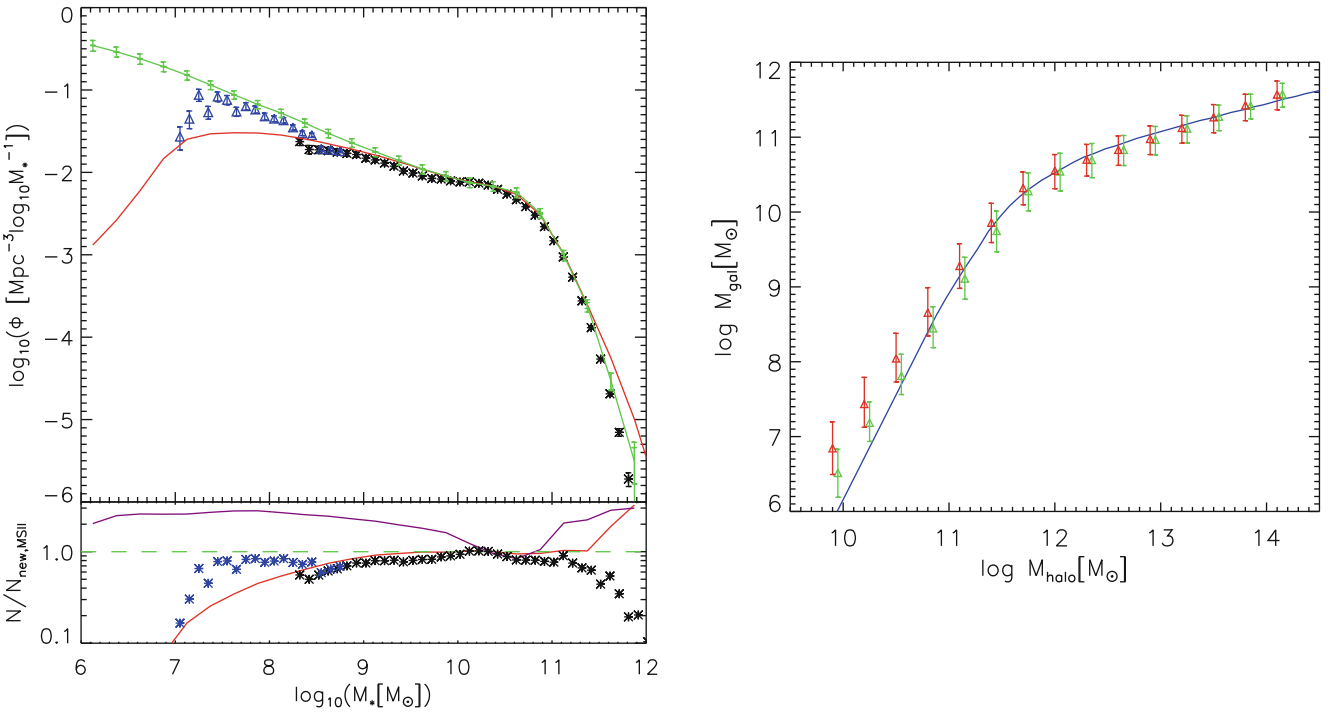


Fig. 10.34 *Upper left panel:* The stellar mass function of galaxies as obtained from a semi-analytic model for which the Millennium Simulation (MS) and the Millennium II (MS-II) simulations have been used for describing the dark matter evolution (see Sect. 7.5.3). The *red* and *green curves* show the model predictions from the MS and the MS-II, respectively. Owing to the better spatial and mass resolution of the MS-II, the stellar mass function can be followed to considerably smaller masses. *Black* and *blue points* show observational results as obtained from the SDSS; at the lowest mass end, the observed galaxies come from a very small local volume, and are therefore subject to a substantial 'cosmic variance'. *The lower left panel* shows the ratio of the mass functions relative to the predictions from the MS-II. Clearly, the semi-analytic model can reproduce the observed mass function accurately over some 4 orders of magnitude (the *purple curve* in the lower panel shows the corresponding results from an earlier incarnation of semi-analytic modelling, where in particular the feedback was assumed to be weaker). *The right panel* shows the mean stellar mass and its dispersion as a function of the halo mass, as obtained from the simulations. *Green symbols* are for central galaxies of halos, whereas the *red symbols* correspond to satellite galaxies (where the corresponding halo mass is the mass of their subhalos at the time the satellite has merged with the main halo). The *blue curve* is obtained if the dark matter halo abundance is directly matched to the stellar mass function, assuming a monotonic dependence between these two quantities. Note that the slope of the relation is considerably steeper than unity at the low-mass end, and much flatter at the high-mass end. This relation therefore explains the different shapes of the halo mass and stellar mass functions shown in Fig. 10.2. Source: Q. Guo et al. 2011, *From dwarf spheroidals to cD galaxies: simulating the galaxy population in a ΛCDM cosmology*, MNRAS 413, 101, p. 115, 117, Figs. 7, 9. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

Tully–Fisher relation and that the number counts of galaxies match those observed. Although these models are too simplistic to trace the processes of galaxy evolution in detail, they are highly successful in describing the basic aspects of the galaxy population, and they are continually being refined. These refinements make use of empirical results (such as the Schmidt-Kennicutt law for star formation) and theoretical progress, such as detailed simulations of the merger process between pairs of galaxies. The outcome from such simulations are summarized in analytic expressions which are then applied to the semi-analytic models. In this section, we want to show some of the results from these models.

Red versus blue galaxies. For instance, all semi-analytic models predict that galaxies in clusters basically consist of old stellar populations, because here the interaction processes concluded already quite early in cosmic history. Therefore, at later times cold gas is no longer available for the formation of stars. Fig. 10.33 shows the outcome of such a model in which the merger history of the individual halos has been taken straight from the numerical N-body simulation, hence the spatial locations of the individual galaxies are also described by these simulations.

By comparison of the results from such semi-analytic models with the observed properties of galaxies and their spatial distribution, the models can be increasingly refined. In this way, we obtain more realistic descriptions of those processes which are included in the models in a parametrized form. This comparison is of central importance for achieving further progress in our understanding of the complex processes that are occurring in galaxy evolution, which can not be studied in detail by observations.

Stellar mass function and stellar-to-total mass ratio. Combining two dark matter simulations with the same cosmological parameters but different box size and spatial resolution (namely the Millennium and Millennium-II simulations; see Sect. 7.5.3), the properties of galaxies can be predicted over a very wide range of masses. For example, the left-hand side of Fig. 10.34 shows the predicted stellar mass function of galaxies at redshift $z = 0$, compared to results from observations. We see that the model can reproduce the observations over a range of several orders of magnitude in stellar mass. Key to this achievement are the feedback processes, as already discussed in connection with Fig. 10.31; together with the temperature- (and mass-)dependent cooling function of gas, they determine the overall efficiency of turning gas into stars, and thus lead to a preferred mass scale where the stellar-to-total mass of halos is maximized (see Fig. 10.2). This can also be seen in the right panel of Fig. 10.34 which plots the mean stellar mass as a function of halo mass. There, one can also see the characteristic mass scale where the slope of this relation changes sharply.

Tully–Fischer relation. Traditionally, galaxy evolution models had problems of reproducing the Tully–Fisher relation for disk galaxies (see Sect. 3.4.1). The implementation of the aforementioned result from numerical simulations, namely that the rotational velocity of a disk is well approximated by the maximum velocity of the corresponding NFW halo, largely solves this problem, as can be seen in Fig. 10.35 which shows the predicted relation between luminosity and rotational velocity for disk-dominated galaxies, compared to the observed Tully–Fisher relation. The agreement between these two distributions is fairly good, in particular concerning the overall amplitude. Whereas the shape of the Tully–Fisher relation in the model is not truly a power law, this may be related to a slightly too efficient feedback in massive galaxies, which decreases their luminosity.

Spatial distribution and correlation function. Since the spatial location of the galaxies is known from such simulations, one can compare their spatial distribution with that of the galaxies from redshift surveys. This is illustrated in Fig. 10.36, which shows a comparison of wedge diagrams from redshift surveys with those obtained from the semi-analytic models applied to the dark matter distribution of the Millennium simulation. At least at first sight, the statistical properties of the 'red' and 'blue' wedge diagrams are the same. The model predicts the occurrence of 'Great Walls', as well as the system of voids and filaments in the overall galaxy distribution. This comparison can be made more quantitative, for example by comparing the two-point correlation function of model galaxies with

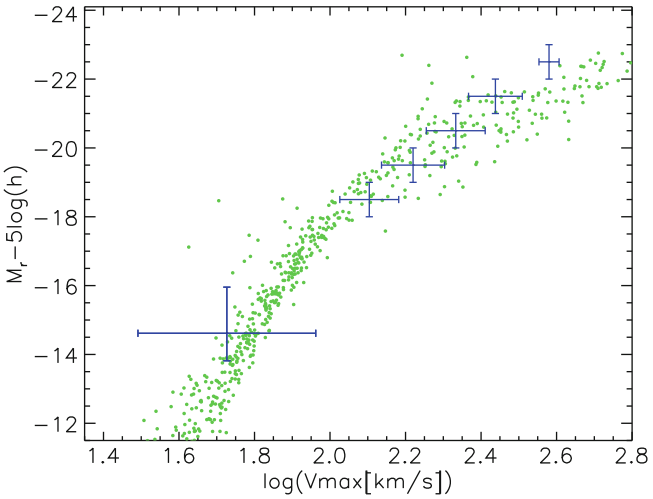


Fig. 10.35 The Tully–Fisher relation in the r-band. *Green points* show the absolute r-band magnitude of disk-dominated galaxies as a function of maximum rotational velocity of their host halos, as obtained from the same semi-analytic model as shown in Fig. 10.34. This is compared to observational results indicated by the *blue crosses*. Semi-analytic models are thus able to reproduce the zero point and approximate shape of the Tully–Fisher relation over a range of about 8 magnitudes. Source: Q. Guo et al. 2011, *From dwarf spheroidals to cD galaxies: simulating the galaxy population in a ΛCDM cosmology*, MNRAS 413, 101, p. 119, Fig. 13. Reproduced by permission of Oxford University Press on behalf of the Royal Astronomical Society

that obtained from observations. Also here, good qualitative agreement is found, though the simulation slightly overpredicts the amplitude of the correlation function. This, however, may be due to the fact that the normalization of the power spectrum was chosen to be $\sigma_8 = 0.9$, slightly larger than the current best estimates for our Universe (see Sect. 8.7).

The correlation function of galaxies has a rather different behavior as a function of scale and redshift than that of the dark matter. In Fig. 10.37, the correlation function of luminous galaxies and that of the overall matter distribution is shown for four different redshifts. Several issues are remarkable. First, the dark matter correlation function $\xi_m(r)$ is not well approximated by a power law, whereas the galaxy correlation function $\xi_g(r)$ shows a power-law behavior over many decades of spatial scale, in agreement with observed galaxy correlation functions. At $z = 0$ (red curve), $\xi_g(r)$ almost traces the correlation function of matter on scales $r \gtrsim 1h^{-1}$ Mpc, but they disagree substantially on smaller scales. This implies that the bias of galaxies is strongly scale-dependent, at least on small scales. In fact, the question arises as to which processes in the evolution of galaxies may produce such a perfect power law: why does the bias factor behave just such that ξ_g attains this simple shape. The answer is found by analyzing galaxies with and without active star formation separately; for each of these sub-populations of

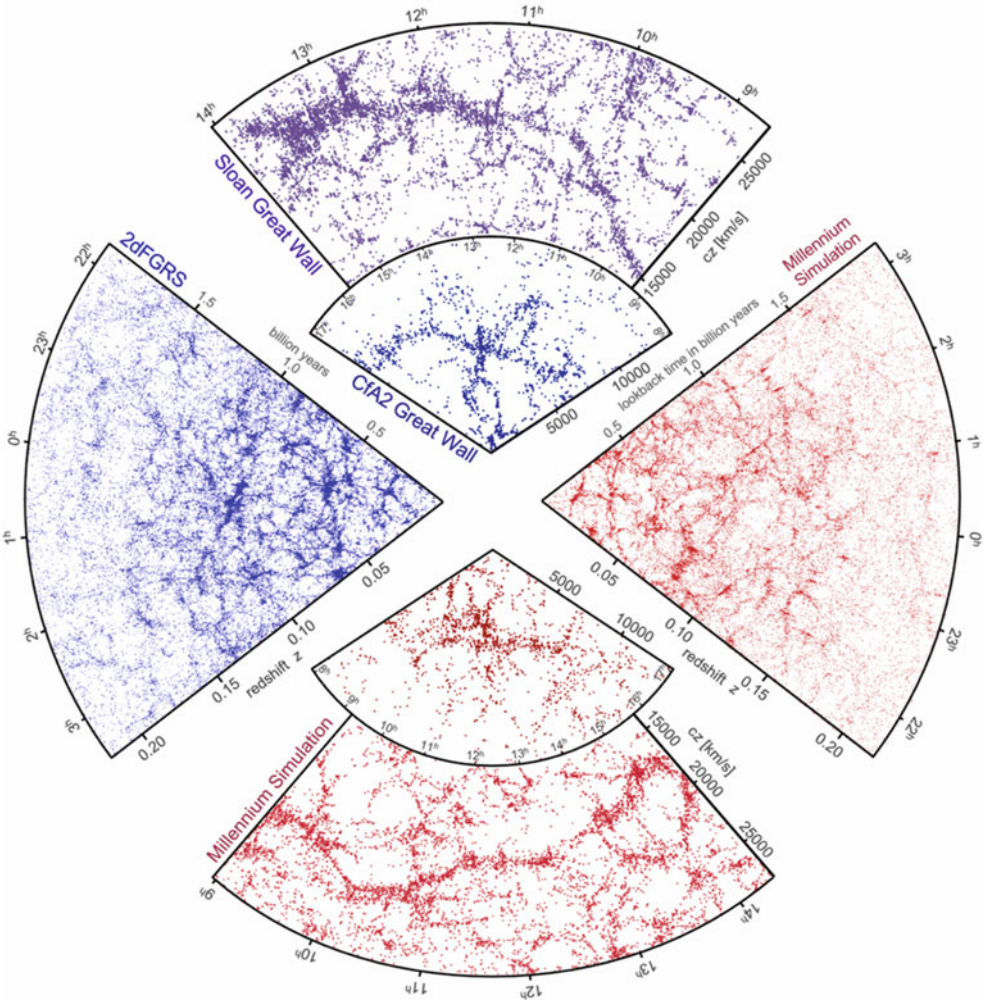


Fig. 10.36 Large-scale distribution of galaxies as obtained from redshift surveys (in *blue*) and from semi-analytic models of galaxies in the Millennium simulation (in *red*). On the *left*, one hemisphere of the 2dFGRS is shown (cf. Fig. 7.1), whereas on the *top* the small wedge diagram shows the CfA2 redshift survey (Fig. 7.2) with the Coma cluster at its center, and the large wedge is part of the SDSS. In much the same way as the observed distributions are obtained, the galaxy distribution from the Millennium simulation has been transformed into wedge diagrams shown in *red*. They are very similar to the observed ones—they show great walls, fingers of god (since the model galaxies are plotted in redshift space, as their peculiar velocity is given by the simulation), as well as the cellular structure of filaments and voids. Source: V. Springel et al. 2006, *The large-scale structure of the Universe*, Nature 440, 1137, Fig. 1. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2006

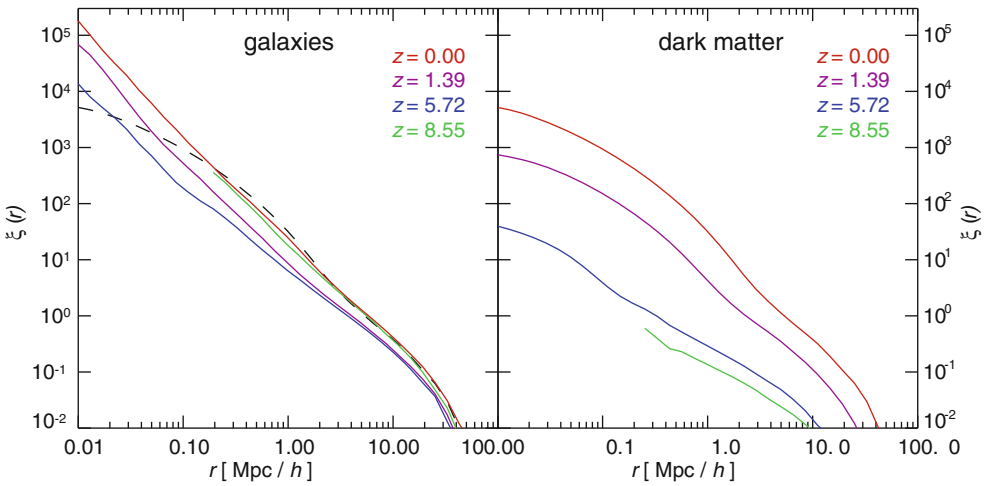


Fig. 10.37 The correlation function of galaxies (*left*) and dark matter (*right*) in the Millennium simulation, for different redshifts. The *dashed curve* in the *left panel* shows the $z = 0$ dark matter correlation, for easier comparison. The galaxies are selected above a given I-band luminosity. There are striking differences between these two correlations. As expected from structure growth, the dark matter correlation function decreases with increasing redshift (remember, on large scales where structure evolution follows linear perturbation theory, $\xi(r, z) \propto D_+^2(z)$. In contrast to that, the evolution of the galaxy correlation function is much smaller, and it is not monotonic with redshift: the correlation at the highest redshift is almost the same as the one at $z = 0$. Source: V. Springel et al. 2006, *The large-scale structure of the Universe*, Nature 440, 1137, Fig. 5. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2006

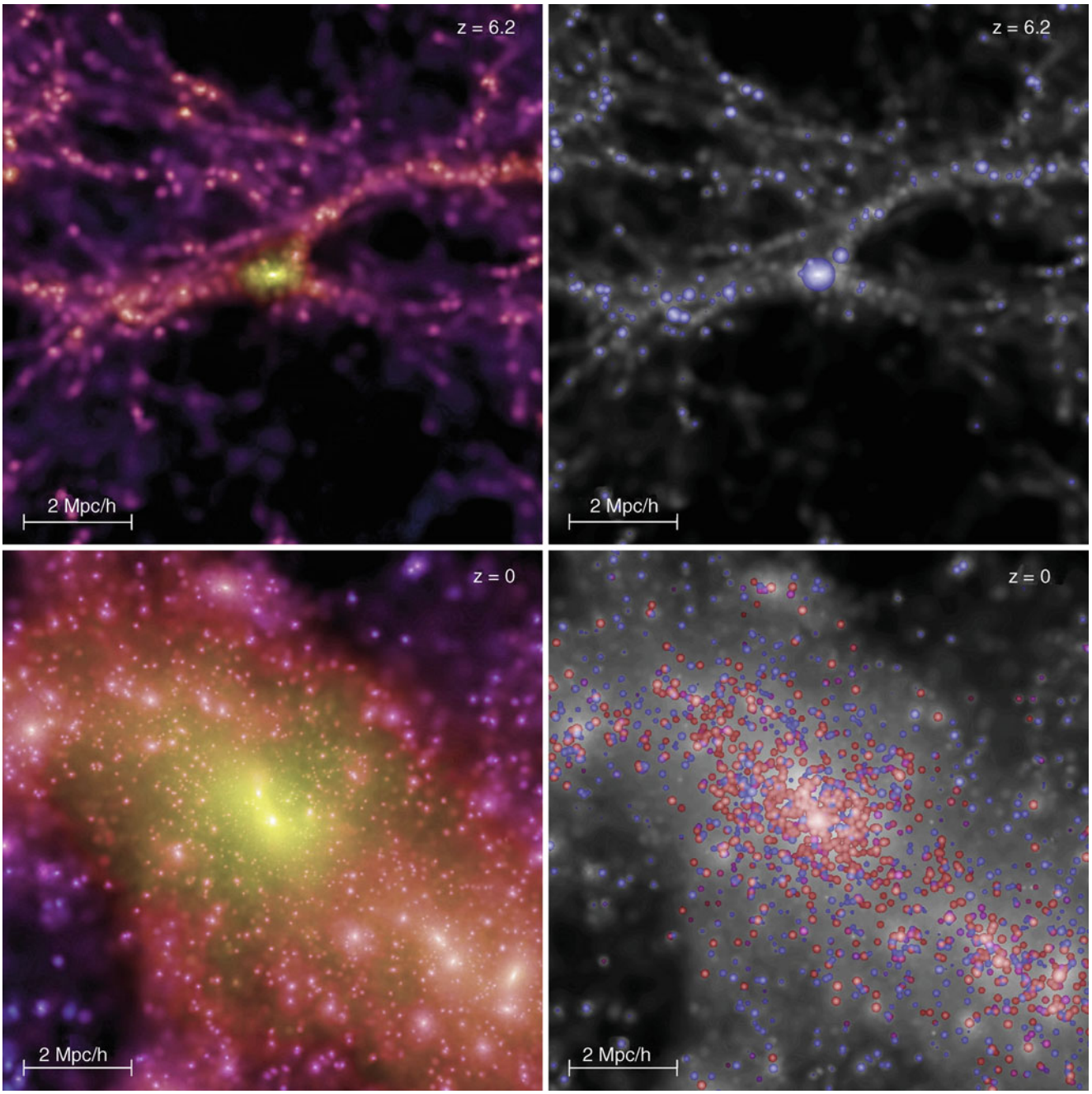


Fig. 10.38 In the *top panels*, one of the most massive halos at $z = 6.2$ from the Millennium simulation (see Fig. 7.13) is shown, whereas in the *bottom panels*, the corresponding distribution in this spatial region at $z = 0$ is shown. Thus, this early massive halo is now located in the center of a very massive galaxy cluster. In the *panels on the left*, the mass distribution is displayed. The corresponding distribution of galaxies as determined from a semi-analytic model is shown in the *right-hand panels*. Galaxies at $z = 6.2$ are all blue since their stellar population must be young, whereas at $z = 0$, most galaxies contain an old stellar population, here indicated by the *red color*. Each of the panels shows the projected distribution of a cube with a comoving side length of $10\,h^{-1}$ Mpc. Source: V. Springel et al. 2005, *Simulating the joint evolution of quasars, galaxies and their large-scale distribution*, Nature 435, 629, Fig. 3. Reprinted by permission of Macmillan Publishers Ltd: Nature, ©2005

galaxies, ξ_g is *not* a power law. Therefore, the simple shape of the correlation function shown in Fig. 10.37 is probably a mere coincidence ('cosmic conspiracy').

Second, the matter correlation function strongly decreases with increasing redshift, whereas ξ_g evolves much slower with z. This implies that the bias of galaxies is redshift dependent; for a given galaxy luminosity (or stellar mass), the bias increases with redshift. In fact, we see that ξ_g at $z = 8.55$ is almost identical with the one at zero redshift—the dependence of ξ_g on redshift is not monotonic.

Early QSOs. Another result from such models is presented in Fig. 10.38, also from the Millennium simulation. Here, one of the most massive dark matter halos in the simulation box at redshift $z = 6.2$ is shown, together with the mass distribution in this spatial region at redshift $z = 0$. In both cases, besides the distribution of dark matter, the galaxy distribution is also displayed, obtained from semi-analytic models. Massive halos which have formed early in cosmic history are currently found predominantly in the centers of very massive galaxy clusters. Assuming that the luminous QSOs at $z \sim 6$ are harbored in the most massive halos of that epoch, we might suppose that these may today be identified as the central galaxies in clusters.[12] This may provide an explanation as to why so many central, dominating cluster galaxies show AGN activity, though with a smaller luminosity due to small accretion rates.

From what we presented in this section, we can summarize that semi-analytic modelling of galaxies is a very useful method to make the link between the dark matter distribution on the one hand, and the properties of the galaxy population on the other. Since semi-analytic models are computational inexpensive, compared to gas-dynamical simulations, one can experiment with them and study in detail the dependence of galaxy properties on certain assumptions and parameter choices. Furthermore, these models allow us to include our best knowledge and understanding of the various complex baryonic processes in a unified way which yields quantitative results. We also have a fairly good understanding of the mean properties of galaxies and their central black holes. Much of this knowledge has been obtained only in recent years, and there is no doubt that future observational results will lead to further refinements, and perhaps qualitative modifications, of our understanding.

[12] This will not be true in every single case; as can be seen from Fig. 10.26, the most massive SMBHs at high redshifts are not necessarily the mass record holder at later epochs.

Outlook

11

The past decade has been a tremendously active and fruitful time for extragalactic astronomy and cosmology, as hopefully is well documented in the previous chapters. Here, we will try to see what progress we may expect for the near and not-so near future.

11.1 Continuous progress

Progress in (extragalactic) astronomy is achieved through information obtained from increasingly improving instruments and by refining our theoretical understanding of astrophysical processes, which in turn is driven by observational results. It is easy to foresee that the evolution of instrumental capabilities will continue rapidly in the near future, enabling us to perform better and more detailed studies of cosmic sources. Before we will mention some of the forthcoming astronomical facilities, it should be pointed out that some of the recently started projects have at best skimmed the cream, and the bulk of the results is yet to come. This concerns the scientific output from the Herschel and Planck satellite missions, as well as the recently commissioned ALMA interferometer, which has already provided exciting results. The great scientific capabilities of these facilities have been impressively documented, so it is easy to predict that far more scientific breakthroughs are waiting to be achieved with them.

Within a relatively short period of $\sim$ 15 years, the total collecting area of large optical telescopes has increased by a large factor, as is illustrated in Fig. 11.1. At the present

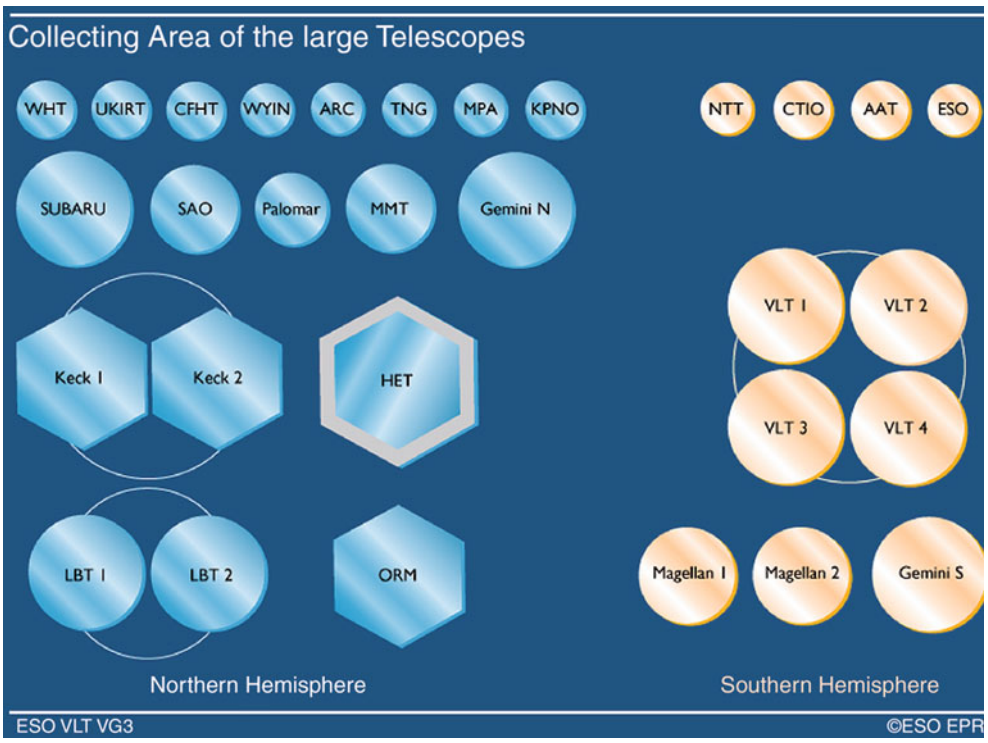


Fig. 11.1 The collecting area of large optical telescopes is displayed. Those in the Northern hemisphere are shown on the *left*, whereas southern telescopes are shown on the *right*. The joint collecting area of these telescopes has been increased by a large factor over the past two decades: only the telescopes shown in the *upper row* plus the 5-m Palomar telescope and the 6-m SAO were in operation before 1993. If, in addition, the parallel development of detectors is considered, it is easy to understand why observational astronomy is making such rapid progress. Credit: European Southern Observatory

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7_11,

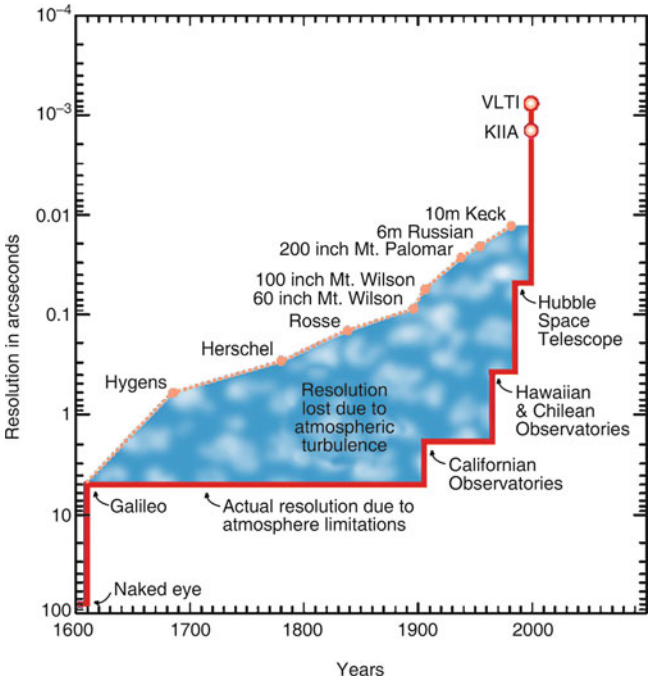


Fig. 11.2 This figure illustrates the evolution of angular resolution as a function of time. The *upper dotted curve* describes the angular resolution that would be achieved in the case of diffraction-limited imaging, which depends, at fixed wavelength, only on the aperture of the telescope. Some historically important telescopes are indicated. The *lower curve* shows the angular resolution actually achieved. This is mainly limited by atmospheric turbulence, i.e., seeing, and thus is largely independent of the size of the telescope. Instead, it mainly depends on the quality of the atmospheric conditions at the observatories. For instance, we can clearly recognize how the opening of the observatories on Mount Palomar, and later on Mauna Kea, La Silla and Paranal have lead to breakthroughs in resolution. A further large step was achieved with HST, which is unaffected by atmospheric turbulence and is therefore diffraction limited. Adaptive optics and interferometry characterize the next essential improvements. Credit: European Southern Observatory

time, 13 telescopes with apertures above 8 m (and four more with an aperture of 6.5 m) are in operation, the first of which, Keck I, was put into operation in 1993. In addition, the development of adaptive optics allows us to obtain diffraction-limited angular resolution from ground-based observations (see Fig. 11.2).

The capability of existing telescopes gets continuously improved by installing new sensitive instrumentation. The successful first generation of instruments for the 10-m class telescopes gets replaced step-by-step by more powerful instruments. As an example, the Subaru telescope will be equipped with Hyper Suprime-Cam, a 1.5 $\deg^2$ camera, by far the largest of its kind on 10-m class telescope. This instrument will allow the conduction of large-area deep imaging surveys, e.g., for cosmic shear studies.

In another step to improve angular resolution, optical and NIR interferometry will increasingly be employed. For example, the two Keck telescopes (Fig. 1.38) are mounted such that they can be used for interferometry. The four unit telescopes of the VLT can be combined, either with each other or with additional (auxiliary) smaller telescopes, to act as an interferometer (see Fig. 1.39). The auxiliary telescopes can be placed at different locations, thus yielding different baselines and thereby increasing the coverage in angular resolution. Finally, the Large Binocular Telescope (LBT, see Fig. 1.44), which consists of two 8.4-m telescopes mounted on the same platform, was developed and constructed for the specific purpose of optical and NIR interferometry.

From HST to its successor. The Hubble Space Telescope has turned out to be the most successful astronomical observatory of all time (although it certainly was also the most expensive one).[1] The importance of HST for extragalactic astronomy is not least based on the characteristics of galaxies at high redshifts. Before the launch of HST, it was not known that such objects are small and therefore have, at a given flux, a high surface brightness. This demonstrates the advantage of the high resolution that is achieved with HST. Several service missions to the observatory led to the installation of new and more powerful instruments which have continuously improved the capacity of HST. With the Space Shuttle program abandoned, no more service to Hubble is possible, and it is only a matter of time before essential parts will start to malfunction.

Fortunately, the successor of HST is already at an advanced stage of construction and is currently scheduled to be launched in 2018. This Next Generation Space Telescope (which was named James Webb Space Telescope—JWST; see Fig. 11.3) will have a mirror of 6.5-m diameter and therefore will be substantially more sensitive than HST. Furthermore, JWST will be optimized for observations in the NIR (1–5 μm) and thus be able, in particular, to observe sources at high redshifts whose stellar light is redshifted into the NIR regime of the spectrum.

We hope that JWST will be able to observe the first galaxies and the first AGNs, i.e., those sources responsible for reionizing the Universe. Besides a NIR camera, JWST will carry the first multi-object spectrograph in space, which is optimized for spectroscopic studies of high-redshift galaxy samples and whose sensitivity will exceed that of all previous instruments by a huge factor. Furthermore, JWST will carry

[1] The total price tag on the HST project will probably be on the order of 10 billion US dollars. This is comparable to the total cost of the Large Hadron Collider and its detectors. I am convinced that a particle physicist and an astrophysicist can argue for hours which of the two investments is more valuable for science—but how to compare the detection of the Higgs boson with the manifold discoveries of HST! However, both, the particle physicist and the astrophysicist, easily agree that the two price tags are a bargain, when compared to an estimated cost of 45 billion US dollar for the B2 stealth bomber program.

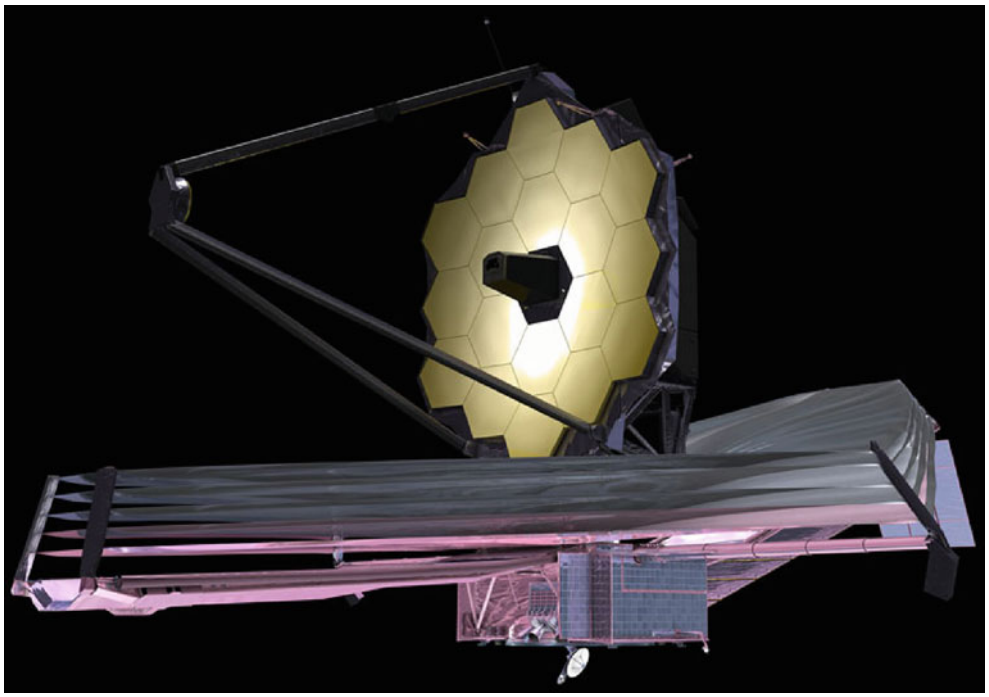


Fig. 11.3 Artist's impression of the 6.5-m James Webb Space Telescope. Like the Keck telescopes, the mirror is segmented and protected against Solar radiation by a giant heat shield, having the size of a tennis court. Keeping the mirror and the instruments permanently in the shadow will permit a passive cooling at a temperature of $\sim 35\,\mathrm{K}$. This will be ideal for conducting observations at NIR wavelengths, with unprecedented sensitivity. Credit: NASA

a MIR instrument which was developed for imaging and spectroscopy in the wavelength range $5\,\mu\mathrm{m} \leq \lambda \leq 28\,\mu\mathrm{m}$.

11.2 New facilities

There are research fields where a single instrument or telescope can yield a breakthrough—an example would be the determination of cosmological parameters from measurements of the CMB anisotropies. However, most of the questions in (extragalactic) astronomy can only be solved by using observations over a broad range of wavelengths; for example, our understanding of AGNs would be much poorer if we did not have the panchromatic view, from the radio regime to TeV energies. Our inventory of powerful facilities is going to be further improved, as the following examples should illustrate.

New radio telescopes. The Square Kilometer Array (SKA) will be the largest radio telescope in the world and will use a technology which is quite different from that of current telescopes. For SKA, the beams of the telescope will be digitally generated on computers. Such digital radio interferometers not only allow a much improved sensitivity and angular resolution, but they also enable us to observe many different sources in vastly different sky regions simultaneously. SKA will consist of about 3000 15-m dishes as well as two other types of radio wave receivers, known as aperture array antennas. Together, the receiving area amounts to about one square kilometer. The telescopes are spread over a region $\sim 3000\,\mathrm{km}$ is size, yielding an angular resolution of $0\farcs02$ at $\nu = 1.4\,\mathrm{GHz}$, and are linked by optical fibers (with a total length of almost $10^5\,\mathrm{km}$). The instantaneous field-of-view at frequencies $\gtrsim 1\,\mathrm{GHz}$ will be $\sim 1\,\mathrm{deg}^2$, increasing to $\sim 200\,\mathrm{deg}^2$ for lower frequencies. This huge (in terms of current radio interferometers) field-of-view is achieved by synthesizing multiple beams using software. The limits of such instruments are no longer bound by the properties of the individual antennas, but rather by the capacity of the computers which analyze the data. SKA will provide a giant boost to astronomy; for the first time ever, the achievable number density of sources on the radio sky will be comparable to or even larger than that in the optical.

SKA is not the first of this new kind of radio telescopes. The first one is the Low-Frequency Array (LOFAR), centered in the Netherlands but with several stations located in neighboring countries to increase the baseline and thus the angular resolution. LOFAR, operating at $\nu \lesssim 250\,\mathrm{MHz}$, can be considered as a pathfinder for the low-frequency part of SKA. LOFAR began its routine operation at the end of 2012. Other pathfinder observatories for SKA include the Australian Square Kilometre Array Pathfinder (ASKAP), and MeerKAT in South Africa.

To avoid terrestrial radio emission as much as possible, the SKA will we constructed in remote places in Australia and South Africa. The remoteness brings with it several great challenges—to mention just one, the power supply will most likely be decentralized, i.e., obtained through Solar panels near the telescopes to generate electricity. The data rate to be transmitted is far larger than the current global

Internet traffic! To process the huge data stream, one needs a computer capable of about 100 petaflops per second—such a computer does not yet exist (at least not with access for scientists).

The scientific outcome from SKA and its pathfinders is expected to be truly revolutionary. To mention just a few: The epoch of reionization will be studied by the redshifted 21-cm hydrogen line; a detailed time- and spatially dependent picture of the reionization process will be obtained. Normal galaxies can be studied via their 21-cm and their continuum (synchrotron) emission out to large redshifts, with an angular resolution better than that of HST. Since the HI-line yields the redshift of the galaxies, large redshift surveys can be employed for studies of the large-scale structure, including baryonic acoustic oscillations. The beam of the interferometer represents the point-spread function, and is very well known. Weak gravitational lensing studies using the radio emission from normal galaxies can thus make use of that knowledge to correct the measured image shapes.

New (Sub-)millimeter telescopes. The Large Millimeter Telescope (LMT) on the Volcán Sierra Negra, Mexico, is a 50-m radio telescope that recently went into operation (though in its first phase, the inner 32-m diameter of its primary surface will be fully installed). The LMT will observe in the range $0.85\,\mathrm{mm} \leq \lambda \leq 4\,\mathrm{mm}$. In addition to the much increased surface area compared to existing single-dish telescopes in this wavelength regime, the large aperture will provide an important step forward in angular resolution, and thus provide far more accurate positions of (sub-)mm sources. The Cerro Chajnantor Atacama Telescope (CCAT, Fig. 11.4) is a planned 25-m sub-millimeter telescope, to be built close to the site of ALMA, but at a higher altitude of $\sim 5600\,\mathrm{m}$. This $\sim 600\,\mathrm{m}$ difference in altitude yields a further decrease of the water vapor column, and thus increases the sensitivity of the observatory. Equipped with powerful instruments, and a 20 arcmin field-of-view, CCAT will be able to map large portions of the sky quickly; it is expected that the CCAT will have a survey speed ~ 1000 times faster than the SCUBA-2 camera (see Sect. 1.3.1). CCAT will carry out large surveys for SMGs over a broad redshift range, and may be able to probe the earliest bursts of dusty star formation out to $z \sim 10$. CCAT will also be a powerful telescope for studying the Sunyaev–Zeldovich effect in galaxy clusters, and thus conduct cluster cosmology surveys. Last but not least, it will provide targets for observations with the ALMA interferometer, and in combination allows the joint reconstruction of compact and extended source components.

The next step in astrometry: Gaia. The ESA satellite mission Gaia, which was launched in Dec. 2013, will conduct astrometry of $\sim 10^9$ stars in the Milky Way, and provide very precise positions, proper motions and parallaxes of these stars. It is thus much more powerful than the previous astrometry satellite Hipparcos, and will provide us with a highly detailed three-dimensional map of our Galaxy, allowing precise dynamical studies, including the study of the total (dark +luminous) matter in the Milky Way and providing tests of General Relativity. Gaia will determine the distances to a large number of Cepheids, thus greatly improving the calibration of the period-luminosity relation which is one of the key elements for determining the Hubble constant in the local Universe. In addition, Gaia is expected to detect $\sim 5 \times 10^5$ AGNs.

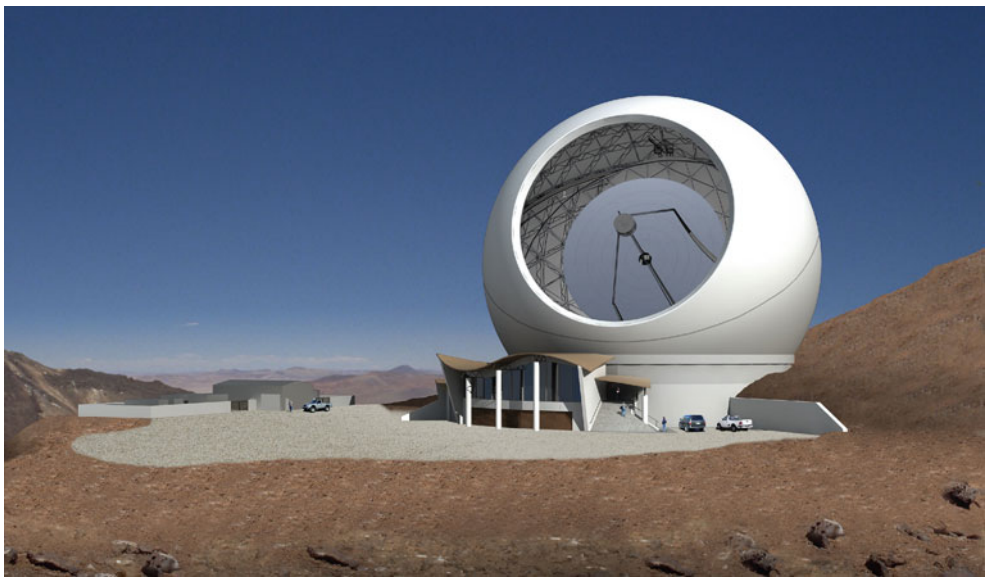


Fig. 11.4 Artist's impression of the CCAT telescope, a planned 25-m sub-millimeter telescope to be built in Chile. At an altitude of 5600 m, it will be highest altitude ground-based telescope world-wide. Credit: Cornell University & Caltech

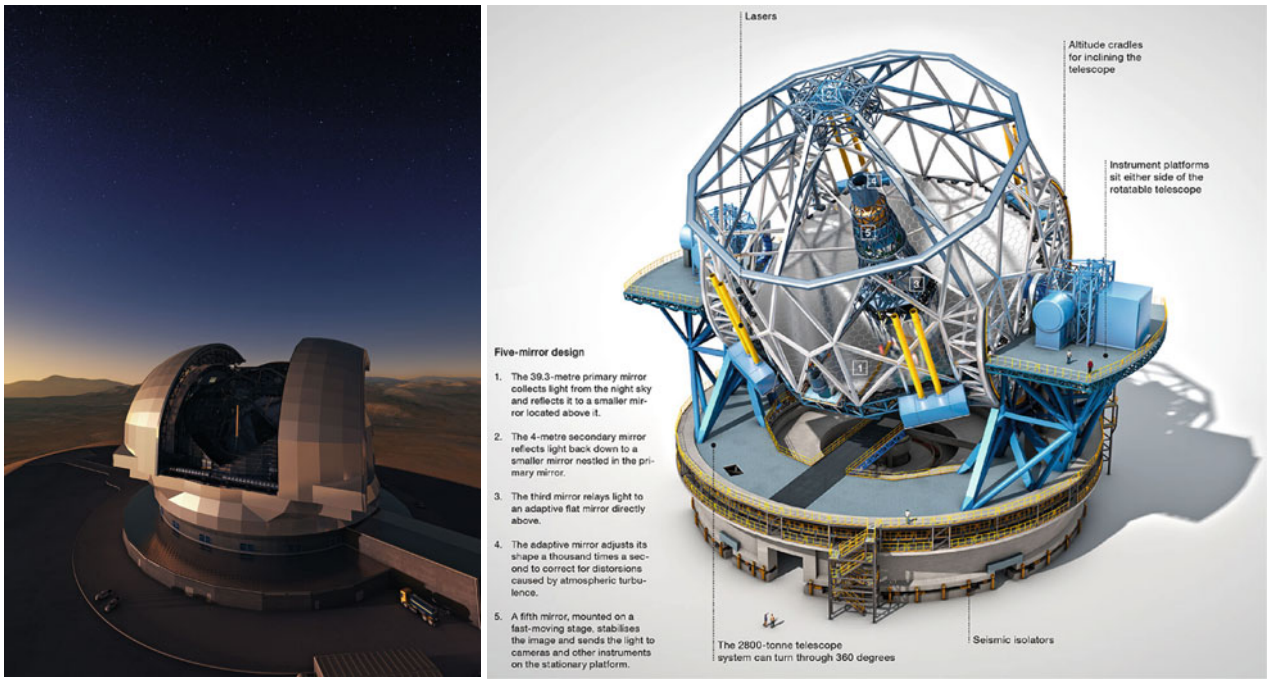


Fig. 11.5 Artist impression of the planned European Extremely Large Telescope, a 39-m telescope. Credit: European Southern Observatory

Giant near-IR/optical telescopes. A new era in observational astronomy will start with the installation of optical telescopes with an aperture of $\sim$ 30 m or larger. Three such giant telescope projects are currently in their planning stage. One of them is ESO's European Extremely Large Telescope (E-ELT, see Fig. 11.5), with a 39-m primary mirror (this mirror has about the same area as all the telescopes displayed in Fig. 11.1 together!). It will be built on Cerro Armazones, Chile, at an altitude of slightly more than 3000 m, not very far away from the Paranal, the mountain that hosts the VLT. The other two projects are the Thirty Meter Telescope (TMT), to be built on Mauna Kea in Hawaii, and the Giant Magellan Telescope (GMT) close to Gemini-South in Chile. All these telescopes will have segmented mirrors, similar to the Keck telescopes (Fig. 1.38), and sophisticated adaptive optics, to achieve the angular resolution corresponding to the diffraction limit of the telescope. For example, the E-ELT primary mirror will consist of 798 hexagonal segments, each with a size of 1.4 m. The secondary mirror will have a diameter of 4 m, which by itself would be a sizeable telescope aperture.

With their huge light-gathering power, these giant telescopes will open totally new opportunities. One expects to observe the most distant objects in the Universe, the first galaxies and AGNs which ionized the Universe. The cosmic expansion rate can be studied directly, by measuring the change of redshift of Lyα forest lines with time. High-resolution spectroscopy will enable studying the time evolution of chemical enrichment throughout cosmic history.

Optical/near-IR wide-field survey telescopes. The huge impact of the SDSS has demonstrated the versatile use of large surveys in astronomy. Several multi-band wide-field deep imaging surveys are currently ongoing, of which we mention two: The Kilo Degree Survey carried out with the VLT Survey Telescope on Paranal (see Fig. 1.39) and complemented with the near-IR VIKING survey with the VISTA telescope, covering $\sim$ 1500 deg^2 in nine bands. The Dark Energy Survey (DES) will image $\sim$ 5000 deg^2 with a newly designed camera on the Blanco Telescope, located at Cerro Tololo Inter-American Observatory (CTIO) in Chile. Both of these surveys have a wide range of science goals, including weak lensing and cosmic shear, the identification of galaxy clusters found at different wavebands (X-ray surveys, Sunyaev–Zeldovich surveys), and the large-scale distribution of galaxies and AGN, to name a few.

However, a revolution in survey astronomy will occur with the Euclid satellite (Fig. 11.6). Motivated by the empirical study of the properties of dark energy, Euclid will observe essentially the full extragalactic sky ($\sim$ 15 000 deg^2) in one broad optical band, and three near-IR bands, making use of its $\sim$ 0.5 deg^2 focal plane. The optical images will be used for a cosmic shear study and thus profit from the high resolution images obtainable from space. The fact that Euclid will be in an orbit around L2 implies that one can expect a very high stability of the telescope and instruments, which is an essential aspect for the correction of galaxy images with respect to PSF effects. The near-IR images are essential for obtaining photometric redshifts of the sources whose shapes

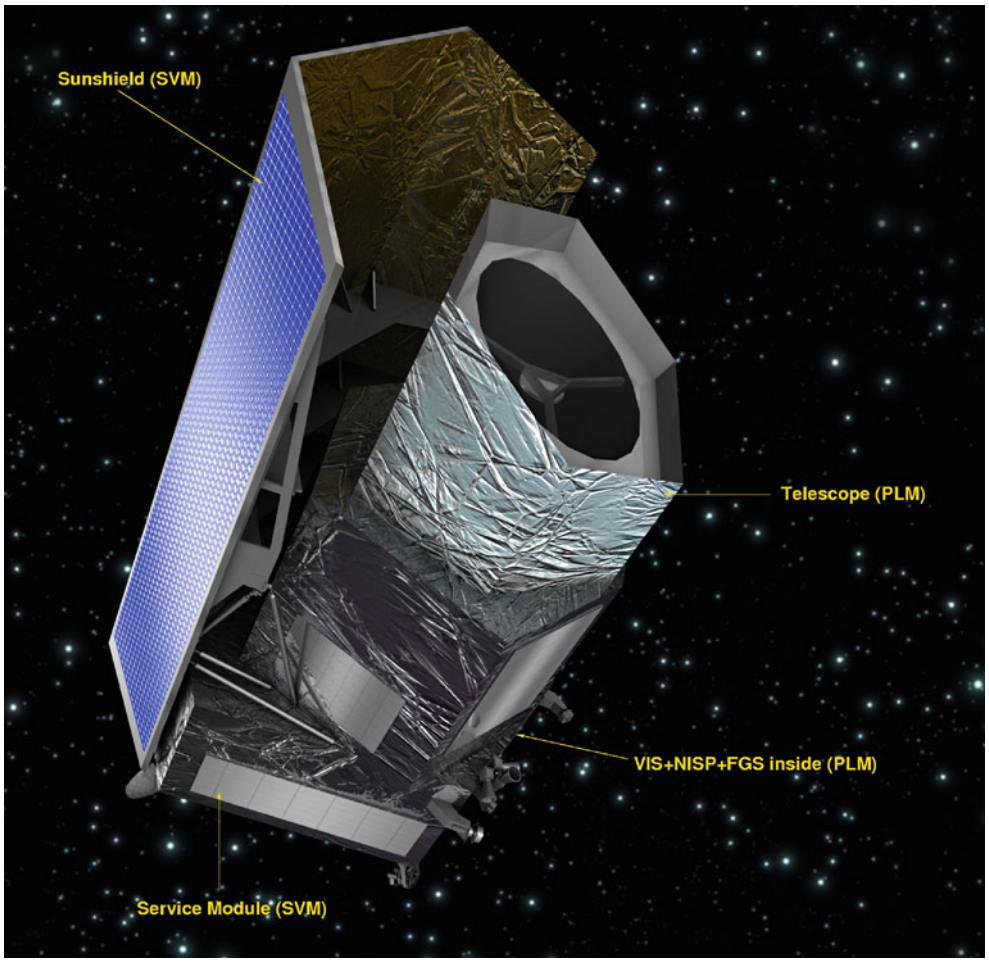


Fig. 11.6 The Euclid satellite, designed to provide a $\sim 15\,000\,\mathrm{deg}^2$ imaging survey in the optical and near-IR, together with NIR slitless spectroscopy. Credit: European Space Agency

are used in the lensing analysis. For the same reason, the Euclid data will be supplemented with images in additional optical bands, to be obtained with ground-based telescopes. Altogether, we expect to measure the shapes of some two billion galaxies. In addition, Euclid will conduct slitless spectroscopy of some 50 million galaxies over a broad range in redshift, to yield a measurement of baryonic acoustic oscillations at $z > 1$. Euclid is scheduled for launch by the end of this decade.

Whereas the primary science goal of Euclid is dark energy, the mission will leave an enormous scientific legacy. Compared to the 2MASS all-sky infrared survey (Fig. 1.52), Euclid is about 7 magnitudes deeper, and will image almost 1000 times more galaxies in the near-IR bands. Together with the optical multi-color information, this data set will indeed provide a giant step in studies of the cosmos at large.

Sometimes nice surprises happen! We have stressed many times the incredible value of the Hubble Space Telescope for astronomy. In the summer of 2012, NASA received an offer from the US National Reconnaissance Office (NRO) intelligence agency for two telescopes which are very similar to the HST, originally planned as spy satellites, but never used. These telescopes have a shorter focal length than HST, allowing a much wider field-of-view. This offer came in handy, as it fits well into the plan to build a Wide Field Infrared Space Telescope (WFIRST), operating in the near-IR. Its planned NIR camera has a field-of-view ~ 200 times larger than the WFC3 onboard HST. Its science goals are manifold, including supernova cosmology, baryonic acoustic oscillations, cosmic shear, and galaxy evolution studies.

A further ambitious project regarding optical sky surveys is the Large Synoptic Survey Telescope (LSST), an 8.4-m telescope in Chile equipped with a huge camera covering $9.6\,\mathrm{deg}^2$. It is designed to survey half the sky in six optical bands every 4 days, taking short exposures. After 10 years, the coadded image from all the short exposures will yield a sky map at a depth of $m \sim 27.5$. A projected start of operations is 2022.

Apart from the telescopes and instrumentation, these projects are extremely demanding in terms of data storage and computing power. To wit, LSST will yield a data rate of $\sim 15\,\mathrm{TB}$ per night, and this raw data has to be reduced also within a day—to avoid any serious backlock. It is expected that most of the resources for this project will be invested into computer power and software for data storage and analysis.

A new X-ray all-sky survey. Twenty-five years after Rosat carried out its all-sky survey, the German-Russian mission Spectrum-X-Gamma will carry the eROSITA (extended ROentgen Survey with an Imaging Telescope Array) instrument, an X-ray telescope operating between 0.5 and $\sim$ 10 keV. eROSITA, expected to be launched at the end of 2015, will be located at L2 and carry out eight full sky surveys. The coadded sky survey will be $\sim$ 20 deeper than the RASS, and extend to higher photon energies. The all-sky survey is expected to detect $\sim 10^5$ clusters of galaxies, many of them at redshifts larger than unity, and $\sim 3 \times 10^6$ AGNs. For the latter, the higher energy band will be particularly useful. This will not only provide a much improved statistical basis for studying the AGN population over cosmic time, but also allow us to map the large-scale structure as traced by AGNs, including the study of baryonic acoustic oscillations. Together with cluster cosmology, eROSITA will be of great value for extragalactic astronomy and cosmology.

The scientific exploitation of the eROSITA data will depend significantly on the availability of auxiliary data for the identification and redshift estimates of the detected sources. The wide-field optical and near-IR imaging surveys described before—KiDS/VIKING, DES, PanSTARRS, and later, Euclid and LSST—will be invaluable for the eROSITA analysis. Furthermore, spectrographs with large multiplexing capabilities will allow to obtain spectra for millions of sources in the e-ROSITA catalog, including the AGNs and galaxies in the detected clusters.

New windows to the Universe will be opened. The first gravitational wave antennas are already in place, and their upgraded versions, to go into operation around 2016, will probably be able to discover the signals from relatively nearby supernova explosions or mergers of compact binaries. With the Laser Interferometer Space Antenna (LISA), mergers of supermassive black holes will become detectable throughout the visible Universe, as we mentioned before. Giant neutrino detectors will open the field of neutrino astronomy and will be able, for example, to observe processes in the innermost parts of AGNs.

Theory. Parallel to these developments in telescopes and instruments, theory is progressing steadily. The continuously increasing capacity of computers available for numerical simulations is only one aspect, albeit an important one. New approaches for modeling, triggered by new observational results, are of equal importance. The close connection between theory, modeling, and observations will become increasingly important since the complexity of data requires an advanced level of modeling and simulations for their quantitative interpretation.

Data availability; virtual observatories. The huge amount of data obtained with current and future instruments is useful not only for the observers taking the data, but also for others in the astronomical community. Realizing this fact, many observatories have set up archives from which data can be retrieved—essentially by everyone. Space observatories pioneered such data archives, and a great deal of science results from the use of archival data. Examples here are the use of the HST deep fields by a large number of researchers, or the analysis of serendipitous sources in X-ray images which led to the EMSS (see Sect. 6.4.5). Together with the fact that an understanding of astronomical sources usually requires data taken over a broad range of frequencies, there is a strong motivation for the creation of *virtual observatories*: infrastructures which connect archives containing astronomical data from a large variety of instruments and which can be accessed electronically. In order for such virtual observatories to be most useful, the data structures and interfaces of the various archives need to become mutually compatible. Intensive activities in creating such virtual observatories are ongoing; they will doubtlessly play in increasingly important role in the future.

11.3 Challenges

Understanding galaxy evolution. One of the major challenges for the next few years will certainly be the investigation of the very distant Universe, studying the evolution of cosmic objects and structures at very high redshift up to the epoch of reionization. To relate the resulting insights of the distant Universe to those obtained more locally and thus to obtain a consistent view about our cosmos, major theoretical investigations will be required as well as extensive observations across the whole redshift range, using the broadest wavelength range possible. Furthermore, the new astrometry satellite Gaia will offer us the unique opportunity to study cosmology in our Milky Way. With Gaia, the aforementioned stellar streams, which were created in the past by the tidal disruption of satellite galaxies during their merging with the Milky Way, can be verified. New insights gained with Gaia will certainly also improve our understanding of other galaxies.

Dark matter. The second major challenge for the near future is the fundamental physics on which our cosmological model is based. From observations of galaxies and galaxy clusters, and also from our determinations of the cosmological parameters, we have verified the presence of dark matter. Since there seem to be no plausible astrophysical explanations for its nature, dark matter most likely consists of new kinds of elementary particles. Two different strategies to find these particles are currently being followed. First, experiments aim at directly detecting these particles, which should also be present in the immediate vicinity of the

Earth. These experiments are located in deep underground laboratories, thus shielded from cosmic rays. Several such experiments, which are an enormous technical challenge due to the sensitivity they are required to achieve, are currently running. They will obtain increasingly tighter constraints on the properties of WIMPs with respect to their mass and interaction cross-section. Such constraint will, however, depend on the mass model of the dark matter in our Galaxy. As a second approach, the Large Hadron Collider at CERN will continue to search for indications for an extension of the Standard Model of particle physics, which would predict the presence of additional particles, including the dark matter candidate.

Dark energy. Whereas at least plausible ideas exist about the nature of dark matter which can be experimentally tested in the coming years, the presence of a non-vanishing density of dark energy, as evidenced from cosmology, presents an even larger mystery for fundamental physics. Though from quantum physics we might expect a vacuum energy density to exist, its estimated energy density is tremendously larger than the cosmic dark energy density. The interpretation that dark energy is a quantum mechanical vacuum energy therefore seems highly implausible. As astrophysical cosmologists, we could take the view that vacuum energy is nothing more than a cosmological constant, as originally introduced by Einstein; this would then be an additional fundamental constant in the laws of nature. From a physical point of view, it would be much more satisfactory if the nature of dark energy could be derived from the laws of fundamental physics. The huge discrepancy between the density of dark energy and the simple estimate of the vacuum energy density clearly indicates that we are currently far from a physical understanding of dark energy. To achieve this understanding, we might well assume that a new theory must be developed which unifies quantum physics and gravity—in a manner similar to the way other 'fundamental' interactions (like electromagnetism and the weak force) have been unified within the standard model of particle physics. Deriving such a theory of quantum gravity turns out to be enormously problematic despite intensive research over several decades. However, the density of dark energy is so incredibly small that its effects can only be recognized on the largest length-scales, implying the necessity of further astronomical and cosmological experiments. Only astronomical techniques are able to probe the properties of dark energy empirically. We have outlined in Sect. 8.8 the most promising ways of studying the properties of dark energy, and the new facilities described above will allow us to make essential progress over the next decade.

Inflation. Although inflation is currently part of the standard model of cosmology, the physical processes occurring during the inflationary phase have not been understood up to now. The fact that different field-theoretical models of inflation yield very similar cosmological consequences is an asset for cosmologists: from their point-of-view, the details of inflation are not immediately relevant, as long as a phase of exponential expansion occurred. But the same fact indicates the size of the problem faced in studying the process of inflation, since different physical models yield rather similar outcomes with regard to cosmological observables. Perhaps the most promising probe of inflation is the polarization of the cosmic microwave background, since it allows us to study whether, and with what amplitude, gravitational waves were generated during inflation. Predictions of the ratio between the amplitudes of gravitational waves and that of density fluctuations are different in different physical models of inflation. A successor of the Planck satellite, in form of a mission which is able to measure the CMB polarization with sufficient accuracy for testing inflation, will probably be considered.

Baryon asymmetry. Another cosmological observation poses an additional challenge to fundamental physics. We observe baryonic matter in our Universe, but we see no signs of appreciable amounts of antimatter. If certain regions in the Universe consisted of antimatter, there would be observable radiation from matter-antimatter annihilation at the interface between the different regions. The question therefore arises, what processes caused an excess of matter over antimatter in the early Universe? We can easily quantify this asymmetry—at very early times, the abundance of protons, antiprotons and photons were all quite similar, but after proton-antiproton annihilation at $T \sim 1\,\mathrm{GeV}$, a fraction of $\sim 10^{-10}$—the current baryon-to-photo ratio—was left over. This slight asymmetry of the abundance of protons and neutrons over their antiparticles in the early Universe, often called baryogenesis, has not been explained in the framework of the standard model of particle physics. Furthermore, we would like to understand why the densities of baryons and dark matter are essentially the same, differing by a mere factor of ~ 6.

The aforementioned issues are arguably the best examples of the increasingly tight connection between cosmology and fundamental physics. Progress in either field can only be achieved by the close collaboration between theoretical and experimental particle physics and astronomy.

Sociological challenges. Astronomy has become Big Science, not only in the sense that our facilities are getting more expensive, in parallel to their increased capabilities, but also in terms of the efforts needed to conduct individual science projects. Although most research projects are still carried out in small collaborations, this is changing drastically for some of the most visible projects. One indication is the growing average number of authors per publication, which doubled between 1990 and 2006 from 3 to 6, with a clearly increasing

trend. Whereas many papers are authored by less than a handful of people, there is an increasing number of publications with long author lists: the typical H.E.S.S. publication now has $\gtrsim$200 authors, the Planck papers of order 250. One consequence of these large collaborations is that a young postdoc or PhD student may find it more difficult to find her or his name as lead author, and thus to become better known to the astrophysical community. We need to cope with this non-reversible trend; other scientific communities, like the particle physicists, have done so successfully.

Is cosmology on the right track? Finally, and perhaps too late in the opinion of some readers, we should note again that this book has assumed throughout that the physical laws, as we know them today, can be used to interpret cosmic phenomena. We have no real proof that this assumption is correct, but the successes of this approach justify this assumption in hindsight. Constraints on possible variation of physical 'constants' with time are getting increasingly tighter, providing additional justification. If this assumption had been grossly violated, there would be no reason why the values of the cosmological parameters, estimated with vastly different methods and thus employing very different physical processes, mutually agree. The price we pay for the acceptance of the standard model of cosmology, which results from this approach, is high though: the standard model implies that we accept the existence and even dominance of dark matter and dark energy in the Universe.

Not every cosmologist is willing to pay this price. For instance, M. Milgrom introduced the hypothesis that the flat rotation curves of spiral galaxies are not due to the existence of dark matter. Instead, they could arise from the possibility that the Newtonian law of gravity ceases to be valid on scales of 10 kpc—on such large scales, and the correspondingly small accelerations, the law of gravity has not been tested. Milgrom's *Modified Newtonian Dynamics (MOND)* is therefore a logically possible alternative to the postulate of dark matter on scales of galaxies. Indeed, MOND offers an explanation for the Tully–Fisher relation of spiral galaxies.

There are, however, several reasons why only a few astrophysicists follow this approach. MOND has an additional free parameter which is fixed by matching the observed rotation curves of spiral galaxies with the model, without invoking dark matter. Once this parameter is fixed, MOND cannot explain the dynamics of galaxies in clusters without needing additional matter—dark matter. Thus, the theory has just enough freedom to fix a problem on one length- (or mass-)scale, but apparently fails on different scales. We can circumvent the problem again by postulating warm dark matter, which would be able to fall into the potential wells of clusters, but not into the shallower ones of galaxies, thereby replacing one kind of dark matter (CDM) with another. In addition, the fluctuations of the cosmic microwave background radiation cannot be explained without the presence of dark matter.

In fact, the consequences of accepting MOND would be far reaching: if the law of gravity deviates from the Newtonian law, the validity of General Relativity would be questioned, since it contains the Newtonian force law as a limiting case of weak gravitational fields. General Relativity, however, forms the basis of our world models. Rejecting it as the correct description of gravity, we would lose the physical basis of our cosmological model—and thus the impressive quantitative agreement of results from vastly different observations that we described in Chap. 8. The acceptance of MOND therefore demands an even higher price than the existence of dark matter, but it is an interesting challenge to falsify MOND empirically.

This example shows that the modification of one aspect of our standard model has the consequence that the whole model is threatened: due to the large internal consistency of the standard model, modifying one aspect has a serious impact on all others. This does not mean that there cannot be other cosmological models which can provide as consistent an explanation of the relevant observational facts as our standard model does. However, an alternative explanation of a single aspect cannot be considered in isolation, but must be seen in its relation to the others. Of course, this poses a true challenge to the promoters of alternative models: whereas the overwhelming majority of cosmologists are working hard to verify and to refine the standard model and to construct the full picture of cosmic evolution, the group of researchers working on alternative models is small[2] and thus hardly able to put together a convincing and consistent model of cosmology. This fact finds its justification in the successes of the standard model, and in the agreement of observations with the predictions of this model.

We have, however, just uncovered an important sociological aspect of the scientific enterprise: there is a tendency to 'jump on the bandwagon'. This results in the vast majority of research going into one (even if the most promising) direction—and this includes scientific staff, research grants, observing time etc. The consequence is that new and unconventional ideas have a hard time getting heard. Hopefully (and in the view of this author, very likely), the bandwagon is heading in the right direction. There are historical examples to the contrary, though—we now know that Rome is not at the center of the cosmos, nor the Earth, nor the Sun, nor the Milky Way, despite long epochs when the vast majority of scientists were convinced of the veracity of these ideas.

[2]However, there has been a fairly recent increase in research activity on MOND. This was triggered mainly by the fact that after many years of research, a theory called TeVeS (for Tensor-Vector-Scalar field) was invented, containing General Relativity, MOND and Newton's law in the respective limits—though at the cost of introducing three new arbitrary functions.

The electromagnetic radiation field

A

In this appendix, we will briefly review the most important properties of a radiation field. We thereby assume that the reader has encountered these quantities already in a different context.

A.1 Parameters of the radiation field

The electromagnetic radiation field is described by the *specific intensity* I_ν, which is defined as follows. Consider a surface element of area $\mathrm{d}A$. The radiation energy which passes through this area per time interval $\mathrm{d}t$ from within a solid angle element $\mathrm{d}\omega$ around a direction described by the unit vector $\boldsymbol{n}$, with frequency in the range between ν and $\nu + \mathrm{d}\nu$, is

$$\mathrm{d}E = I_\nu \, \mathrm{d}A \, \cos\theta \, \mathrm{d}t \, \mathrm{d}\omega \, \mathrm{d}\nu \; , \tag{A.1}$$

where θ describes the angle between the direction $\boldsymbol{n}$ of the light and the normal vector of the surface element. Then, $\mathrm{d}A \, \cos\theta$ is the area projected in the direction of the infalling light. The specific intensity depends on the considered position (and, in time-dependent radiation fields, on time), the direction $\boldsymbol{n}$, and the frequency ν. With the definition (A.1), the dimension of I_ν is energy per unit area, time, solid angle, and frequency, and it is typically measured in units of $\mathrm{erg\,cm^{-2}\,s^{-1}\,ster^{-1}\,Hz^{-1}}$. The specific intensity of a cosmic source describes its surface brightness.

The *specific net flux* F_ν passing through an area element is obtained by integrating the specific intensity over all solid angles,

$$F_\nu = \int \mathrm{d}\omega \; I_\nu \, \cos\theta \; . \tag{A.2}$$

The flux that we receive from a cosmic source is defined in exactly the same way, except that cosmic sources usually subtend a very small solid angle on the sky. In calculating the flux we receive from them, we may therefore drop the factor $\cos\theta$ in (A.2); in this context, the specific flux is also denoted as S_ν. However, in this Appendix (and only here!), the notation S_ν will be reserved for another quantity. The flux is measured in units of $\mathrm{erg\,cm^{-2}\,s^{-1}\,Hz^{-1}}$. If the radiation field is isotropic, F_ν vanishes. In this case, the same amount of radiation passes through the surface element in both directions.

The *mean specific intensity* J_ν is defined as the average of I_ν over all angles,

$$J_\nu = \frac{1}{4\pi} \int \mathrm{d}\omega \; I_\nu \; , \tag{A.3}$$

so that, for an isotropic radiation field, $I_\nu = J_\nu$. The *specific energy density* u_ν is related to J_ν according to

$$u_\nu = \frac{4\pi}{c} J_\nu \; , \tag{A.4}$$

where u_ν is the energy of the radiation field per volume element and frequency interval, thus measured in $\mathrm{erg\,cm^{-3}\,Hz^{-1}}$. The total energy density of the radiation is obtained by integrating u_ν over frequency. In the same way, the intensity of the radiation is obtained by integrating the specific intensity I_ν over ν.

A.2 Radiative transfer

The specific intensity of radiation in the direction of propagation between source and observer is constant, as long as no emission or absorption processes are occurring. If s measures the length along a line-of-sight, the above statement can be formulated as

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}s} = 0 \; . \tag{A.5}$$

An immediate consequence of this equation is that the surface brightness of a source is independent of its distance. The observed flux of a source depends on its distance, because the solid angle, under which the source is observed, decreases with the square of the distance, $F_\nu \propto D^{-2}$ [see (A.2)]. However, for light propagating through a medium, emission and absorption (or scattering of light) occurring along the path over which the light travels may change the specific

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7,

intensity. These effects are described by the *equation of radiative transfer*

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}s} = -\kappa_\nu I_\nu + j_\nu \ . \tag{A.6}$$

The first term describes the absorption of radiation and states that the radiation absorbed within a length interval $\mathrm{d}s$ is proportional to the incident radiation. The factor of proportionality is the *absorption coefficient* κ_ν, which has the unit of cm^{-1}. The *emission coefficient* j_ν describes the energy that is added to the radiation field by emission processes, having a unit of $\mathrm{erg\,cm^{-3}\,s^{-1}\,Hz^{-1}\,ster^{-1}}$; hence, it is the radiation energy emitted per volume element, time interval, frequency interval, and solid angle. Both, κ_ν and j_ν depend on the nature and state (such as temperature, chemical composition) of the medium through which light propagates.

The absorption and emission coefficients both account for true absorption and emission processes, as well as the scattering of radiation. Indeed, the scattering of a photon can be considered as an absorption that is immediately followed by an emission of a photon.

The *optical depth* τ_ν along a line-of-sight is defined as the integral over the absorption coefficient,

$$\tau_\nu(s) = \int_{s_0}^{s} \mathrm{d}s'\, \kappa_\nu(s') \ , \tag{A.7}$$

where s_0 denotes a reference point on the sightline from which the optical depth is measured. Dividing (A.6) by κ_ν and using the relation $\mathrm{d}\tau_\nu = \kappa_\nu\ \mathrm{d}s$ in order to introduce the optical depth as a new variable along the light ray, the equation of radiative transfer can be written as

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}\tau_\nu} = -I_\nu + S_\nu \ , \tag{A.8}$$

where the source function

$$S_\nu = \frac{j_\nu}{\kappa_\nu} \tag{A.9}$$

is defined as the ratio of the emission and absorption coefficients. In this form, the equation of radiative transport can be formally solved; as can easily be tested by substitution, the solution is

$$\begin{aligned} I_\nu(\tau_\nu) &= I_\nu(0)\, \exp\left(-\tau_\nu\right) \\ &\quad + \int_0^{\tau_\nu} \mathrm{d}\tau'_\nu\ \exp\left(\tau'_\nu - \tau_\nu\right)\ S_\nu(\tau'_\nu) \ . \end{aligned} \tag{A.10}$$

This equation has a simple interpretation. If $I_\nu(0)$ is the incident intensity, it will have decreased by absorption to a value $I_\nu(0)\,\exp\left(-\tau_\nu\right)$ after an optical depth of τ_ν. On the other hand, energy is added to the radiation field by emission, accounted for by the τ'-integral. Only a fraction $\exp\left(\tau'_\nu - \tau_\nu\right)$ of this additional energy emitted at τ' reaches the point τ, the rest is absorbed.

In the context of (A.10), we call this a *formal* solution for the equation of radiative transport. The reason for this is based on the fact that both the absorption and the emission coefficient depend on the physical state of the matter through which radiation propagates, and in many situations this state depends on the radiation field itself. For instance, κ_ν and j_ν depend on the temperature of the matter, which in turn depends, by heating and cooling processes, on the radiation field it is exposed to. Hence, one needs to solve a coupled system of equations in general: on the one hand the equation of radiative transport, and on the other hand the equation of state for matter. In many situations, very complex problems arise from this, but we will not consider them further in the context of this book.

A.3 Blackbody radiation

For matter in thermal equilibrium, the source function S_ν is solely a function of the matter temperature,

$$S_\nu = B_\nu(T) \ , \ \text{or} \ \ j_\nu = B_\nu(T)\,\kappa_\nu \ , \tag{A.11}$$

independent of the composition of the medium (Kirchhoff's law). We will now consider radiation propagating through matter in thermal equilibrium at constant temperature T. Since in this case $S_\nu = B_\nu(T)$ is constant, the solution (A.10) can be written in the form

$$\begin{aligned} I_\nu(\tau_\nu) &= I_\nu(0)\, \exp\left(-\tau_\nu\right) \\ &\quad + B_\nu(T) \int_0^{\tau_\nu} \mathrm{d}\tau'_\nu\ \exp\left(\tau'_\nu - \tau_\nu\right) \\ &= I_\nu(0)\, \exp\left(-\tau_\nu\right) + B_\nu(T)\ \left[1 - \exp\left(-\tau_\nu\right)\right] \ . \end{aligned} \tag{A.12}$$

From this it follows that $I_\nu = B_\nu(T)$ is valid for sufficiently large optical depth τ_ν. The radiation propagating through matter which is in thermal equilibrium is described by the function $B_\nu(T)$ if the optical depth is sufficiently large, independent of the composition of the matter. A specific case of this situation can be illustrated by imagining the radiation field inside a box whose opaque walls are kept at a constant temperature T. Due to the opaqueness of the walls, their optical depth is infinite, hence the radiation field within the box is given by $I_\nu = B_\nu(T)$. This is also valid if the volume is filled with matter, as long as the latter is in thermal equilibrium at temperature T. For these reasons, this kind of radiation field is also called blackbody radiation.

The function $B_\nu(T)$ was first obtained in 1900 by Max Planck, and in his honor, it was named the *Planck function*; it reads

$$B_\nu(T) = \frac{2h_\mathrm{P}\nu^3}{c^2}\frac{1}{\mathrm{e}^{h_\mathrm{P}\nu/k_\mathrm{B}T}-1}\,, \qquad \text{(A.13)}$$

where $h_\mathrm{P} = 6.625\times10^{-27}\,\mathrm{erg\,s}$ is the *Planck constant* and $k_\mathrm{B} = 1.38\times10^{-16}\,\mathrm{erg\,K^{-1}}$ is the Boltzmann constant. The shape of the spectrum (Fig. A.1) can be derived from statistical physics. *Blackbody radiation* is defined by $I_\nu = B_\nu(T)$, and *thermal radiation* by $S_\nu = B_\nu(T)$. For large optical depths in the case of thermal radiation, the specific intensity converges to blackbody radiation. For small optical depth, the radiation field is approximated by an integral over the emissivity j_ν, which can deviate strongly from that of blackbody spectrum even in the case of a thermal source; an example is the optically thin bremsstrahlung from the hot gas in galaxy clusters (see Sect. 6.4).

The Planck function has its maximum at

$$\frac{h_\mathrm{P}\nu_\mathrm{max}}{k_\mathrm{B}T} \approx 2.82\,, \qquad \text{(A.14)}$$

i.e., the frequency of the maximum is proportional to the temperature. This property is called *Wien's law*. This law can also be written in more convenient units,

$$\nu_\mathrm{max} = 5.88\times10^{10}\,\mathrm{Hz}\,\frac{T}{1\,K}\,. \qquad \text{(A.15)}$$

The Planck function can also be formulated depending on wavelength $\lambda = c/\nu$, such that $B_\lambda(T)\,\mathrm{d}\lambda = B_\nu(T)\,\mathrm{d}\nu$,

$$B_\lambda(T) = \frac{2h_\mathrm{P}c^2/\lambda^5}{\exp\left(h_\mathrm{P}c/\lambda k_\mathrm{B}T\right)-1}\,. \qquad \text{(A.16)}$$

Two limiting cases of the Planck function are of particular interest. For low frequencies, $h_\mathrm{P}\nu \ll k_\mathrm{B}T$, one can apply the expansion of the exponential function for small arguments in (A.13). The leading-order term in this expansion then yields

$$B_\nu(T) \approx B_\nu^\mathrm{RJ}(T) = \frac{2\nu^2}{c^2}k_\mathrm{B}T\,, \qquad \text{(A.17)}$$

which is called the *Rayleigh–Jeans approximation* of the Planck function. We point out that the Rayleigh–Jeans equation does not contain the Planck constant, and this law had been known even before Planck derived his exact equation. In the other limiting case of very high frequencies, $h_\mathrm{P}\nu \gg k_\mathrm{B}T$, the exponential factor in the denominator in (A.13) becomes very much larger than unity, so that we obtain

$$B_\nu(T) \approx B_\nu^\mathrm{W}(T) = \frac{2h_\mathrm{P}\nu^3}{c^2}\mathrm{e}^{-h_\mathrm{P}\nu/k_\mathrm{B}T}\,, \qquad \text{(A.18)}$$

called the *Wien approximation* of the Planck function.

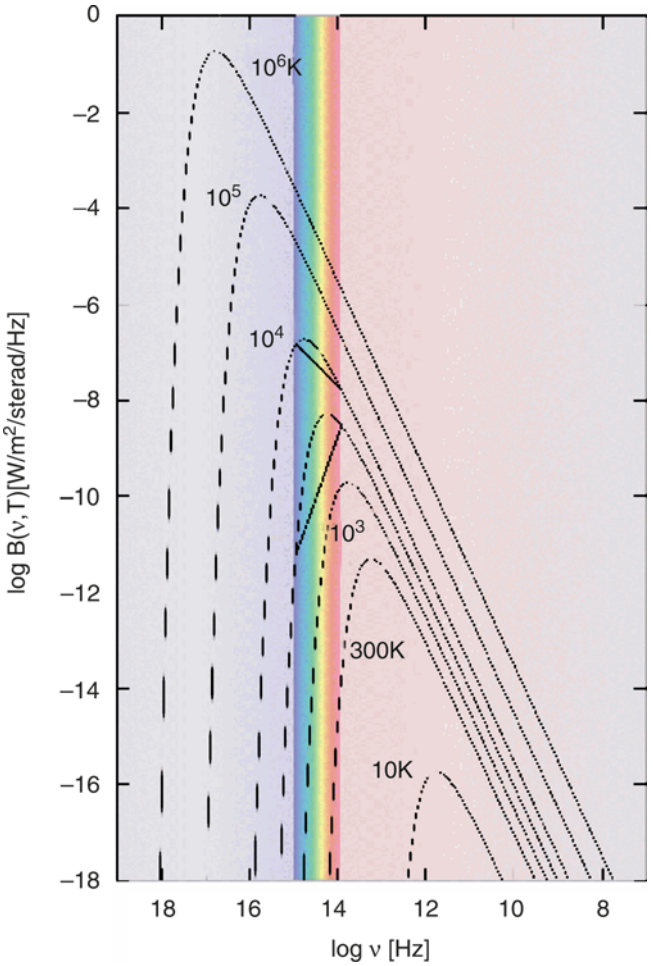


Fig. A.1 The Planck function (A.13) for different temperatures T. The plot shows $B_\nu(T)$ as a function of frequency ν, where high frequencies are plotted towards the left (thus large wavelengths towards the right). The exponentially decreasing Wien part of the spectrum is visible on the *left*, the Rayleigh–Jeans part on the *right*. The *shape* of the spectrum in the Rayleigh–Jeans part is independent of the temperature, which is determining the amplitude however. Credit: T. Kaempf & M. Altmann, Argelander-Institut für Astronomie, Universität Bonn

The energy density of blackbody radiation depends only on the temperature, of course, and is calculated by integration over the Planck function,

$$u = \frac{4\pi}{c}\int_0^\infty \mathrm{d}\nu\; B_\nu(T) = \frac{4\pi}{c}\,B(T) = a\,T^4\,, \qquad \text{(A.19)}$$

where we defined the frequency-integrated Planck function

$$B(T) = \int_0^\infty \mathrm{d}\nu\; B_\nu(T) = \frac{a\,c}{4\pi}\,T^4\,, \qquad \text{(A.20)}$$

and where the constant a has the value

$$a = \frac{8\pi^5 k_\mathrm{B}^4}{15c^3h_\mathrm{P}^3} = 7.56\times10^{-15}\,\mathrm{erg\,cm^{-3}\,K^{-4}}\,. \qquad \text{(A.21)}$$

The flux which is emitted by the surface of a blackbody per unit area is given by

$$F = \int_0^\infty \mathrm{d}\nu\ F_\nu = \pi \int_0^\infty \mathrm{d}\nu\ B_\nu(T) = \pi B(T) = \sigma_{\mathrm{SB}} T^4\ , \tag{A.22}$$

where the *Stefan–Boltzmann constant* σ_{SB} has a value of

$$\sigma_{\mathrm{SB}} = \frac{ac}{4} = \frac{2\pi^5 k_{\mathrm{B}}^4}{15 c^2 h_{\mathrm{P}}^3} = 5.67 \times 10^{-5}\,\mathrm{erg\,cm^{-2}\,K^{-4}\,s^{-1}}\ . \tag{A.23}$$

A.4 The magnitude scale

Optical astronomy was being conducted well before methods of quantitative measurements became available. The brightness of stars had been cataloged more than 2000 years ago, and their observation goes back as far as the ancient world. Stars were classified into magnitudes, assigning a magnitude of 1 to the brightest stars and higher magnitudes to the fainter ones. Since the apparent magnitude as perceived by the human eye scales roughly logarithmically with the radiation flux (which is also the case for our hearing), the magnitude scale represents a logarithmic flux scale. To link these visually determined magnitudes in historical catalogs to a quantitative measure, the magnitude system has been retained in optical astronomy, although with a precise definition. Since no historical astronomical observations have been conducted in other wavelength ranges, because these are not accessible to the unaided eye, only optical astronomy has to bear the historical burden of the magnitude system.

A.4.1 Apparent magnitude

We start with a relative system of flux measurements by considering two sources with fluxes S_1 and S_2. The *apparent magnitudes* of the two sources, m_1 and m_2, then behave according to

$$m_1 - m_2 = -2.5\,\log\left(\frac{S_1}{S_2}\right)\ ; \quad \frac{S_1}{S_2} = 10^{-0.4(m_1-m_2)}\ . \tag{A.24}$$

This means that the brighter source has a smaller apparent magnitude than the fainter one: the larger the apparent magnitude, the fainter the source.[1] The factor of 2.5 in this definition is chosen so as to yield the best agreement of the magnitude system with the visually determined magnitudes. A difference of $|\Delta m| = 1$ in this system corresponds to a flux ratio of ~ 2.51, and a flux ratio of a factor 10 or 100 corresponds to 2.5 or 5 magnitudes, respectively.

[1] Of course, this convention is confusing, particularly to someone just becoming familiar with astronomy, and it frequently causes confusion and errors, as well as problems in the communication with non-astronomers—but we have to get along with that.

A.4.2 Filters and colors

Since optical observations are performed using a combination of a filter and a detector system, and since the flux ratios depend, in general, on the choice of the filter (because the spectral energy distribution of the sources may be different), apparent magnitudes are defined for each of these filters. The most common filters are shown in Fig. A.2 and listed in Table A.1, together with their characteristic wavelengths and the widths of their transmission curves. The apparent magnitude for a filter X is defined as m_X, frequently written as X. Hence, for the B-band filter, $m_B \equiv B$.

Next, we need to specify how the magnitudes measured in different filters are related to each other, in order to define the color indices of sources. For this purpose, a particular class of stars is used, main-sequence stars of spectral type A0, of which the star Vega is an archetype. For such a star, by definition, $U = B = V = R = I = \ldots$, i.e., every color index for such a star is defined to be zero.

For a more precise definition, let $T_X(\nu)$ be the transmission curve of the filter-detector system. $T_X(\nu)$ specifies which fraction of the incoming photons with frequency ν are registered by the detector. The apparent magnitude of a source with spectral flux S_ν is then

$$m_X = -2.5\,\log\left(\frac{\int \mathrm{d}\nu\ T_X(\nu)\,S_\nu}{\int \mathrm{d}\nu\ T_X(\nu)}\right) + \mathrm{const.}\ , \tag{A.25}$$

where the constant needs to be determined from reference stars.

Another commonly used definition of magnitudes is the AB system. In contrast to the Vega magnitudes, no stellar spectral energy distribution is used as a reference here, but instead one with a constant flux at all frequencies, $S_\nu^{\mathrm{ref}} = S_\nu^{\mathrm{AB}} = 2.89 \times 10^{-21}\,\mathrm{erg\,s^{-1}\,cm^{-2}\,Hz^{-1}}$. This value has been chosen such that A0 stars like Vega have the same magnitude in the original Johnson-V-band as they have in the AB system, $m_V^{\mathrm{AB}} = m_V$. With (A.25), one obtains for the conversion between the two systems

$$\begin{aligned} m_{\mathrm{AB}\to\mathrm{Vega}} &:= m_X^{\mathrm{AB}} - m_X^{\mathrm{Vega}} \\ &= -2.5\,\log\left(\frac{\int \mathrm{d}\nu\ T_X(\nu)\,S_\nu^{\mathrm{AB}}}{\int \mathrm{d}\nu\ T_X(\nu)\,S_\nu^{\mathrm{Vega}}}\right) \end{aligned} \tag{A.26}$$

For the filters at the ESO Wide-Field Imager, which are designed to resemble the Johnson set of filters, the following

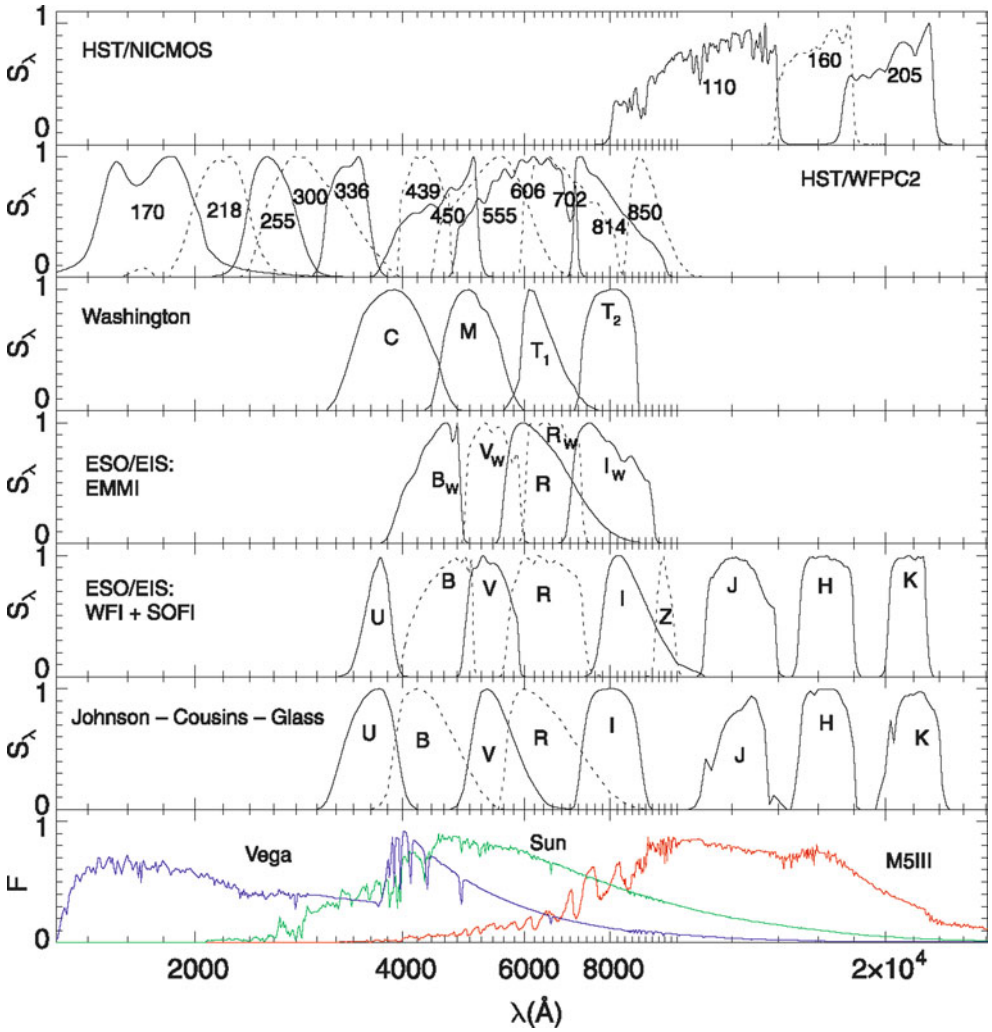


Fig. A.2 Transmission curves of various filter-detector systems. *From top to bottom*: the filters of the NICMOS camera and the WFPC2 on-board HST, the Washington filter system, the filters of the EMMI instrument at ESO's NTT, the filters of the WFI at the ESO/MPG 2.2-m telescope and those of the SOFI instrument at the NTT, and the Johnson-Cousins filters. In the *bottom* diagram, the spectra of three stars with different effective temperatures are displayed. Source: L. Girardi et al. 2002, *Theoretical isochrones in several photometric systems. I. Johnson-Cousins-Glass, HST/WFPC2, HST/NICMOS, Washington, and ESO Imaging Survey filter sets*, A&A 391, 195, p. 204, Fig. 3. ©ESO. Reproduced with permission

Table A.1 For some of the best-established filter systems—Johnson, Strömgren, and the filters of the Sloan Digital Sky Surveys—the central (more precisely, the effective) wavelengths and the widths of the filters are listed

Johnson	U	B	V	R	I	J	H	K	L	M
$\lambda_{\rm eff}$(nm)	367	436	545	638	797	1220	1630	2190	3450	4750
$\Delta\lambda$(nm)	66	94	85	160	149	213	307	39	472	460

Strömgren	u	v	b	y	β_w	β_n
$\lambda_{\rm eff}$(nm)	349	411	467	547	489	489
$\Delta\lambda$(nm)	30	19	18	23	15	3

SDSS	u′	g′	r′	i′	z′
$\lambda_{\rm eff}$(nm)	354	477	623	762	913
$\Delta\lambda$(nm)	57	139	138	152	95

prescriptions are then to be applied: $U_{\rm AB} = U_{\rm Vega} + 0.80$; $B_{\rm AB} = B_{\rm Vega} - 0.11$; $V_{\rm AB} = V_{\rm Vega}$; $R_{\rm AB} = R_{\rm Vega} + 0.19$; $I_{\rm AB} = I_{\rm Vega} + 0.59$.

A.4.3 Absolute magnitude

The apparent magnitude of a source does not in itself tell us anything about its luminosity, since for the determination of the latter we also need to know its distance D in addition to the radiative flux. Let L_ν be the specific luminosity of a source, i.e., the energy emitted per unit time and per unit frequency interval, then the flux is given by (note that from here on we switch back to the notation where S denotes the flux, which was denoted by F earlier in this appendix)

$$S_\nu = \frac{L_\nu}{4\pi D^2}\,, \tag{A.27}$$

where we implicitly assumed that the source emits isotropically. Having the apparent magnitude as a measure of S_ν (at the frequency ν defined by the filter which is applied), it is desirable to have a similar measure for L_ν, specifying the physical properties of the source itself. For this purpose, the *absolute magnitude* is introduced, denoted as M_X, where X refers to the filter under consideration. By definition, M_X

is equal to the apparent magnitude of a source if it were to be located at a distance of 10 pc from us. The absolute magnitude of a source is thus independent of its distance, in contrast to the apparent magnitude. With (A.27) we find for the relation of apparent to absolute magnitude

$$m_X - M_X = 5 \log \left(\frac{D}{1\,\mathrm{pc}} \right) - 5 \equiv \mu \,, \tag{A.28}$$

where we have defined the *distance modulus* μ in the final step. Hence, the latter is a logarithmic measure of the distance of a source: $\mu = 0$ for $D = 10\,\mathrm{pc}$, $\mu = 10$ for $D = 1\,\mathrm{kpc}$, and $\mu = 25$ for $D = 1\,\mathrm{Mpc}$. The difference between apparent and absolute magnitude is independent of the filter choice, and it equals the distance modulus if no extinction is present. In general, this difference is modified by the wavelength- (and thus filter-)dependent extinction coefficient—see Sect. 2.2.4.

A.4.4 Bolometric parameters

The total luminosity L of a source is the integral of the specific luminosity L_ν over all frequencies. Accordingly, the total flux S of a source is the frequency-integrated specific flux S_ν. The *apparent bolometric magnitude* m_bol is defined as a logarithmic measure of the total flux,

$$m_\mathrm{bol} = -2.5 \log S + \mathrm{const.} \,, \tag{A.29}$$

where here the constant is also determined from reference stars. Accordingly, the *absolute bolometric magnitude* is defined by means of the distance modulus, as in (A.28). The absolute bolometric magnitude depends on the bolometric luminosity L of a source via

$$M_\mathrm{bol} = -2.5 \log L + \mathrm{const.} \,. \tag{A.30}$$

The constant can be fixed, e.g., by using the parameters of the Sun: its apparent bolometric magnitude is $m_{\odot\mathrm{bol}} = -26.83$, and the distance of one Astronomical Unit corresponds to a distance modulus of $\mu = -31.47$. With these values, the absolute bolometric magnitude of the Sun becomes

$$M_{\odot\mathrm{bol}} = m_{\odot\mathrm{bol}} - \mu = 4.74 \,, \tag{A.31}$$

so that (A.30) can be written as

$$M_\mathrm{bol} = 4.74 - 2.5 \log \left(\frac{L}{L_\odot} \right) \,, \tag{A.32}$$

and the luminosity of the Sun is then

$$L_\odot = 3.85 \times 10^{33}\,\mathrm{erg\,s^{-1}} \,. \tag{A.33}$$

The direct relation between bolometric magnitude and luminosity of a source can hardly be exploited in practice, because the apparent bolometric magnitude (or the flux S) of a source cannot be observed in most cases. For observations of a source from the ground, only a limited window of frequencies is accessible. Nevertheless, in these cases one also likes to quantify the total luminosity of a source. For sources for which the spectrum is assumed to be known, like for many stars, the flux from observations at optical wavelengths can be extrapolated to larger and smaller wavelengths, and so m_bol can be estimated. For galaxies or AGNs, which have a much broader spectral distribution and which show much more variation between the different objects, this is not feasible. In these cases, the flux of a source in a particular frequency range is compared to the flux the Sun would have at the same distance and in the same spectral range. If M_X is the absolute magnitude of a source measured in the filter X, the X-band luminosity of this source is defined as

$$L_X = 10^{-0.4(M_X - M_{\odot X})}\, L_{\odot X} \,. \tag{A.34}$$

Thus, when speaking of, say, the 'blue luminosity of a galaxy', this is to be understood as defined in (A.34). For reference, the absolute magnitude of the Sun in optical filters is $M_{\odot\mathrm{U}} = 5.55$, $M_{\odot\mathrm{B}} = 5.45$, $M_{\odot\mathrm{V}} = 4.78$, $M_{\odot\mathrm{R}} = 4.41$, and $M_{\odot\mathrm{I}} = 4.07$.

In this appendix, we will summarize the most important properties of stars as they are required for understanding the contents of this book. Of course, this brief overview cannot replace the study of other textbooks in which the physics of stars is covered in much more detail.

B.1 The parameters of stars

To a good approximation, stars are gas spheres, in the cores of which light atomic nuclei are transformed into heavier ones (mainly hydrogen into helium) by thermonuclear processes, thereby producing energy. The external appearance of a star is predominantly characterized by its radius R and its characteristic temperature T. The properties and evolution of a star depend mainly on its mass M.

In a first approximation, the spectral energy distribution of the emission from a star can be described by a blackbody spectrum. This means that the specific intensity I_ν is given by a Planck spectrum (A.13) in this approximation. The luminosity L of a star is the energy radiated per unit time. If the spectrum of star was described by a Planck spectrum, the luminosity would depend on the temperature and on the radius according to

$$L = 4\pi R^2 \, \sigma_{\rm SB} \, T^4 \, , \tag{B.1}$$

where (A.22) was applied. However, the spectra of stars deviate from that of a blackbody (see Fig. 3.33 and the bottom panel of Fig. A.2). One defines the *effective temperature* $T_{\rm eff}$ of a star as the temperature a blackbody of the same radius would need to have to emit the same luminosity as the star, thus

$$\sigma_{\rm SB} \, T_{\rm eff}^4 \equiv \frac{L}{4\pi R^2} \, . \tag{B.2}$$

The luminosities of stars cover a huge range; the weakest are a factor $\sim 10^4$ times less luminous than the Sun, whereas the brightest emit $\sim 10^5$ times as much energy per unit time as the Sun. This big difference in luminosity is caused either by a variation in radius or by different temperatures. We know from the colors of stars that they have different temperatures: there are blue stars which are considerably hotter than the Sun, and red stars that are very much cooler. The temperature of a star can be estimated from its color. From the flux ratio at two different wavelengths or, equivalently, from the color index $X - Y \equiv m_X - m_Y$ in two filters X and Y, the temperature $T_{\rm c}$ is determined such that a blackbody at $T_{\rm c}$ would have the same color index. $T_{\rm c}$ is called the *color temperature* of a star. If the spectrum of a star was a Planck spectrum, then the equality $T_{\rm c} = T_{\rm eff}$ would hold, but in general these two temperatures differ.

B.2 Spectral class, luminosity class, and the Hertzsprung–Russell diagram

The spectra of stars can be classified according to the atomic (and, in cool stars, also molecular) spectral lines that are present. Based on the line strengths and their ratios, the Harvard sequence of stellar spectra was introduced. These spectral classes follow a sequence that is denoted by the letters O, B, A, F, G, K, M; besides these, some other spectral classes exist that will not be mentioned here. The sequence corresponds to a sequence of color temperature of stars: O stars are particularly hot, around 50 000 K, M stars very much cooler with $T_{\rm c} \sim 3500$ K. For a finer classification, each spectral class is supplemented by a number between 0 and 9. An A1 star has a spectrum very similar to that of an A0 star, whereas an A5 star has as many features in common with an A0 star as with an F0 star.

Plotting the spectral type versus the absolute magnitude for those stars for which the distance and hence the absolute magnitude can be determined, a striking distribution of stars becomes apparent in such a *Hertzsprung–Russell diagram* (HRD). Instead of the spectral class, one may also plot the color index of the stars, typically $B - V$ or $V - I$. The resulting *color-magnitude diagram* (CMD) is essentially equivalent to an HRD, but is based solely on photometric data. A different but very similar diagram plots the luminosity versus the effective temperature.

In Fig. B.1, a color-magnitude diagram is plotted, compiled from data observed by the HIPPARCOS satellite. Instead of filling the two-dimensional parameter space

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7,

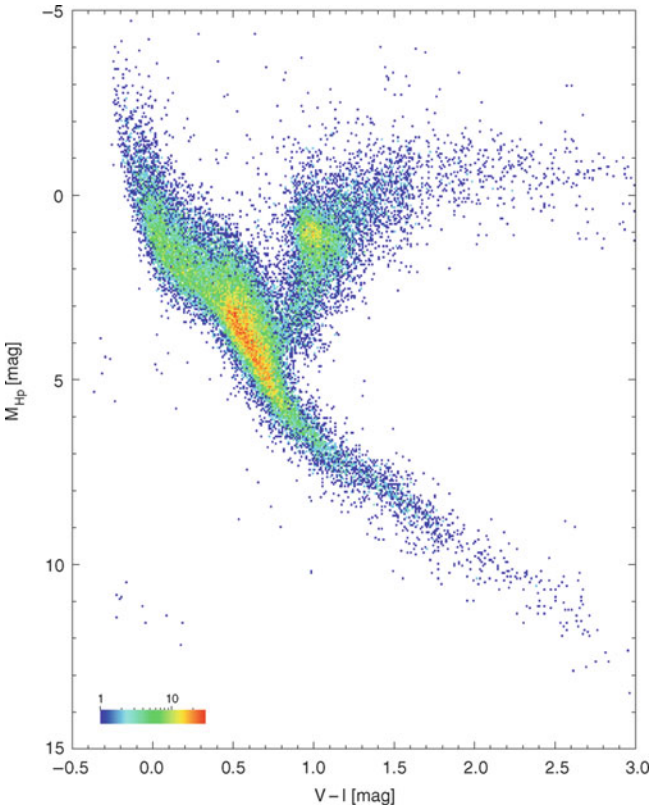


Fig. B.1 Color-magnitude diagram for 41 453 individual stars, whose parallaxes were determined by the Hipparcos satellite with an accuracy of better than 20 %. Since the stars shown here are subject to unavoidable strong selection effects favoring nearby and luminous stars, the relative number density of stars is not representative of their true abundance. In particular, the lower main sequence is much more densely populated than is visible in this diagram. Credit: European Space Agency, Web page of the Hipparcos project

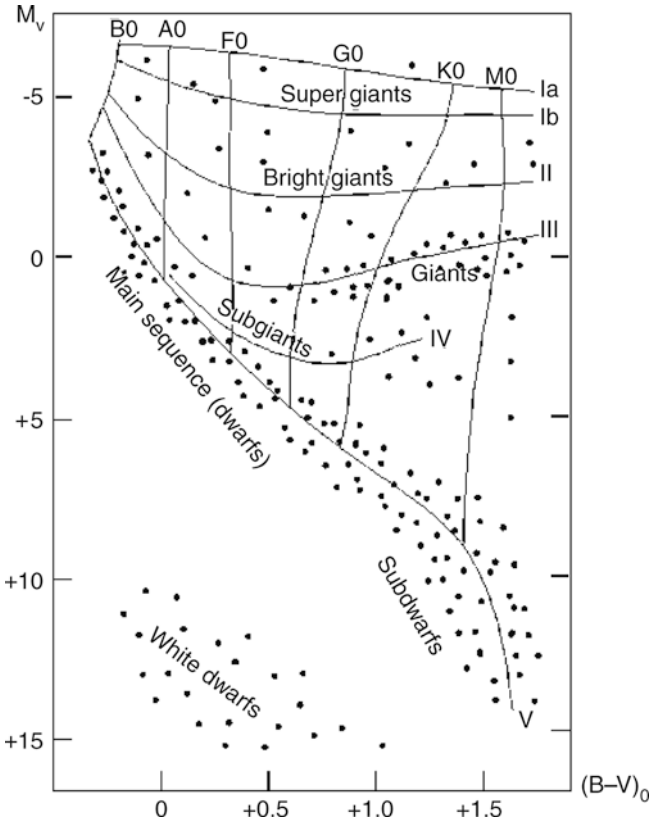


Fig. B.2 Schematic color-magnitude diagram in which the spectral types and luminosity classes are indicated. Source: http://de.wikipedia.org

rather uniformly, characteristic regions exist in such color-magnitude diagrams in which nearly all stars are located. Most stars can be found in a thin band called the *main sequence*. It extends from early spectral types (O, B) with high luminosities ('top left') down to late spectral types (K, M) with low luminosities ('bottom right'). Branching off from this main sequence towards the 'top right' is the domain of red giants, and below the main sequence, at early spectral types and very much lower luminosities than on the main sequence itself, we have the domain of white dwarfs. The fact that most stars are arranged along a one-dimensional sequence—the main sequence—is probably one of the most important discoveries in astronomy, because it tells us that the properties of stars are determined basically by a single parameter: their mass.

Since stars exist which have, for the same spectral type and hence the same color temperature (and roughly the same effective temperature), very different luminosities, we can deduce immediately that these stars have different radii, as can be read from (B.2). Therefore, stars on the red giant branch, with their much higher luminosities compared to main-sequence stars of the same spectral class, have a much larger radius than the corresponding main-sequence stars. This size effect is also observed spectroscopically: the gravitational acceleration on the surface of a star (surface gravity) is

$$g = \frac{GM}{R^2} \,. \tag{B.3}$$

We know from models of stellar atmospheres that the width of spectral lines depends on the gravitational acceleration on the star's surface: the lower the surface gravity, the narrower the stellar absorption lines. Hence, a relation exists between the line width and the stellar radius. Since the radius of a star—for a fixed spectral type or effective temperature—specifies the luminosity, this luminosity can be derived from the width of the lines. In order to calibrate this relation, stars of known distance are required.

Based on the width of spectral lines, stars are classified into *luminosity classes*: stars of luminosity class I are called supergiants, those of luminosity class III are giants, main-sequence stars are denoted as dwarfs and belong to luminosity class V; in addition, the classification can be further broken down into bright giants (II), subgiants (IV), and subdwarfs (VI). Any star in the Hertzsprung–Russell

diagram can be assigned a luminosity class and a spectral class (Fig. B.2). The Sun is a G2 star of luminosity class V.

If the distance of a star, and thus its luminosity, is known, and if in addition its surface gravity can be derived from the line width, we obtain the stellar mass from these parameters. By doing so, it turns out that for main-sequence stars the luminosity is a steep function of the stellar mass, approximately described by

$$\frac{L}{L_\odot} \approx \left(\frac{M}{M_\odot}\right)^{3.5} . \tag{B.4}$$

Therefore, a main-sequence star of $M = 10 M_\odot$ is ~ 3000 times more luminous than our Sun.

B.3 Structure and evolution of stars

To a very good approximation, stars are spherically symmetric. Therefore, the structure of a star is described by the radial profile of the parameters of its stellar plasma. These are density, pressure, temperature, and chemical composition of the matter. During almost the full lifetime of a star, the plasma is in hydrostatic equilibrium, so that pressure forces and gravitational forces are of equal magnitude and directed in opposite directions, so as to balance each other.

The density and temperature are sufficiently high in the center of a star that thermonuclear reactions are ignited. In main-sequence stars, hydrogen is fused into helium, thus four protons are combined into one ^{4}He nucleus. For every helium nucleus that is produced this way, 26.73 MeV of energy are released. Part of this energy is emitted in the form of neutrinos which can escape unobstructed from the star due to their very low cross section.[1] The energy production rate is approximately proportional to T^4 for temperatures below about 15×10^6 K, at which the reaction follows the so-called pp-chain. At higher temperatures, another reaction chain starts to contribute, the so-called CNO cycle, with an energy production rate which is much more strongly dependent on temperature—roughly proportional to T^{20}.

The energy generated in the interior of a star is transported outwards, where it is then released in the form of electromagnetic radiation. This energy transport may take place in two different ways: first, by radiation transport, and second, it can be transported by macroscopic flows of the stellar plasma. This second mechanism of energy transport is called convection; here, hot elements of the gas rise upwards, driven by buoyancy, and at the same time cool ones sink downwards. The process is similar to that observed in heating water on a stove. Which of the two processes is responsible for the energy transport depends on the temperature profile inside the star. The intervals in a star's radius in which energy transport takes place via convection are called convection zones. Since in convection zones stellar material is subject to mixing, the chemical composition is homogeneous there. In particular, chemical elements produced by nuclear fusion are transported through the star by convection.

Stars begin their lives with a homogeneous chemical composition, resulting from the composition of the molecular cloud out of which they are formed. If their mass exceeds about $0.08 M_\odot$, the temperature and pressure in their core are sufficient to ignite the fusion of hydrogen into helium. Gas spheres with a mass below $\sim 0.08 M_\odot$ will not satisfy these conditions, hence these objects—they are called brown dwarfs—are not stars in a proper sense.[2] At the onset of nuclear fusion, the star is located on the zero-age main sequence (ZAMS) in the HRD (see Fig. B.3). The energy production by fusion of hydrogen into helium alters the chemical composition in the stellar interior; the abundance of hydrogen decreases by the same rate as the abundance of helium increases. As a consequence, the duration of this phase of central hydrogen burning is limited. As a rough estimate, the conditions in a star will change noticeably when about 10 % of its hydrogen is used up. Based on this criterion, the lifetime of a star on the main sequence can now be estimated. The total energy produced in this phase can be written as

$$E_{\mathrm{MS}} = 0.1 \times M\, c^2 \times 0.007 , \tag{B.5}$$

where $M\, c^2$ is the rest-mass energy of the star, of which a fraction of 0.1 is fused into helium, which is supposed to occur with an efficiency of 0.007. Phrased differently, in the fusion of four protons into one helium nucleus, an energy of $\sim 0.007 \times 4 m_{\mathrm{p}} c^2$ is generated, with m_{p} denoting the proton mass. In particular, (B.5) states that the total energy produced during this main-sequence phase is proportional to the mass of the star. In addition, we know from (B.4) that the luminosity is a steep function of the stellar mass. The lifetime of a star on the main sequence can then be estimated by equating the available energy E_{MS} with the product of luminosity and lifetime. This yields

$$t_{\mathrm{MS}} = \frac{E_{\mathrm{MS}}}{L} \approx 8 \times 10^9 \frac{M/M_\odot}{L/L_\odot}\ \mathrm{yr} \approx 8 \times 10^9 \left(\frac{M}{M_\odot}\right)^{-2.5} \mathrm{yr} . \tag{B.6}$$

Using this argument, we observe that stars of higher mass conclude their lives on the main sequence much faster than

[1] The detection of neutrinos from the Sun in terrestrial detectors was the final proof for the energy production mechanism being nuclear fusion.

[2] If the mass of a brown dwarf exceeds $\sim 0.013 M_\odot$, the central density and temperature are high enough to enable the fusion of deuterium (heavy hydrogen) into helium. However, the abundance of deuterium is smaller by several orders of magnitude than that of normal hydrogen, rendering the fuel reservoir of a brown dwarf very small.

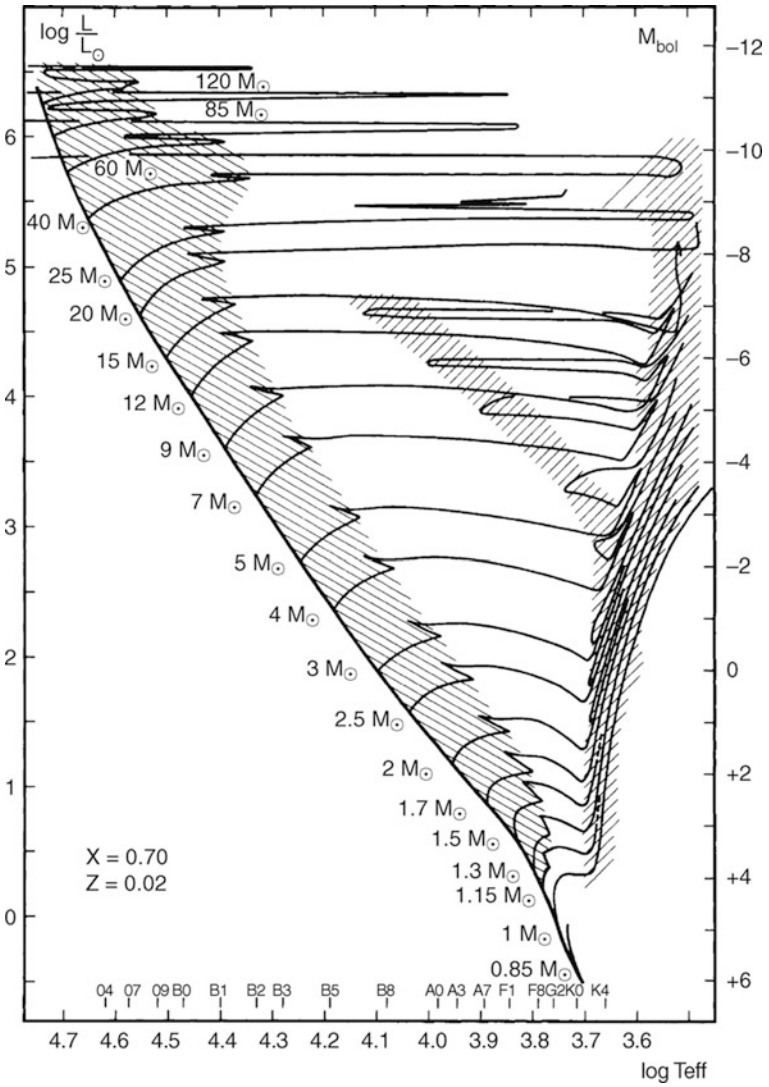


Fig. B.3 Theoretical temperature-luminosity diagram of stars. The *solid curve* is the zero age main sequence (ZAMS), on which stars ignite the burning of hydrogen in their cores. The evolutionary tracks of these stars are indicated by the various *lines* which are labeled with the stellar mass. The *hatched areas* mark phases in which the evolution proceeds only slowly, so that many stars are observed to be in these areas. Source: A. Maeder & G. Meynet 1989, *Grids of evolutionary models from 0.85 to 120 solar masses - Observational tests and the mass limits*, A&A 210, 155, p. 166, Fig. 15. ©ESO. Reproduced with permission

stars of lower mass. The Sun will remain on the main sequence for about eight to ten billion years, with about half of this time being over already. In comparison, very luminous stars, like O and B stars, will have a lifetime on the main sequence of only a few million years before they have exhausted their hydrogen fuel.

In the course of their evolution on the main sequence, stars move away only slightly from the ZAMS in the HRD, towards somewhat higher luminosities and lower effective temperatures. In addition, the massive stars in particular can lose part of their initial mass by stellar winds. The evolution after the main-sequence phase depends on the stellar mass. Stars of very low mass, $M \lesssim 0.7 M_\odot$, have a lifetime on the main sequence which is longer than the age of the Universe, therefore they cannot have moved away from the main sequence yet.

For massive stars, $M \gtrsim 2.5 M_\odot$, central hydrogen burning is first followed by a relatively brief phase in which the fusion of hydrogen into helium takes place in a shell outside the center of the star. During this phase, the star quickly moves to the 'right' in the HRD, towards lower temperatures, and thereby expands strongly. After this phase, the density and temperature in the center rise so much as to ignite the fusion of helium into carbon. A central helium-burning zone will then establish itself, in addition to the source in the shell where hydrogen is burned. As soon as the helium in the core has been exhausted, a second shell source will form fusing helium. In this stage, the star will become a red giant or supergiant, ejecting part of its mass into the interstellar medium in the form of stellar winds. Its subsequent evolutionary path depends on this mass loss. A star with an initial mass $M \lesssim 8 M_\odot$ will evolve into a white dwarf, which will be discussed further below.

For stars with initial mass $M \lesssim 2.5 M_\odot$, the helium burning in the core occurs explosively, in a so-called helium flash. A large fraction of the stellar mass is ejected in the course of

this flash, after which a new stable equilibrium configuration is established, with a helium shell source burning beside the hydrogen-burning shell. Expanding its radius, the star will evolve into a red giant or supergiant and move along the asymptotic giant branch (AGB) in the HRD.

The configuration in the helium shell source is unstable, so that its burning will occur in the form of pulses. After some time, this will lead to the ejection of the outer envelope which then becomes visible as a *planetary nebula*. The remaining central star moves to the left in the HRD, i.e., its temperature rises considerably (to more than 10^5 K). Finally, its radius gets smaller by several orders of magnitude, so that the stars move downwards in the HRD, thereby slightly reducing its temperature: a white dwarf is born, with a mass of about $0.6 M_\odot$ and a radius roughly corresponding to that of the Earth.

If the initial mass of the star is $\gtrsim 8\, M_\odot$, the temperature and density at its center become so large that carbon can also be fused. Subsequent stellar evolution towards a core-collapse supernova is described in Sect. 2.3.2.

The individual phases of stellar evolution have very different time-scales. As a consequence, stars pass through certain regions in the HRD very quickly, and for this reason stars at those evolutionary stages are never or only rarely found in the HRD. By contrast, long-lasting evolutionary stages like the main sequence or the red giant branch exist, with those regions in an observed HRD being populated by numerous stars.

Units and constants

C

In this book, we consistently used, besides astronomical units, the Gaussian cgs system of units, with lengths measured in cm, masses in g, and energies in erg. This is the commonly used system of units in astronomy. In these units, the speed of light is $c = 2.998 \times 10^{10}\,\mathrm{cm\,s^{-1}}$, the masses of protons, neutrons, and electrons are $m_\mathrm{p} = 1.673 \times 10^{-24}\,\mathrm{g}$, $m_\mathrm{n} = 1.675 \times 10^{-24}\,\mathrm{g}$, and $m_\mathrm{e} = 9.109 \times 10^{-28}\,\mathrm{g}$, respectively.

Frequently used units of length in astronomy include the Astronomical Unit, thus the average separation between the Earth and the Sun, where $1\,\mathrm{AU} = 1.496 \times 10^{13}\,\mathrm{cm}$, and the parsec (see Sect. 2.2.1 for the definition), $1\,\mathrm{pc} = 3.086 \times 10^{18}\,\mathrm{cm}$. A year has $1\,\mathrm{yr} = 3.156 \times 10^{7}\,\mathrm{s}$. In addition, masses are typically specified in Solar masses, $1M_\odot = 1.989 \times 10^{33}\,\mathrm{g}$, and the bolometric luminosity of the Sun is $L_\odot = 3.846 \times 10^{33}\,\mathrm{erg\,s^{-1}}$.

In cgs units, the value of the elementary charge is $e = 4.803 \times 10^{-10}\,\mathrm{cm^{3/2}\,g^{1/2}\,s^{-1}}$, and the unit of the magnetic field strength is one Gauss, where $1\,\mathrm{G} = 1\,\mathrm{g^{1/2}\,cm^{-1/2}\,s^{-1}} = 1\,\mathrm{erg^{1/2}\,cm^{-3/2}}$. One of the very convenient properties of cgs units is that the energy density of the magnetic field in these units is given by $\rho_B = B^2/(8\pi)$—the reader may check that the units of this equation are consistent.

X-ray astronomers measure energies in electron Volts, where $1\mathrm{eV} = 1.602 \times 10^{-12}\,\mathrm{erg}$. Temperatures can also be measured in units of energy, because $k_\mathrm{B}T$ has the dimension of energy. They are related according to $1\,\mathrm{eV} = 1.161 \times 10^4\,k_\mathrm{B}\,\mathrm{K}$. Since we always use the Boltzmann constant k_B in combination with a temperature, its actual value is never needed. The same holds for Newton's constant of gravity which is always used in combination with a mass. Here one has

$$\frac{G\,M_\odot}{c^2} = 1.495 \times 10^5\,\mathrm{cm}\,, \tag{C.1}$$

which can also be written in the form

$$G = 4.35 \times 10^{-3}\,\frac{\mathrm{pc}}{M_\odot}\left(\frac{\mathrm{km}}{s}\right)^2\,. \tag{C.2}$$

The frequency of a photon is linked to its energy according to $h_\mathrm{P}\nu = E$, and we have the relation $1\mathrm{eV}\,h_\mathrm{P}^{-1} = 2.418 \times 10^{14}\,\mathrm{s^{-1}} = 2.418 \times 10^{14}\,\mathrm{Hz}$. Accordingly, we can write the wavelength $\lambda = c/\nu = h_\mathrm{P}c/E$ in the form

$$\frac{h_\mathrm{P}c}{1\,\mathrm{eV}} = 1.2400 \times 10^{-4}\,\mathrm{cm} = 12\,400\,\text{Å}\,.$$

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7,

Recommended literature

D

In the following, we will give some recommendations for further study of the literature on astrophysics. For readers who have been in touch with astronomy only occasionally until now, the general textbooks may be of particular interest. The choice of literature presented here is a very subjective one which represents the preferences of the author, and of course it represents only a small selection of the many astronomy texts available.

D.1 General textbooks

There exist a large selection of general textbooks in astronomy which present an overview of the field at a non-technical level. A classic one (though by now becoming of age) and an excellent presentation of astronomy is

- F. Shu: *The Physical Universe: An Introduction to Astronomy*, University Science Books, Sausalito, 1982.

Turning to more technical books, at about the level of the present text, my favorite is

- B.W. Carroll & D.A. Ostlie: *An Introduction to Modern Astrophysics*, Addison-Wesley, Reading, 2006;

its $\sim$ 1400 pages cover the whole range of astronomy. The texts

- M.L. Kutner: *Astronomy: A physical perspective*, Cambridge University Press, Cambridge, 2003,
- J.O. Bennett, M.O. Donahue, N. Schneider & M. Voit: *The Cosmic Perspective*, Addison-Wesley, 2013,

also cover the whole field of astronomy. A text with a particular focus on stellar and Galactic astronomy is

- A. Unsöld & B. Baschek: *The New Cosmos*, Springer-Verlag, Berlin, 2002;

The book

- M.H. Jones & R.J.A. Lambourne: *An Introduction to Galaxies and Cosmology*, Cambridge University Press, Cambridge, 2003

covers the topics described in this book and is also highly recommended; it is less technical than the present text.

D.2 More specific literature

More specific monographs and textbooks exist for the individual topics covered in this book, some of which shall be suggested below. Again, this is just a brief selection. The technical level varies substantially among these books and, in general, exceeds that of the present text.

Astrophysical processes:

- M. Harwit: *Astrophysical Concepts*, Springer, New-York, 2006,
- G.B. Rybicki & A.P. Lightman: *Radiative Processes in Astrophysics*, John Wiley & Sons, New York, 1979,
- F. Shu: *The Physics of Astrophysics I: Radiation*, University Science Books, Mill Valley, 1991,
- F. Shu: *The Physics of Astrophysics II: Gas Dynamics*, University Science Books, Mill Valley, 1991,
- S.N. Shore: *The Tapestry of Modern Astrophysics*, Wiley-VCH, Berlin, 2002,
- D.E. Osterbrock & G.J. Ferland: *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei*, University Science Books, Mill Valley, 2005.

Furthermore, there is a three-volume set of books,

- T. Padmanabhan: *Theoretical Astrophysics: I. Astrophysical Processes. II. Stars and Stellar Systems. III. Galaxies and Cosmology*, Cambridge University Press, Cambridge, 2000.

Galaxies and gravitational lenses:

- L.S. Sparke & J.S. Gallagher: *Galaxies in the Universe: An Introduction*, Cambridge University Press, Cambridge, 2007,
- J. Binney & M. Merrifield: *Galactic Astronomy*, Princeton University Press, Princeton, 1998,
- J. Binney & S. Tremaine: *Galactic dynamics*, Princeton University Press, Princeton, 2008,
- R.C. Kennicutt, Jr., F. Schweizer & J.E. Barnes: *Galaxies: Interactions and Induced Star Formation*, Saas-Fee Advanced Course 26, Springer-Verlag, Berlin, 1998,

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7,

- B.E.J. Pagel: *Nucleosynthesis and Chemical Evolution of Galaxies*, Cambridge University Press, Cambridge, 2009,
- F. Combes, P. Boissé, A. Mazure & A. Blanchard: *Galaxies and Cosmology*, Springer-Verlag, 2004,
- P. Schneider, J. Ehlers & E.E. Falco: *Gravitational Lenses*, Springer-Verlag, New York, 1992.
- P. Schneider, C.S. Kochanek & J. Wambsganss: *Gravitational Lensing: Strong, Weak & Micro*, Saas-Fee Advanced Course 33, G. Meylan, P. Jetzer & P. North (Eds.), Springer-Verlag, Berlin, 2006.

Active galaxies:

- B.M. Peterson: *An Introduction to Active Galactic Nuclei*, Cambridge University Press, Cambridge, 1997,
- R.D. Blandford, H. Netzer & L. Woltjer: *Active Galactic Nuclei*, Saas-Fee Advanced Course 20, Springer-Verlag, 1990,
- J. Krolik: *Active Galactic Nuclei*, Princeton University Press, Princeton, 1999,
- J. Frank, A. King & D. Raine: *Accretion Power in Astrophysics*, Cambridge University Press, Cambridge, 2012.

Cosmology:

- M.S. Longair: *Galaxy Formation*, Springer-Verlag, Berlin, 2008,
- J.A. Peacock: *Cosmological Physics*, Cambridge University Press, Cambridge, 1999,
- T. Padmanabhan: *Structure formation in the Universe*, Cambridge University Press, Cambridge, 1993,
- E.W. Kolb and M.S. Turner: *The Early Universe*, Addison Wesley, 1990,
- S. Dodelson: *Modern Cosmology*, Academic Press, San Diego, 2003,
- P.J.E. Peebles: *Principles of Physical Cosmology*, Princeton University Press, Princeton, 1993,
- G. Börner: *The Early Universe*, Springer-Verlag, Berlin, 2003,
- D.H. Lyth & A.R. Liddle: *The Primordial Density Perturbation: Cosmology, Inflation and the Origin of Structure*, Cambridge University Press, Cambridge, 2009.

D.3 Review articles, current literature, and journals

Besides textbooks and monographs, review articles on specific topics are particularly useful for getting extended information about a special field. A number of journals and series exist in which excellent review articles are published. Among these are *Annual Reviews of Astronomy and Astrophysics* (ARA&A) and *Astronomy & Astrophysics Reviews* (A&AR), both publishing astronomical articles only. In *Physics Reports* (Phys. Rep.) and *Reviews of Modern Physics* (RMP), astronomical review articles are also frequently found. Such articles are also published in the lecture notes of international summer/winter schools and in the proceedings of conferences; of particular note are the Lecture Notes of the *Saas-Fee Advanced Courses*. A very useful archive containing review articles on the topics covered in this book is the Knowledgebase for Extragalactic Astronomy and Cosmology, which can be found at

http://nedwww.ipac.caltech.edu/level5.

Original astronomical research articles are published in the relevant scientific journals; most of the figures presented in this book are taken from these journals. The most important of them are Astronomy & Astrophysics (A&A), The Astronomical Journal (AJ), The Astrophysical Journal (ApJ), Monthly Notices of the Royal Astronomical Society (MNRAS), and Publications of the Astronomical Society of the Pacific (PASP). Besides these, a number of smaller, regional, or more specialized journals exist, such as Astronomische Nachrichten (AN), Acta Astronomica (AcA), or Publications of the Astronomical Society of Japan (PASJ). Some astronomical articles are also published in the journals Nature and Science. The Physical Review D and Physical Review Letters contain an increasing number of papers on astrophysical cosmology.

Since many years now, the primary source of astronomical information by far is the electronic archive

http://arxiv.org/archive/astro-ph

which is freely accessible. This archive, now hosted at Cornell University and supported by the Simons Foundation and the Allianz der deutschen Wissenschaftsorganisationen, koordiniert durch TIB, MPG und HGF, has existed since 1992, with an increasing number of articles being stored at this location. In particular, in the fields of extragalactic astronomy and cosmology, almost all articles that are published in the major journals can be found in this archive. A large number of review articles and coference proceedings are also available here.

The SAO/NASA Astrophysics Data System (ADS) is a Digital Library portal for Astronomy and Physics, operated by the Smithsonian Astrophysical Observatory (SAO) under a NASA grant. It can be accessed via the Internet at, e.g.,

http://cdsads.u-strasbg.fr/abstract_service.html,
http://adsabs.harvard.edu/abstract_service.html,

and it provides the best access to astronomical literature. Besides tools to search for authors and keywords, ADS offers also direct access to older articles that have been scanned. The access to more recent articles, and to all articles in some other journals, is restricted to IP addresses that are associated with a subscription for the respective journals—but ADS also contains a link to the article in the arXiv (if it has been posted there), so also these articles are accessible.

Acronyms used

E

In this Appendix, we compile some of the acronyms that are used, and references to the sections in which these acronyms have been introduces or explained.

2dF(GRS) – 2 degree Field Galaxy Redshift Survey (Sect. 8.1.2)
2MASS – Two Micron All Sky Survey (Sect. 1.4)
AAS – American Astronomical Society
AAT – Anglo-Australian Telescope (Sect. 1.3.3)
ACBAR – Arcminute Cosmology Bolometer Array – Receiver (Sect. 8.6.5)
ACO – Abell, Corwin & Olowin (catalogue of clusters of galaxies, Sect. 6.2.1)
ACS – Advanced Camera for Surveys (HST instrument—Sect. 1.3.3)
ACT – Atacama Cosmology Telescope (Sect. 8.6.6)
ADAF – Advection-Dominated Accretion Flow (Sect. 5.3.2)
AGB – Asymptotic Giant Branch (Sect. 3.5.2)
AGN – Active Galactic Nucleus (Sect. 5)
ALMA – Atacama Large Millimeter/sub-millimeter Array (Sect. 1.3.1)
AMR – Adaptive Mesh Refinement (Sect. 10.6.1)
APEX – Atacama Pathfinder Experiment (Sect. 1.3.1)
ASP – Astronomical Society of the Pacific
AU – Astronomical Unit
BAL – Broad Absorption Line (-Quasar, Sect. 5.7)
BAOs – Baryonic Acoustic Oscillations (Sect. 7.4.3)
BATSE – Burst And Transient Source Experiment (CGRO instrument, Sect. 9.7)
BBB – Big Blue Bump (Sect. 5.4.1)
BBN – Big Bang Nucleosynthesis (Sect. 4.4.5)
BCD – Blue Compact Dwarf (Sect. 3.2.1)
BCG – Brightest Cluster Galaxy (Sect. 6.2.4)
BH – Black Hole (Sect. 5.3.5)
BLR – Broad Line Region (Sect. 5.4.2)
BLRG – Broad Line Radio Galaxy (Sect. 5.2.4)
BOOMERANG – Balloon Observations Of Millimetric Extragalactic Radiation and Geophysics (Sect. 8.6.4)
BTP diagram – Baldwin–Phillips–Terlevich diagram (Sect. 5.4.3)
CBI – Cosmic Background Imager (Sect. 8.6.5)
CCAT – Cerro Chajnantor Atacama Telescope (Chap. 11)
CCD – Charge Coupled Device
CDF – Chandra Deep Field (Sect. 9.2.1)
CDM – Cold Dark Matter (Sect. 7.4.1)
CERN – Conseil European pour la Recherché Nucleaire
CfA – Harvard-Smithsonian Center for Astrophysics
CFHT – Canada-France-Hawaii Telescope (Sect. 1.3.3)
CFRS – Canada-France Redshift Survey (Sect. 8.1.2)
COSMOS – Cosmological Evolution Survey (Sect. 9.2.1)
CTIO – Cerro Tololo Inter-American Observatory
CXB – Cosmic X-ray Background (Sect. 9.5.3)
DASI – Degree Angular Scale Interferometer (Sect. 8.6.4)
DES – Dark Energy Survey (Chap. 11)
DIRBE – Diffuse Infrared Background Experiment (instrument onboard COBE)
DLA system – Damped Lyman Alpha system (Sect. 9.3.4)
DRG – Distant Red Galaxy (Sect 9.1.3)
dSph – dwarf Spheroidal (Sect. 3.2.1)
DSS – Digital Sky Survey (Sect. 1.4)
ECDFS – Extended Chandra Deep Field South (Sect. 9.3.3)
EdS – Einstein–de Sitter (Sect. 4.3.4)
E-ELT – European Extremely Large Telescope (Chap. 11)
EMSS – Extended Medium Sensitivity Survey (Sect. 6.4.5)

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7,

EPIC – European Photon Imaging Camera (XMM-Newton instrument)
ERO – Extremely Red Object (Sect. 9.3.2)
EROS – Expérience pour la Recherche d'Objets Sombres (microlensing collaboration, Sect. 2.5)
ESA – European Space Agency
ESO – European Southern Observatory
FFT – Fast Fourier Transform (Sect. 7.5.3)
FIR – Far Infrared
FIRAS – Far Infrared Absolute Spectrophotometer (instrument onboard COBE; see Fig. 4.3)
FJ – Faber–Jackson (Sect. 3.4.2)
FOC – Faint Object Camera (HST instrument)
FORS – Focal Reducer / Low Dispersion Spectrograph (VLT instrument)
FOS – Faint Object Spectrograph (HST instrument)
FP – Fundamental Plane (Sect. 3.4.3)
FR (I/II) – Fanaroff–Riley Type (Sect. 5.1.2)
FUSE – Far Ultraviolet Spectroscopic Explorer (Sect. 1.3.4)
FWHM – Full Width Half Maximum (Sect. 5.1.4)
GALEX – Galaxy Evolution Explorer (Sect. 1.3.4)
GBM – Gamma-ray Burst Monitor (instrument on Fermi—Sect. 1.3.6)
GC – Galactic Center (Sects. 2.3, 2.6)
GEMS – Galaxy Evolution from Morphology and Spectral Energy Distributions (Sect. 9.2.1)
GGL – Galaxy-Galaxy Lensing (Sect. 7.7)
GMT – Giant Magellan Telescope (Chap. 11)
GOODS – Great Observatories Origins Deep Survey (Sect. 9.2.1)
GR – General Relativity
GRB – Gamma-Ray Burst (Sect. 1.3.5, 9.7)
GTC – Gran Telescopio Canarias (Sect. 1.3.3)
GUT – Grand Unified Theory (Sect. 4.5.3)
Gyr – Gigayear $= 10^9$ years
GZK – Greisen–Zatsepin–Kuzmin (Sect. 2.3.4)
HB – Horizontal Branch
HCG – Hickson Compact Group (catalogue of galaxy groups, Sect. 6.2.3)
HDF(N/S) – Hubble Deep Field (North/South) (Sect. 1.3.3, 9.2.1)
HDM – Hot Dark Matter (Sect. 7.4.1)
HEAO – High Energy Astrophysical Observatory (Sect. 1.3.5)
H.E.S.S. – High Energy Stereoscopic System (Sect. 1.3.6)
HFI – High-Frequency Instrument (onboard the Planck satellite; Sect. 8.6.6)
HIFI – Heterodyne Instrument for the Far Infrared (Herschel instrument—Sect. 1.3.2)
HLS – Herschel Lensing Survey (Sect. 9.2.3)
HRD – Hertzsprung–Russell Diagram (Appendix B)
HRI – High Resolution Imager (ROSAT instrument)
HST – Hubble Space Telescope (Sect. 1.3.3)
HVC – High Velocity Cloud (Sect. 2.3.6)
HUDF – Hubble Ultradeep Survey (Sect. 9.2.1)
IAU – International Astronomical Union
ICL – IntraCluster Light (Sect. 6.3.4)
ICM – Intra-Cluster Medium (Chap. 6)
IFU – Integral Field Unit (Sect. 1.3.3)
IGM – Intergalactic Medium (Sect. 10.3)
IMF – Initial Mass Function (Sect. 3.5.1)
IoA – Institute of Astronomy (Cambridge)
IR – Infrared (Sect. 1.3.2)
IRAC – Infrared Array Camera (instrument on Spitzer—Sect. 1.3.2)
IRAS – Infrared Astronomical Observatory (Sect. 1.3.2)
IRS – Infrared Spectrograph (instrument on Spitzer—Sect. 1.3.2)
ISM – Interstellar Medium
ISO – Infrared Space Observatory (Sect. 1.3.2)
ISW effect – Integrated Sachs–Wolfe effect (Sect. 8.6.1)
IUE – International Ultraviolet Explorer (Sect. 1.3.4)
IVC – Intermediate-Velocity Cloud (Sect. 2.3.7)
JCMT – James Clerk Maxwell Telescope (Sect. 1.3.1)
JVAS – Jodrell Bank-VLA Astrometric Survey (Sect. 3.11.3)
JWST – James Webb Space Telescope (Chap. 11)
KAO – Kuiper Airborne Observatory (Sect. 1.3.2)
KiDS – KiLO Degree Survey (Chap. 11)
LAB – Leiden-Argentine-Bonn (Sect. 1.4)
LAE – Lyman Alpha Emitter (Sect 9.1.3)
LAT – Large Area Telescope (instrument on Fermi—Sect. 1.3.6)
LBG – Lyman-Break Galaxy (Sect. 9.1.1)
LBT – Large Binocular Telescope (Sect. 1.3.3)
LCRS – Las Campanas Redshift Survey (Sect. 8.1.2)
LFI – Low-Frequency Instrument (onboard the Planck satellite; Sect. 8.6.6)
LHC – Large Hadron Collider
LINER – Low-Ionization Nuclear Emission-Line Region (Sect. 5.2.3)
LIRG – Luminous InfraRed Galaxy (Sect. 9.4.1)
LISA – Laser Interferometer Space Antenna (Chap. 11)
LMC – Large Magellanic Cloud
LMT – Large Millimeter Telescope (Chap. 11)
LOFAR – Low Frequency Array (Chap. 11)
LSB galaxy – Low Surface Brightness galaxy (Sect. 3.3.2)

LSR – Local Standard of Rest (Sect. 2.4.1)
LSS – Large-Scale Structure (Chap. 8)
LSST – Large Synoptic Survey Telescope (Chap. 11)
MACHO – Massive Compact Halo Object (and collaboration of the same name, Sect. 2.5)
MAGIC – Major Atmospheric Gamma-ray Imaging Cherenkov telescope (Sect. 1.3.6)
MAMBO – Max-Planck Millimeter Bolometer (Sect. 9.3.3)
MAXIMA – Millimeter Anisotropy Experiment Imaging Array (Sect. 8.6.4)
MDM – Mixed Dark Matter (Sect. 7.4.2)
MIPS – Multiband Imaging Photometer for Spitzer (instrument on Spitzer— Sect. 1.3.2)
MIR – Mid-Infrared
MLCS – Multi-Color Light Curve Shape (Sect. 3.9.4)
MMT – Multi-Mirror Telescope
MOND – Modified Newtonian Dynamics (Chap. 11)
MS – used for the 'Main Sequence' of stars, or the 'Millennium Simulation' (Sect. 7.5.3)
MW – Milky Way
MXXL – Millennium XXL simulation (Sect. 7.5.3)
NAOJ – National Astronomical Observatory of Japan
NFW – Navarro, Frenk & White (-profile, Sect. 7.6.1)
NGC – New General Catalog (Chap. 3)
NGP – North Galactic Pole (Sect. 2.1)
NICMOS – Near Infrared Camera and Multi-Object Spectrometer (HST instrument— Sect. 1.3.3)
NIR – Near Infrared
NLR – Narrow Line Region (Sect. 5.4.3)
NLRG – Narrow Line Radio Galaxy (Sect. 5.2.4)
NOAO – National Optical Astronomy Observatory
NRAO – National Radio Astronomy Observatory
NTT – New Technology Telescope (Sect. 1.3.3)
NVSS – NRAO VLA Sky Survey (Sect. 1.4)
OGLE – Optical Gravitational Lensing Experiment (microlensing collaboration, Sect. 2.5)
OVV – Optically Violently Variable (Sect. 5.2.5)
PACS – Photodetector Array Camera and Spectrometer (Herschel instrument— Sect. 1.3.2)
PL – Period-Luminosity (Sect. 2.2.7)
PLANET – Probing Lensing Anomalies Network (microlensing collaboration, Sect. 2.5)
PM – Particle-Mesh (Sect. 7.5.3)
P^3M – Particle-Particle Particle-Mesh (Sect. 7.5.3)
PN – Planetary Nebula
POSS – Palomar Observatory Sky Survey (Sect. 1.4)
PSF – Point Spread Function
PSPC – Position-Sensitive Proportional Counter (ROSAT instrument)
QCD – Quantum Chromodynamics (Sect. 4.4.1)
QSO – Quasi-Stellar Object (Sect. 5.2.1)
RASS – ROSAT All-Sky Survey (Sect. 6.4.5)
RCS – Red Cluster Sequence (Sect. 6.8)
REFLEX – ROSAT-ESO Flux-Limited X-Ray survey
RGB – Red Giant Branch (Sect. 3.5.2)
ROSAT – Roentgen Satellite (Sect. 1.3.5)
SAO – Smithsonian Astrophysical Observatory
SCUBA – Sub-millimeter Common-User Bolometer Array (Sect. 1.3.1)
SDSS – Sloan Digital Sky Survey (Sects. 1.4, 8.1.2)
SFR – Star Formation Rate (Sect. 9.6.1)
SGP – South Galactic Pole (Sect. 2.1)
SIS – Singular Isothermal Sphere (Sect. 3.11.2)
SKA – Square Kilometer Array (Chap. 11)
SLACS – Sloan Lens Advanced Camera for Surveys (Sect. 3.11.3)
SMBH – Supermassive Black Hole (Sect. 5.3)
SMC – Small Magellanic Cloud
SMG – Sub-Millimeter Galaxy (Sect. 9.3.3)
SN(e) – Supernova(e) (Sect. 2.3.2)
SNR – Supernova Remnant
SOFIA – Stratospheric Observatory for Infrared Astronomy (Sect. 1.3.2)
SPH – Smooth Particle Hydrodynamics (Sect. 10.6.1)
SPIRE – Spectral and Photometric Imaging REceiver (Herschel instrument— Sect. 1.3.2)
SPT – South Pole Telescope (Sect. 1.3.1)
SQLS – SDSS Quasar Lens Search (Sect. 3.11.3)
STIS – Space Telescope Imaging Spectrograph (HST instrument)
STScI – Space Telescope Science Institute (Sect. 1.3.3)
SZ – Sunyaev–Zeldovich (-effect, Sect. 6.4.4)
TDE – Tidal Disruption Event (Sect. 5.5.6)
TeVeS – Tensor-Vector-Scalar (Chap. 11)
TF – Tully–Fisher (Sect. 3.4)
TMT – Thirty Meter Telescope (Chap. 11)
TP-AGB star – Thermally Pulsating AGB star (Sect. 3.5.5)
UDF – Ultra Deep Field (Sect. 9.2.1)
UHECRs – Ultra-High Energy Cosmic Rays (Sect. 2.3.4)
ULIRG – Ultraluminous Infrared Galaxy (Sect. 9.3.1)
ULX – Ultraluminous Compact X-ray Source (Sect. 9.3.1)
UV – Ultraviolet
VISTA – Visible and Infrared Survey Telescope (Sect. 1.3.3)
VLA – Very Large Array (Sect. 1.3.1)
VLBA – Very Long Baseline Array (Sect. 1.3.1)
VLBI – Very Long Baseline Interferometer (Sect. 1.3.1)

VLT – Very Large Telescope (Sect. 1.3.3)
VST – VLT Survey Telescope (Sect. 1.3.3)
VVDS – VIMOS VLT Deep Survey (Sect. 8.1.2)
WD – White Dwarf (Sect. 2.3.2)
WDM – Warm Dark Matter (Sect. 7.8)
WFIRST – Wide Field Infrared Space Telescope (Chap. 11)
WIMP – Weakly Interacting Massive Particle (Sect. 4.4.3)
WISE – Wide-field Infrared Survey Explorer(Sect. 1.3.2)
WFC3 – Wide Field Camera 3 (HST instrument—Sect. 1.3.3)
WFI – Wide Field Imager (camera at the ESO/MPG 2.2m telescope, La Silla, Sect. 6.6.2)
WFPC2 – Wide Field and Planetary Camera 2 (HST instrument– Sect. 1.3.3)
WMAP – Wilkinson Microwave Anisotropy Probe (Sect. 8.6.5)
XDF – eXtremely Deep Field (Sect. 9.2.1)
XMM – X-ray Multi-Mirror Mission (Sect. 1.3.5)
XRB – X-Ray Background (Sect. 9.5.3)
ZAMS – Zero Age Main Sequence (Sect. 3.5.2)

Solution to 1.1. If the object at current distance D had a constant velocity $v = H_0 D$ for all times, it needed a time $t = D/v = 1/H_0$ to reach separation D. This time is independent of D. Using (1.7), we find that

$$H_0^{-1} = \frac{3.086 \times 10^{24}\,\mathrm{cm}}{h\,10^7\,\mathrm{cm}}\,\mathrm{s} = 9.77\,h^{-1} \times 10^9\,\mathrm{yr}\,,$$

where we used that $1\,\mathrm{yr} = 3.16 \times 10^7\,\mathrm{s}$. For a value of $h \approx 0.71$, this time is comparable to, but slightly larger than the age of the oldest stars. Light can propagate a distance c/H_0 over the time-scale H_0^{-1}, where

$$\frac{c}{H_0} = \frac{2.998 \times 10^5\,\mathrm{km\,s^{-1}}}{100h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}} = 2.998 h^{-1}\,\mathrm{Gpc}\,.$$

Solution to 1.2. The number of galaxies in a sphere of radius $r_0 = 1\,h^{-1}\,\mathrm{Gpc}$ is $N = (4\pi/3) r_0^3\, n_0$, where $n_0 = 2 \times 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$. Thus,

$$\begin{aligned} N &= (4\pi/3) h^{-3}\,\mathrm{Gpc}^3\, 2 \times 10^{-2}\,h^3\,\mathrm{Mpc}^{-3} \\ &= (8\pi/3)\,10^7 \approx 8 \times 10^7\,. \end{aligned}$$

The number of these galaxies per square degree on the sky is obtained by dividing N by the solid angle of the sky, which is 4π steradian. Since 180° corresponds to π rad, we have that $1\,\mathrm{steradian} = (180/\pi)^2\,\mathrm{deg}^2$, so that the full sky has a solid angle of $4\pi\,(180/\pi)^2\,\mathrm{deg}^2 = 41253\,\mathrm{deg}^2$. This yields a number density of $\sim 2 \times 10^3\,\mathrm{deg}^{-2}$.

To calculate the fraction of the sky covered by the luminous region of these galaxies, we consider first a thin spherical shell of radius r and thickness $\mathrm{d}r$ around us. In this shell, there are $\mathrm{d}N = 4\pi r^2\, n_0\, \mathrm{d}r$ galaxies, each of them subtending a solid angle of $\pi R^2/r^2$, where $R = 10\,\mathrm{kpc}$ is the radius of the luminous region. Thus, the solid angle covered by all galaxies in the shell is $\mathrm{d}\omega = 4\pi^2\, R^2\, n_0\, \mathrm{d}r$. The solid angle covered by all galaxies within distance r_0 is obtained by integrating this expression over r,

$$\omega = \int_0^{r_0} \mathrm{d}\omega = 4\pi^2\, R^2\, n_0 \int_0^{r_0} \mathrm{d}r = 4\pi^2\, R^2\, n_0\, r_0\,.$$

The fraction of the sky covered by these galaxies is $f = \omega/(4\pi) = \pi\, R^2\, n_0\, r_0 \approx 0.6\,\%$.

Solution to 1.3.

(1) The mean baryon density of the Universe is $\rho_\mathrm{b} = 0.15\Omega_\mathrm{m}\, 3H_0^2/(8\pi G)$. Making use of (1.14), this yields $\rho_\mathrm{b} = 4.3 \times 10^{-31}\,\mathrm{g\,cm^{-3}}$. The estimate to the local mass density yields $\rho_{*\mathrm{local}} = 1\,M_\odot\,\mathrm{pc}^{-3} \approx 2 \times 10^{33}\,\mathrm{g}\,(3 \times 10^{18}\,\mathrm{cm})^{-3} = (2/27) \times 10^{-21}\,\mathrm{g/cm^3} \approx 7 \times 10^{-23}\,\mathrm{g/cm^3}$. Thus, $\rho_{*\mathrm{local}}/\rho_\mathrm{b} \approx 1.6 \times 10^8$.

(2) According to (1.1), the mass of the Galaxy inside R_0 is $M = R_0\, V_0^2/G$, yielding a mean density within R_0 of

$$\rho_8 = \frac{M}{(4\pi/3)\,R_0^3} = \frac{V_0^2}{(4\pi/3)\,G\,R_0^2}\,.$$

The mean matter density of the Universe is $\rho_\mathrm{m} = \Omega_\mathrm{m}\, 3H_0^2/(8\pi G)$. With $\Omega_\mathrm{m} \approx 0.3$, this yields

$$\begin{aligned} \frac{\rho_8}{\rho_\mathrm{m}} &= \frac{V_0^2}{(4\pi/3)\,G\,R_0^2}\,\frac{8\pi G}{\Omega_\mathrm{m}\,3H_0^2} = \frac{2V_0^2}{\Omega_\mathrm{m}\,R_0^2\,H_0^2} \\ &= \frac{2}{h^2\,\Omega_\mathrm{m}}\left(\frac{220\,\mathrm{km/s}}{8\,\mathrm{kpc}\;100\,(\mathrm{km/s})\,\mathrm{Mpc}^{-1}}\right)^2 \\ &= \frac{2}{h^2\,\Omega_\mathrm{m}}\left(0.275 \times 10^3\right)^2 \approx 1.0 \times 10^6\,. \end{aligned}$$

(3) The mean number of baryons N_b in the box is the volume of the box times the mean number density of baryons. The latter is given by the mean mass density ρ_b of baryons in the Universe, divided by the mass per baryon, which is $m_\mathrm{b} \approx 1.7 \times 10^{-24}\,\mathrm{g}$. Thus, making use of value of ρ_b derived above,

$$\begin{aligned} N_\mathrm{b} &\approx 1\,\mathrm{m^3}\,4.3 \times 10^{-31}\,\mathrm{g/cm^3}\,(1.7 \times 10^{-24}\,\mathrm{g})^{-1} \\ &\approx (0.43/1.7) \approx 0.25\,. \end{aligned}$$

Thus, the mean baryon density in the Universe is about $(1/4)\,\mathrm{m}^{-3}$.

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7,

Solution to 1.4.

(1) Differentiating the ansatz for $r(t)$, one finds $\dot{r} = -\alpha(r_0/t_\mathrm{f})\,(1-t/t_\mathrm{f})^{\alpha-1}$, and $\ddot{r} = \alpha(\alpha-1)(r_0/t_\mathrm{f}^2)\,(1-t/t_\mathrm{f})^{\alpha-2}$. Inserting this into the equation of motion yields

$$\alpha(\alpha-1)(r_0/t_\mathrm{f}^2)\,(1-t/t_\mathrm{f})^{\alpha-2} = -GM\,r_0^{-2}\,(1-t/t_\mathrm{f})^{-2\alpha}\,.$$

The powers of the time-dependent term must be the same one both sides, yielding $\alpha - 2 = -2\alpha$, or $\alpha = 2/3$. Equating the prefactors then yields

$$\frac{2r_0}{9t_\mathrm{f}^2} = \frac{GM}{r_0^2} \;\Rightarrow\; t_\mathrm{f} = \sqrt{\frac{2r_0^3}{9GM}}\,.$$

The solution $r(t)$ has infinite radius for $t \to -\infty$, but the inflow velocity $\dot{r}$ tends to zero as $t \to -\infty$. At $t = 0$, $r(0) = r_0$, but the velocity $\dot{r}(0)$ is finite and determined by M and r_0. For $t = t_\mathrm{f}$, the radius shrinks to zero.

(2) Replacing the mass by the mean initial density $\bar{\rho}$ of the sphere leads to

$$t_\mathrm{f} = \sqrt{\frac{1}{6\pi G\bar{\rho}}}\,,$$

so that the time-scale t_f depends only on the initial density. We can now compare t_f to the orbital time $t_\mathrm{orb} = 2\pi R_0/V_0$. Using (1.1), and replacing M by the mean density, we get $t_\mathrm{orb} = \sqrt{3\pi/(G\bar{\rho})}$. Hence, $t_\mathrm{f}/t_\mathrm{orb} = 1/(\sqrt{18}\pi) \approx 0.075$. Thus, $t_\mathrm{f} \approx 1.75 \times 10^7$ yr.

(3) Using the mean matter density of the Universe, $\bar{\rho} = \rho_\mathrm{m}$, as given by (1.10), we obtain

$$t_\mathrm{f} = \frac{2}{3\sqrt{\Omega_\mathrm{m}}H_0}\,.$$

Hence, t_f is very similar to the estimated age of the Universe, H_0^{-1}, and agrees with the age of the Einstein–de Sitter model (1.13), for which $\Omega_\mathrm{m} = 1$. Indeed, the expansion history of the Einstein–de Sitter model follows exactly the same equation of motion, except that the relevant solution is an expanding one, $r(t) \propto t^{2/3}$, which is obtained from our infalling solution by inverting the arrow of time (note that the equation of motion is invariant against $t \to -t$), and shifting the origin of the time axis.

Solution to 2.1. The angular diameter is $\delta = 3476/385000 \approx 9.03 \times 10^{-3}$. To convert this to degrees, we recall that $\pi = 180°$, so that $\delta \approx 9.03 \times 10^{-3}\,(180°/\pi) \approx 0.517° \approx 31'$. The solid angle covered by the Moon is $(0.517°)^2\pi/4 \approx 0.21\,\mathrm{deg}^2$, so the fraction it covers of the full sky is (cf. problem 1.2) $0.21/41253 \approx 5.1 \times 10^{-6}$.

Solution to 2.2. The total energy radiated throughout the galaxy's lifetime is $E = L\,t$, where $t = 10^{10}$ yr is the assumed age. The energy generated is the mass that is converted into helium in nuclear fusion, times the energy released per unit mass. The former is $Y\,M$, where Y is the helium mass fraction generated by nuclear fusion. The energy released per unit mass is $\epsilon\,c^2$, where ϵ is the efficiency of this energy generating process, given by the ratio of the binding energy per nucleon in helium ($\sim 28\,\mathrm{MeV}/4 = 7\,\mathrm{MeV}$) and the mass per nucleon, $m_\mathrm{nuc} \approx m_\mathrm{p} \approx 938\,\mathrm{MeV}$; i.e., $\epsilon \approx 0.008$. Thus, the total energy released is $E = M\,Y\,\epsilon\,c^2$. Equating this to $L\,t$, we obtain

$$Y = \frac{L\,t}{\epsilon\,M\,c^2} = \frac{1}{3\epsilon}\,\frac{L_\odot\,t}{M_\odot\,c^2}\,,$$

where we used the mass-to-light ratio. Inserting the values for $M_\odot \approx 2 \times 10^{33}$ g and $L_\odot \approx 3.8 \times 10^{33}$ erg/s, and using $t \approx 3.1 \times 10^{17}$ s, we obtain $Y \approx 2.7\,\%$. This value is lower by a factor of ten than the observed helium abundance. On the other hand, adding this value to the helium abundance from BBN yields a result which is very close to the currently observed helium abundance in local galaxies.

Solution to 2.3. According to the Kepler rotation, $V(R) = \sqrt{G\,M(R)/R}$, a constant V implies $M(R) \propto R$. The relation between M and ρ is

$$M(R) = 4\pi \int_0^R \mathrm{d}r\; r^2\,\rho(r)\,.$$

Differentiating this w.r.t. R yields $\mathrm{d}M/\mathrm{d}R = 4\pi\,R^2\,\rho(R)$. On the other hand, $M(R) \propto R$ implies $\mathrm{d}M/\mathrm{d}R = \text{const.}$, so that $\rho(r)\,r^2 = \text{const.}$, or $\rho(r) \propto r^{-2}$.

Solution to 2.4. Light at the Solar limb is deflected by $\hat{\alpha}_\odot = 1\farcs74$, which is far smaller than the angular radius $\theta_\odot = 16'$ of the Sun. Hence, if we consider a cone of light rays with vertex at the Earth and an opening angle of $\theta_\odot$, this cone will continue to diverge after being deflected at the Solar limb. If the Sun had a larger distance D, its angular radius would be smaller, namely $\theta = (D/1\,\mathrm{AU})^{-1}\theta_\odot$. If θ equals $\hat{\alpha}_\odot$, the rays of the cone with opening angle θ at Earth would be parallel after light deflection at the Solar limb, and if the distance was slightly larger they would converge after deflection and go through a common focus. If a source was placed at this focus, the Sun would then produce an Einstein ring. Displacing the source slightly, the ring breaks up into a pair of images, with angular separation 2θ. Thus, the minimum distance for lensing is $D = \theta_\odot/\hat{\alpha}_\odot$ AU ≈ 552 AU, and the image splitting at the minimum distance would be $2\hat{\alpha}_\odot \approx 3\farcs48$.

Solution to 2.5. Kepler rotation yields $V(r) = \sqrt{GM_\bullet/r}$, which we rewrite as

$$V(r) = c\sqrt{\frac{GM_\odot}{c^2\,\mathrm{pc}}}\left(\frac{M_\bullet}{M_\odot}\right)^{1/2}\left(\frac{r}{1\,\mathrm{pc}}\right)^{-1/2}$$
$$\approx 2.2\times10^{-7}c\left(\frac{M_\bullet}{M_\odot}\right)^{1/2}\left(\frac{r}{1\,\mathrm{pc}}\right)^{-1/2}.$$

With $M_\bullet \approx 4\times10^6 M_\odot$, we obtain at $r = 4\,\mathrm{pc}$ a rotational velocity of $2.2\times10^{-4}c \approx 66\,\mathrm{km/s}$. Hence, at about this radius, the Keplerian rotation velocity around the black hole equals the velocity dispersion of the stellar cluster.

Solution to 2.6. A light ray from the GC to us which is scattered in the screen at radius R has a geometric length of $L = L_1 + L_2$, where L_1 is the length of the ray path from us to the point R in the screen, and L_2 the length of the path from the GC to that point. Trigonometry then yields $L_2 = \sqrt{D^2+R^2} = D\sqrt{1+(R/D)^2} \approx D\left[1+R^2/(2D^2)\right]$, where we made use of the fact that $R/D \ll 1$ and used a first-order Taylor expansion of the square root. Similarly, $L_1 = D_{\mathrm{sc}}\left(1+R^2/(2D_{\mathrm{sc}}^2)\right)$. Thus, the total length of the ray path is

$$L = L_1 + L_2 = D + D_{\mathrm{sc}} + \frac{R^2}{2}\left(\frac{1}{D}+\frac{1}{D_{\mathrm{sc}}}\right).$$

The light-travel time along this ray is L/c. If we define t as the excess of this light-travel time relative to a straight ray, we obtain

$$t = \frac{R^2}{2c}\left(\frac{1}{D}+\frac{1}{D_{\mathrm{sc}}}\right) = \frac{R^2}{2c}\left(\frac{R_0}{D\,D_{\mathrm{sc}}}\right) \approx \frac{R^2}{2c\,D}$$

where in the final step we used $D \ll R_0$, or $D_{\mathrm{sc}} \approx R_0$. Hence, $R^2 \approx 2cDt$. Differentiating both sides w.r.t. t, we get $2R\dot{R} = 2cD$, or $\dot{R} = cD/R$. This apparent velocity is larger than the speed of light, since $R \ll D$. If the scattering screen is located behind the Galactic center, the situation is very similar to the one described here.

Solution to 3.1.

(1) We consider the central square arcsecond of the galaxy as a source; its apparent magnitude (all magnitudes considered here are in the B-band) is $m = \mu_0$ arcsec^2 $= 21.5$. Thus, its absolute magnitude is $M = m - 5\log(D/10\,\mathrm{pc})$. The corresponding luminosity of that source is $L = L_\odot\,10^{-0.4(M-M_\odot)} = L_\odot\,10^{-0.4(m-M_\odot)}\,(D/10\,\mathrm{pc})^2$. The central square arcsecond corresponds to an area of $(D\,1'')^2$, and so the central surface brightness is $I_0 = L/(D\,1'')^2$. Thus,

$$I_0 = L_\odot\,10^{-0.4\times15.96}\frac{1}{(1''\,10\,\mathrm{pc})^2} \approx 175\,\frac{L_\odot}{\mathrm{pc}^2}.$$

(2) At the distance of 16 Mpc, $50''$ correspond to $h_R \approx 4\,\mathrm{kpc}$. The luminosity is obtained by integrating the surface brightness over the disk,

$$L = 2\pi\int_0^\infty \mathrm{d}R\;R\,I_0\,\mathrm{e}^{-R/h_R} = 2\pi\,h_R^2\,I_0,$$

and by inserting numbers,

$$L = 2\pi(4\,\mathrm{kpc})^2\,\frac{175\,L_\odot}{\mathrm{pc}^2} \approx 1.76\times10^{10}\,L_\odot.$$

Solution to 3.2. According to the assumptions, the mass function of the stars is $n(m)\,\mathrm{d}m \propto m^{-\alpha}\,\mathrm{d}m$, with $\alpha = 2.35$. If half the mass is contained in stars with mass $m_{\mathrm{L}} \le m \le m_{m50}$, then

$$\int_{m_{\mathrm{L}}}^{m_{\mathrm{U}}}\mathrm{d}m\;m\,n(m) = 2\int_{m_{\mathrm{L}}}^{m_{m50}}\mathrm{d}m\;m\,n(m),$$

so that the result for m_{m50} is independent of the normalization of the mass function. Integration then yields

$$m_{\mathrm{L}}^{2-\alpha} - m_{\mathrm{U}}^{2-\alpha} = 2\left(m_{\mathrm{L}}^{2-\alpha} - m_{m50}^{2-\alpha}\right).$$

Solving for m_{m50} yields

$$m_{m50} = 2^{1/(\alpha-2)}\left(m_{\mathrm{L}}^{2-\alpha}+m_{\mathrm{U}}^{2-\alpha}\right)^{1/(2-\alpha)},$$

or $m_{m50} \approx 0.55\,M_\odot$.

To calculate m_{L50}, we need to satisfy

$$\int_{m_{\mathrm{L}}}^{m_{\mathrm{U}}}\mathrm{d}m\;m^3\,n(m) = 2\int_{m_{\mathrm{L}}}^{m_{L50}}\mathrm{d}m\;m^3\,n(m).$$

The same steps than lead to the result

$$m_{L50} = 2^{1/(\alpha-4)}\left(m_{\mathrm{L}}^{4-\alpha}+m_{\mathrm{U}}^{4-\alpha}\right)^{1/(4-\alpha)},$$

or $m_{L50} \approx 46\,M_\odot$. We thus see that a stellar population with such a mass spectrum contains most of its mass in the low-mass stars, whereas most of the luminosity is due to the highest-mass stars.

Solution to 3.3. The volume contained in a solid angle ω out to distance D is $V = \omega\,D^3/3$. If $\dot{n}$ is the supernova event rate per unit volume, the observed SNe rate in the solid angle ω up to distance D is $\dot{N} = \dot{n}\,V = \omega\,\dot{n}\,D^3/3$. Inserting numbers, $\dot{N} = \omega\,10^{-5}\,\mathrm{Mpc}^{-3}\,\mathrm{yr}^{-1}\,(500\,\mathrm{Mpc})^3/3 = \omega\,417\,\mathrm{yr}^{-1}$. Since 1 deg $= \pi/180$, we can write $\omega = (\omega/\mathrm{deg}^2)(\pi/180)^2$,

so that $\dot{N} \approx 0.127\,(\omega/\mathrm{deg}^2)\,\mathrm{yr}^{-1}$. Thus, one needs to survey $\sim 80\,\mathrm{deg}^2$ to find 10 nearby SNe per year.

Solution to 3.4.

(1) The volume of a cone with height $D_{\rm lim}$ and opening solid angle ω is

$$\int_0^{D_{\rm lim}} \mathrm{d}r\ \omega\, r^2 = D_{\rm lim}^3 \omega/3\ .$$

(2) If the luminosity of a source is larger than $L = 4\pi\, S_{\rm min}\, D_{\rm lim}^2$, its flux is larger than $S_{\rm min}$ even out to distances $D_{\rm lim}$, and so they can be seen throughout the search volume. On the other hand, if their luminosity is smaller than this value, they can only be seen out to a distance $D = D_{\rm max}(L) = \sqrt{L/(4\pi S_{\rm min})}$. Thus,

$$D_{\rm max}(L) = \min\left(\sqrt{\frac{L}{4\pi S_{\rm min}}}, D_{\rm lim}\right)\ ;$$

$$V_{\rm max}(L) = \frac{\omega}{3} D_{\rm max}^3(L)\ .$$

(3) If $\Phi(L)\,\mathrm{d}L$ is the number density of galaxies with luminosity within $\mathrm{d}L$ of L, then we observe

$$N(L) = \int_0^L \mathrm{d}L'\ \Phi(L')\, V_{\rm max}(L')$$

galaxies in the survey with luminosity $\le L$. Differentiation of this relation then immediately yields the desired result.

Solution to 3.5.

(1) The deflection angle corresponding to the surface mass density (3.84) consists of two terms. The deflection due to the second term in (3.84) is simply $\lambda\boldsymbol{\alpha}(\boldsymbol{\theta})$. The deflection due to the first term can be calculated from (3.70). Together, we find $\boldsymbol{\alpha}_\lambda(\boldsymbol{\theta}) = (1-\lambda)\boldsymbol{\theta} + \lambda\boldsymbol{\alpha}(\boldsymbol{\theta})$.

(2) The corresponding lens equation reads $\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}_\lambda(\boldsymbol{\theta}) = \lambda\,[\boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})]$. Dividing the lens equation by λ, we obtain $\boldsymbol{\beta}_\lambda = \boldsymbol{\beta}/\lambda = \boldsymbol{\theta} - \boldsymbol{\alpha}(\boldsymbol{\theta})$. That means that the new mass distribution κ_λ yields the same image positions as the original mass model, provided the source position is changed from $\boldsymbol{\beta}$ to $\boldsymbol{\beta}_\lambda$. Since the source position in unobservable, this shift in the source plane can not be observed.

(3) The magnification is given by (3.68), and thus reads for the modified mass model κ_λ

$$\mu_\lambda = \left|\det\left(\frac{\partial\boldsymbol{\beta}_\lambda}{\partial\boldsymbol{\theta}}\right)\right|^{-1} = \frac{1}{\lambda^2}\left|\det\left(\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\theta}}\right)\right|^{-1} = \frac{\mu}{\lambda^2}\ .$$

Hence, the magnification of all images is changed with this new mass model; however, the magnification ratios between images stays the same, and thus the predicted flux ratios.

Solution to 4.1. At BBN, $T \approx 0.1\,\mathrm{MeV} \approx 10^9\,\mathrm{K} \approx 4\times10^8 T_0$, where T_0 is the current temperature of the CMB. Hence, BBN happens at a scale factor of $a_{\rm BBN} \approx 2.5\times10^{-9}$. The current baryon density is determined by (4.68) and the critical density, $\rho_{\rm b}(0) \approx 0.02h^{-2}\rho_{\rm cr}$, and the baryon density at BBN is then $\rho_{\rm b}(\mathrm{BBN}) = a_{\rm BBN}^{-3}\rho_{\rm b}(0)$. Using (4.15), we find $\rho_{\rm b}(\mathrm{BBN}) \sim 2.5\times10^{-5}\,\mathrm{g/cm}^3$. This density is many orders of magnitude lower than in the center of the Sun, or in other stars.

Nuclear burning in stars is slow because the thermal energy of protons (i.e., temperature) is too low to allow overcoming the Coulomb barrier, that is the electrostatic repulsion between equally charged particles. Only very rarely does this process happen; quantum-mechanically, it occurs by a process called 'tunneling'. In contrast, the temperature during BBN was high, which makes it far easier to beat electrostatic repulsion. Second, and most important, stars contain essentially no free neutrons, in contrast to the situation at BBN; the formation of deuterium at BBN thus did not involve a Coulomb barrier, only the transformation of deuterium into helium.

The energy density ρ_E released during BBN is given by the mass fraction of nucleons that ended up in helium, which is $Y \approx 0.25$, times the binding energy per nucleon in helium, corresponding to $\sim 7\,\mathrm{MeV}$, times the number density of nucleon $n_{\rm nucl}$. The latter is given by the baryon density at BBN, $\rho_{\rm b}(\mathrm{BBN})$, divided by the mass of the nucleon, which is about $940\,\mathrm{MeV}/c^2$. Thus,

$$\rho_E = Y\,\frac{7\,\mathrm{MeV}}{940\,\mathrm{MeV}}\,\frac{\rho_{\rm b}(0)}{a_{\rm BBN}^3}\,c^2\ .$$

The energy density of photons at BBN is given by $\rho_\gamma = \Omega_{\rm CMB}\rho_{\rm cr}a_{\rm BBN}^{-4}c^2$. Hence,

$$\frac{\rho_E}{\rho_\gamma} \approx 7.4\times10^{-3} Y\,\frac{\Omega_{\rm b}h^2}{\Omega_{\rm CMB}}\,a_{\rm BBN} \approx 3.7\times10^{-10}\,,$$

where we made use of the results from the first part of this problem and used (4.27). Hence, the energy released during BBN is totally negligible compared to the energy of the photon gas.

Solution to 4.2. From the general definition (4.34), we need to calculate

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + 3P/c^2\right)\ .$$

We have (at the current epoch) $\rho = \rho_{\rm cr}\sum_i \Omega_i$, and $P/c^2 = \rho_{\rm cr}\sum_i \Omega_i\, w_i$, where the sum extends over all N species. Hence, with the definition of $\rho_{\rm cr}$, we obtain $\ddot{a}/a = -(H_0^2/2)\sum_i \Omega_i(1+3w_i)$. Using (4.34), this yields $q_0 = (1/2)\sum_i \Omega_i(1+3w_i)$. With a pressureless matter component ($w = 0$) and the vacuum energy with $w = -1$, (4.35) is recovered.

Solution to 4.3. Since $\dot{a} = H(a)a = H_0E(a)a$, an expansion can change into a contraction (or the opposite) only when $H(a) = 0$. For $\Omega_\Lambda = 0$, we have $E^2 = \Omega_{\rm m}/a^3 + (1-\Omega_{\rm m})/a^2 = \Omega_{\rm m}(1-a)/a^3 + a^{-2}$. This expression is always positive for $0 < a < 1$. From the first form of E^2, we see that this expression is also positive for $a > 1$ if $\Omega_{\rm m} \le 1$. If $\Omega_{\rm m} > 1$, $E^2 = 0$ at $a = a_{\max} = \Omega_{\rm m}/(\Omega_{\rm m}-1)$.
Including a finite Ω_Λ, we have $E^2 = \Omega_{\rm m}/a^3 + (1-\Omega_{\rm m}-\Omega_\Lambda)/a^2 + \Omega_\Lambda$. We rewrite this expression as $E^2 = (1-\Omega_{\rm m})/a^2 + \Omega_{\rm m}/a^3 + \Omega_\Lambda(1-1/a^2)$ and see that all terms are non-negative for all $a > 1$ if $\Omega_{\rm m} \le 1$, i.e., the universe expands forever in the future. Using the form $E^2 = (1-\Omega_\Lambda)/a^2 + \Omega_\Lambda + \Omega_{\rm m}(1-a)/a^3$, we see that all terms are positive provided $0 \le \Omega_\Lambda < 1$.
Assume that at $t_{\rm ex}$, an expansion turns into a contraction. Since $\dot{a}^2 = a^2H^2(a)$, we then have that $\dot{a} = +a\sqrt{H^2(a)}$ for $t < t_{\rm ex}$, and $\dot{a} = -a\sqrt{H^2(a)}$ for $t > t_{\rm ex}$. By integrating this expression, we obtain for $t > 0$:

$$t = \int_{t_{\rm ex}}^{t_{\rm ex}+t} {\rm d}t = -\int_{a(t_{\rm ex})}^{a(t_{\rm ex}+t)} \frac{{\rm d}a}{a\sqrt{H^2}} = \int_{a(t_{\rm ex}+t)}^{a(t_{\rm ex})} \frac{{\rm d}a}{a\sqrt{H^2}}\,,$$

and for $t < 0$

$$t = \int_{t_{\rm ex}-t}^{t_{\rm ex}} {\rm d}t = \int_{a(t_{\rm ex}-t)}^{a(t_{\rm ex})} \frac{{\rm d}a}{a\sqrt{H^2}}\,.$$

Combining these two equation yields

$$\int_{a(t_{\rm ex}+t)}^{a(t_{\rm ex})} \frac{{\rm d}a}{a\sqrt{H^2}} = \int_{a(t_{\rm ex}-t)}^{a(t_{\rm ex})} \frac{{\rm d}a}{a\sqrt{H^2}}\,,$$

which implies $a(t_{\rm ex}-t) = a(t_{\rm ex}+t)$.

Solution to 4.4. The Friedmann equation in a flat universe reads $(\dot{a}/a)^2 = H_0^2(\Omega_{\rm m}/a^3 + \Omega_\Lambda)$, with $\Omega_\Lambda = 1-\Omega_{\rm m}$. With the ansatz $a = v^\beta$, we find $(\dot{a}/a) = \beta(\dot{v}/v)$, in terms of which the expansion equation becomes $\dot{v}^2 = (H_0/\beta)^2\left(\Omega_{\rm m}v^{(2-3\beta)} + \Omega_\Lambda v^2\right)$. The desired form is achieved by setting $\beta = 2/3$, yielding $\dot{v}^2 = (9H_0^2\Omega_{\rm m}/4)\left[1 + (\Omega_\Lambda/\Omega_{\rm m})v^2\right]$.
With the ansatz $v(t) = v_0\sinh(t/t_{\rm a})$, $\dot{v} = (v_0/t_{\rm a})\cosh(t/t_{\rm a})$, and we obtain $(v_0/t_{\rm a})^2\cosh^2(t/t_{\rm a}) = (9H_0^2\Omega_{\rm m}/4)\left[1 + (\Omega_\Lambda/\Omega_{\rm m})v_0^2\sinh^2(t/t_{\rm a})\right]$. In order for this equation to be valid, the time-dependence on the r.h.s. must be of the form $\cosh^2(t/t_{\rm a})$; this can be achieved by choosing v_0 such that $(\Omega_\Lambda/\Omega_{\rm m})v_0^2 = 1$, or $v_0 = \sqrt{\Omega_{\rm m}/\Omega_\Lambda}$. The prefactors on both sides also must agree, which then determines $t_{\rm a} = 2/(3H_0\sqrt{\Omega_\Lambda})$.
Hence, the final solution reads

$$a(t) = v^{2/3} = \left(\frac{\Omega_{\rm m}}{\Omega_\Lambda}\right)^{1/3}\sinh^{2/3}\left(\frac{3H_0\sqrt{\Omega_\Lambda}\,t}{2}\right)\,.$$

Considering the case $t \ll t_{\rm a}$ and making use of $\sinh(x) \approx x$ for $x \ll 1$, we find $a = \left(3H_0\sqrt{\Omega_{\rm m}}\,t/2\right)^{2/3}$, which agrees with (4.70). Hence, for times $t \ll t_{\rm a}$, the expansion law does not contain Ω_Λ explicitly. On the other hand, for $t \gg t_{\rm a}$, using $\sinh(x) \approx {\rm e}^x/2$ for $x \gg 1$, we obtain $a = (1/2)\,(\Omega_{\rm m}/\Omega_\Lambda)^{1/3}\exp\left(H_0\sqrt{\Omega_\Lambda}\,t\right)$. Hence, for late times, the universe expands exponentially. Note that this last solution satisfies $(\dot{a}/a) = H_0\sqrt{\Omega_\Lambda}$, the Friedmann equation for a Λ-dominated universe.
Finally, we consider the sign of the second derivative of a. With $a = v^{2/3}$, we find in turn $\dot{a} = (2/3)v^{-1/3}\dot{v}$, $\ddot{a} = (2/3)v^{-1/3}\ddot{v} - (2/9)v^{-4/3}\dot{v}^2$, and thus $\ddot{a} = 2v_0^2/(9v^{4/3}t_{\rm a}^2)\left[3\sinh^2(t/t_{\rm a}) - \cosh(t/t_{\rm a})\right]$. The prefactor is always positive, and the term in brackets is negative for $t \ll t_{\rm a}$, since $\cosh(x) \approx 1$ for $x \ll 1$, and positive for $t \gg t_{\rm a}$, when both $\sinh(x) \approx {\rm e}^x/2 \approx \cosh(x)$. Hence, the solution describes the transition from an decelerating universe to an accelerating one.

Solution to 4.5. In this case, the Friedmann equation reads $(\dot{a}/a)^2 = H_0^2(\Omega_{\rm r}/a^4 + \Omega_\Lambda)$. Using the same ansatz as in problem 4.4, we now need to choose $\beta = 1/2$, $v_0 = \sqrt{\Omega_{\rm r}/\Omega_\Lambda}$, and $t_{\rm a} = \left(2H_0\sqrt{\Omega_\Lambda}\right)^{-1}$, to obtain

$$a = \left(\frac{\Omega_{\rm r}}{\Omega_\Lambda}\right)^{1/4}\sinh^{1/2}\left(2H_0\sqrt{\Omega_\Lambda}\,t\right)\,.$$

For $t \ll t_{\rm a}$, this becomes $a \approx \sqrt{2H_0\sqrt{\Omega_{\rm r}}\,t}$, which has the $t^{1/2}$-dependence that we derived for the radiation-dominated era. For $t \gg t_{\rm a}$, the expansion is exponential.

Solution to 4.6. From (4.39), we find ${\rm d}a/a = (H/c){\rm d}r = (H/c)a\,{\rm d}x$, where we used the relation between physical length and comoving length, ${\rm d}r = a\,{\rm d}x$. Hence,

$${\rm d}x = \frac{c}{H(a)}\frac{{\rm d}a}{a^2} = \frac{c}{H_0}\,E^{-1}(a)\,\frac{{\rm d}a}{a^2}\,.$$

Integrating both sides from some point along the ray, characterized by the scale factor a or, equivalently, the redshift $z = 1/a - 1$, and corresponding to the comoving distance $x(z)$, and using (4.33), one arrives at (4.53).

Consider two light rays separated by a small angle θ at the observer. For a flat universe, the comoving separation between these two rays is then given by $L(a) = \theta x$. The physical separation is then $aL(a)$. According to the definition of D_A, we then have $D_A = a\,L(a)/\theta = x/(1+z)$, which reproduces (4.54) for the case $K = 0$.

Solution to 4.7.

(1) Consider the case $K > 0$ first. We note from (4.85) that the parameter θ corresponding to $t = t_1$ is $\theta_1 = 0$, since $(\theta - \sin\theta)$ is a monotonically increasing function. The first of (4.85) then yields $f(t_1) = 0$, i.e., the initial condition $f = 0$ at $t = t_1$ is satisfied by (4.85). Denoting derivatives w.r.t. t by a dot, those w.r.t. θ by a prime, (4.84) reads $\dot{f}^2 = C/f - K$. Differentiation yields $f' = C\sin\theta/(2K)$, $t' = C(1-\cos\theta)/(2K^{3/2})$, so that $\dot{f} = f'/t' = K^{1/2}\sin\theta/(1-\cos\theta)$. The l.h.s. of (4.84) then becomes

$$\dot{f}^2 = K\,\frac{\sin^2\theta}{(1-\cos\theta)^2} = K\,\frac{1-\cos^2\theta}{(1-\cos\theta)^2} = K\,\frac{1+\cos\theta}{1-\cos\theta}.$$

The r.h.s. of (4.84) is $C/f - K = 2K/(1-\cos\theta) - K$, which is seen to agree with the above expression. Hence, (4.85) indeed solves (4.84) with the correct initial condition. The case $K < 0$ can be treated in the same way. For $K = 0$, (4.87) yields $f(t_1) = 0$, as required, and $\dot{f} = (2C/3)^{1/3}(t-t_1)^{-1/3}$. Simple algebra then shows that $\dot{f}^2 - C/f = 0$. Considering the case $K > 0$ again, f attains a maximum where $\cos\theta$ has its minimum, which occurs for $\theta = \pi$; hence, $f_{\max} = C/K$, at time $t_{\max} = t(\pi) = t_1 + \pi C/(2K^{3/2})$. Furthermore, for $\theta = 2\pi$, $f = 0$, which happens at time $t_{\rm coll} = t(2\pi) = t_1 + \pi C/K^{3/2}$.

(2) For $\Omega_\Lambda = 0 = \Omega_{\rm r}$, (4.33) reads $\dot{a}^2 = H_0^2\,[\Omega_{\rm m}/a + (1-\Omega_{\rm m})]$, which is seen to have the same form as (4.84), with $C = H_0^2\Omega_{\rm m}$ and $K = H_0^2(\Omega_{\rm m} - 1)$. Setting $t_1 = 0$ then yields $a(0) = 0$, the correct initial condition for Friedmann expansion. Hence, with these parameter values, (4.85) describes the expansion for $\Omega_{\rm m} > 1$, (4.86) the expansion for $\Omega_{\rm m} < 1$, and (4.87) yields the EdS case, $a(t) = (3H_0t/2)^{2/3}$. For $\Omega_{\rm m} > 1$, the previous results then show that $a_{\max} = C/K = \Omega_{\rm m}/(\Omega_{\rm m} - 1)$, occurring at time $t_{\max} = \pi C/(2K^{3/2}) = \pi/(2H_0)\,\Omega_{\rm m}/(\Omega_{\rm m}-1)^{3/2}$, and collapse happens at $t_{\rm coll} = 2t_{\max}$.

(3) Differentiation of (4.84) w.r.t. t yields $2\dot{f}\ddot{f} = -C\dot{f}/f^2$, or $\ddot{f} = -(C/2)/f^2$. The equation for the radius of the sphere is $\ddot{r} = -GM/r^2$, so that we can identify f with r, and $C = 2GM$. The constant K is proportional to the (negative of the) total energy of sphere, $-K/2 = \dot{r}^2/2 - GM/r$, as the sum of specific kinetic and potential energy. With $r_0 = r(0)$, we can write $K = 2GM/r_0 - \dot{r}^2(0)$. The solution found in problem 1.4 was $r = r_0(1-t/t_{\rm f})^{2/3}$, which yields $K = (1-t/t_{\rm f})^{-2/3}\left[2GM/r_0 - 4r_0^2/(9t_{\rm f}^2)\right] = 0$, when using the value for $t_{\rm f}$ derived in problem 1.4. This solution corresponds to one where the sphere at $t = 0$ has an initial infall speed, such that the total energy of the sphere is zero, in full analogy to a time-reversed Einstein–de Sitter model. Setting the initial velocity to zero, $\dot{r}(0) = 0$, yields $K = 2GM/r_0 > 0$. In the context of the equation of motion discussed above, the free-fall time is the time between the maximum expansion and the time of collapse, $t_{\rm ff} = t_{\rm coll} - t_{\max} = \pi C K^{-3/2}/2$, which yields $t_{\rm ff} = \sqrt{\pi^2 r_0^3/(8GM)}$. Replacing M by the mean density of the sphere, we arrive at (4.88).

Solution to 4.8.

(1) Conservation of kinetic plus potential energy (per unit mass) yields $\dot{r}^2/2 - GM_{\rm E}/r = \text{const.}$, or $\dot{r}^2 = 2GM_{\rm E}/r - K$. At initial time t_0, $r(t_0) = r_{\rm E}$, $\dot{r}(t_0) = v_0$, so that $K = v_{\rm esc}^2 - v_0^2$, which is assumed to be positive in the following.

(2) The equation of motion has the form (4.84), with $C = 2GM_{\rm E}$, hence the solution (4.85) applies. Without loss of generality, we set $t_1 = 0$, and denote the time when the object leaves the Earth surface by t_0. The corresponding parameter value θ_0 is found from the first of (4.85), $r_{\rm E} = C(1-\cos\theta_0)/(2K) \approx C\theta_0^2/(4K)$, where we used the leading order of the Taylor expansion and assumed that $\theta_0 \ll 1$, which needs to be verified. Writing the initial velocity as a fraction μ of the escape velocity, $v_0 = \mu v_{\rm esc}$, we see that $\theta_0^2 = 4Kr_{\rm E}/C = 4(1-\mu^2)v_{\rm esc}^2 r_{\rm E}/(2GM_{\rm E}) = 4(1-\mu^2)$. Hence, $\theta_0 \ll 1$ if the initial velocity is sufficiently close to the escape velocity, $1-\mu \ll 1$, which will be assumed in the following (corresponding to the assumption that the flight is 'long'). Then using the second of (4.85), we obtain in the same manner $t_0 = C(\theta_0 - \sin\theta_0)/(2K^{3/2}) \approx C\theta_0^3/(12K^{3/2}) = 2(1-\mu^2)^{3/2}CK^{-3/2}/3$. Since $C/K^{3/2} = (1-\mu^2)^{-3/2}r_{\rm E}/v_{\rm esc}$, we finally obtain $t_0 = (2/3)r_{\rm E}/v_{\rm esc}$.

(3) The return time $t_{\rm ret}$ is given by $t_{\rm coll} - 2t_0$, with $t_{\rm coll} = \pi C/K^{3/2} = \pi(1-\mu^2)^{-3/2}r_{\rm E}/v_{\rm esc}$. Provided $1-\mu \ll 1$, $t_{\rm coll} \gg t_0$, and so we can approximate $t_{\rm ret} \approx t_{\rm coll}$. Inserting numbers, we obtain $t_{\rm ret} \approx 1800\,{\rm s}/(1-\mu^2)^{3/2}$. Setting $t_{\rm ret} = 1\,{\rm d}$, we find $\mu \approx 0.93$, whereas for $t_{\rm ret} = 1\,{\rm yr}$, $\mu \approx 0.994$. Hence, if one wants to have a really long flight, the initial velocity must be extremely well tuned.

Solution to 4.9. The momentum behaves like $p = mv \propto (1+z)$, where m is the rest mass of the (non-relativistic) particle, and v its velocity as measured by a comoving

observer. The temperature T_b is related to the mean velocity dispersion as $(3/2)k_\mathrm{B}T_\mathrm{b} = mv^2/2$. Thus, $T_\mathrm{b} \propto v^2 \propto (1+z)^2$.

Solution to 4.10. Consider a cosmic epoch, characterized by the scale factor a, at which the neutrinos were ultra-relativistic; their momentum then was $p = E/c$. The mean energy of the thermal distribution is about $3T_\nu$, where $T_\nu \sim 0.7T_\gamma$, according to (4.62). With $T_\gamma = 2.73\,\mathrm{K}/a \approx 2.5 \times 10^{-4}\,\mathrm{eV}/a$, we then get $pa \sim 5\times10^{-4}\,\mathrm{eV}/c$. This product is conserved, as shown by (4.47). When the temperature of the universe drops below the neutrino rest mass, the momentum is $p = m_\nu v$. Thus, we obtain for the characteristic velocity of cosmic neutrinos

$$v \sim (1+z)\left(\frac{m_\nu c^2}{1\,\mathrm{eV}}\right)^{-1} 150\,\mathrm{km/s}\,.$$

Solution to 4.11.

(1) From (4.56), we see that $t = t_0(1+z)^{-3/2}$. The look-back time is $\tau(z) = t_0 - t(z) = \left[1-(1+z)^{-3/2}\right]t_0$. The redshift where $\tau(z) = t_0/2$ is then determined by $(1+z)^{3/2} = 2$, or $z = 2^{2/3} - 1 \approx 0.59$.

(2) The volume of the spherical shell is the product of the surface of the sphere at redshift z and the physical thickness $c\,\mathrm{d}t$ of the shell, corresponding to the redshift interval $\mathrm{d}z$. By definition of the angular diameter distance, the surface is $4\pi D_\mathrm{A}^2(z)$, so that

$$\mathrm{d}V = 4\pi D_\mathrm{A}^2(z)\left|\frac{c\,\mathrm{d}t}{\mathrm{d}z}\right|\mathrm{d}z\,.$$

Furthermore, from $\mathrm{d}t = \mathrm{d}a/(a\,H)$ and $\mathrm{d}a/a = -\mathrm{d}z/(1+z)$, we find that $\mathrm{d}t = -\mathrm{d}z/[(1+z)H]$. Using (4.57) and $H = H_0(1+z)^{3/2}$, valid for the EdS model, one finally finds

$$V = 16\pi\left(\frac{c}{H_0}\right)^3 \frac{1}{(1+z)^{9/2}}\left(1-\frac{1}{\sqrt{1+z}}\right)^2\,.$$

(3) The number of objects is $N = \int \mathrm{d}V_\mathrm{com}\, n_\mathrm{com} = n_\mathrm{com}\int \mathrm{d}V_\mathrm{com}$, where the comoving volume element $\mathrm{d}V_\mathrm{com}$ is related to the physical (proper) volume element by $\mathrm{d}V_\mathrm{com} = (1+z)^3\,\mathrm{d}V$. Thus

$$N = 16\pi\left(\frac{c}{H_0}\right)^3 n_\mathrm{com}\int_0^z \frac{\mathrm{d}z'}{(1+z')^{3/2}}\left(1-\frac{1}{\sqrt{1+z'}}\right)^2$$

$$= \frac{32\pi}{3}\left(\frac{c}{H_0}\right)^3 n_\mathrm{com}\left(1-\frac{1}{\sqrt{1+z}}\right)^3\,.$$

Specializing the last result to $z \ll 1$, using $1/\sqrt{1+z} \approx 1 - z/2$, we then obtain $N \approx (4\pi/3)(c\,z/H_0)^3 n_\mathrm{com}$, which is the number of objects in a sphere of radius $c\,z/H_0$.

Solution to 4.12. The reason for the absence of H_0 in (4.61) is that in the Friedmann equation (4.18) curvature and cosmological constant can be neglected at early times, and thus $(\dot a/a)^2 \propto \rho_\mathrm{r} \propto T^4(a)$. In other words, $\dot a/a = CT^2$, where the constant C depends on the number of relativistic species only—see (4.60). The first law of thermodynamics (4.17) then yields that $T \propto 1/a$ [see (4.24)], which leads to $\dot T/T = -\dot a/a = -CT^2$, without any reference to the current universe – neither to its expansion rate, nor to its current temperature. The foregoing equation can be solved with the ansatz $T(t) = xt^\alpha$, which yields $\dot T/T = \alpha/t = -Cx^2t^{2\alpha}$, and thus $\alpha = -1/2$, $x = 1/\sqrt{2C}$.

The helium abundance depends on the density of baryons (given that the density of photons as a function of temperature is known). But $\rho_\mathrm{b} = \Omega_\mathrm{b}\rho_\mathrm{cr}(1+z)^3 \propto \Omega_\mathrm{b}h^2(1+z)^3$. Hence, the combination $\Omega_\mathrm{b}h^2$ determines the physical baryon density.

Solution to 4.13. The scattering optical depth is given by the line-of-sight integral over the product of the Thompson scattering cross section σ_T and the number density n_e of free electrons,

$$\tau = \sigma_\mathrm{T}\int c\,\mathrm{d}t\; n_\mathrm{e}\,.$$

From problem 4.11 we know that $\mathrm{d}t = -\mathrm{d}z/(z\,H)$ (in this problem we set $1+z \approx z$, since all the contributions to the integral comes from $z \gtrsim 800$). Since recombination happens in the matter-dominated epoch, we have $z\,H = H_0\Omega_\mathrm{m}^{1/2}z^{5/2} \propto h\,\Omega_\mathrm{m}^{1/2}z^{5/2}$. The number density of free electrons is equal to the number density of protons n_p and the ionization fraction x. For n_p, we have $n_\mathrm{p} \propto \Omega_\mathrm{b}\rho_\mathrm{cr}z^3 \propto \Omega_\mathrm{b}h^2z^3$, so that

$$\tau(z) \propto \int_0^z \mathrm{d}y\,\frac{\Omega_\mathrm{b}h^2y^3}{h\,\Omega_\mathrm{m}^{1/2}y^{5/2}}\,x(y)\,.$$

Using (4.72), we see that the Ω's and h's drop out, and we obtain

$$\tau(z) \propto \int_0^z \mathrm{d}y\; y^{13.25}\,,$$

which yields the correct functional dependence of (4.73).

Solution to 5.1.

(1) The integral over the emissivity of a single electron $f(\nu/\nu_\mathrm{c})$ over all frequencies is

$$\int_0^\infty \mathrm{d}\nu\ f(\nu/\nu_\mathrm{c}) = \nu_\mathrm{c} \int_0^\infty \mathrm{d}x\ f(x) \propto \gamma^2\ ,$$

where we used (5.3). This dependence on the Lorentz factor γ is thus the same as in (5.5).

(2) The distribution of relativistic electrons is $N(\gamma)\,\mathrm{d}\gamma = a\,\gamma^{-s}\,\mathrm{d}\gamma$, since $E = \gamma m_\mathrm{e} c^2$, and a is a constant $\propto A$. The synchrotron emissivity of this distribution is obtained by integrating the emissivity of a single electron over the electron distribution,

$$\epsilon_\nu = \int_0^\infty \mathrm{d}\gamma N(\gamma)\ f(\nu/\nu_\mathrm{c}) = a \int_0^\infty \mathrm{d}\gamma \gamma^{-s} f\left(\frac{\nu}{\nu_0 \gamma^2}\right),$$

where we defined $\nu_0 = \nu_\mathrm{c}/\gamma^2 = 3eB/(4\pi m_\mathrm{e} c)$. Changing the integration variable to $x = \nu/(\nu_0\gamma^2)$, with $\mathrm{d}\gamma = (-1/2)\sqrt{\nu/\nu_0}\, x^{-3/2}\mathrm{d}x$ then yields

$$\epsilon_\nu = \frac{a}{2}\left(\frac{\nu}{\nu_0}\right)^{-(s-1)/2} \int_0^\infty \mathrm{d}x\ x^{(s-3)/2}\ f(x)\ .$$

Since the final integral is frequency-independent, we find that the emitted spectrum is a power law with index $\alpha = (s-1)/2$.

Solution to 5.2.

(1) Replace $B^2 = 8\pi U_\mathrm{B}$ in (5.5) and use the definition (5.23) for σ_T to arrive at the expression given.

(2) Consider first the case that all photons have the same energy E_γ. The energy loss of a relativistic electron scattering a photon then is on average $\Delta E = -(4/3)\gamma^2 E_\gamma$, and the time between scatterings is $\Delta t = \left(n_\gamma c \sigma_\mathrm{T}\right)^{-1}$. With $U_\gamma = n_\gamma E_\gamma$ and $\mathrm{d}E/\mathrm{d}t = \Delta E/\Delta t$, the desired expression is obtained. Since this expression no longer refers to the photon energy (or frequency), but only to the energy density of photons, the same result is obtained for a spectral distribution of photons.

Solution to 5.3. We write the temperature profile (5.13) as $T(r) = T_0(r/r_0)^{-3/4}$. The specific intensity is that of a blackbody with temperature $T(r)$, so that

$$I_\nu(r) \propto \nu^3 \left[\exp\left(\frac{h_\mathrm{P}\nu}{k_\mathrm{B} T(r)}\right) - 1\right]^{-1}\ .$$

The emitted luminosity L_ν is then the integral of the specific intensity over the surface of the disk,

$$L_\nu \propto \int_0^\infty \mathrm{d}r\ r\ I_\nu(r)\ ,$$

where we neglected boundary effects by setting the integration limits to 0 and ∞. Writing the exponent as

$$x = \frac{h_\mathrm{P}\nu}{k_\mathrm{B} T(r)} = \frac{h_\mathrm{P}\nu}{k_\mathrm{B} T_0}\left(\frac{r}{r_0}\right)^{3/4} =: \frac{\nu}{\nu_0}\left(\frac{r}{r_0}\right)^{3/4}\ ,$$

where in the last step we defined ν_0, we see that $r/r_0 = x^{4/3}(\nu/\nu_0)^{-4/3}$. Therefore, $r\,\mathrm{d}r \propto x^{5/3}(\nu/\nu_0)^{-8/3}\mathrm{d}x$, and

$$L_\nu \propto \left(\frac{\nu}{\nu_0}\right)^{1/3} \int_0^\infty \mathrm{d}x\ \frac{x^{5/3}}{\mathrm{e}^x - 1}\ .$$

Solution to 5.4. At an accretion rate of $\dot{m}$, the growth rate of the black hole is $\dot{M}_\bullet = (1-\epsilon)\dot{m}$, since the fraction ϵ of the accretion rate is converted into luminosity, and thus does not contribute to the increase of the black hole mass. With (5.15), one then finds

$$\dot{M}_\bullet = \frac{1-\epsilon}{\epsilon\, c^2}\frac{L}{L_\mathrm{edd}} L_\mathrm{edd} = \frac{1-\epsilon}{\epsilon}\frac{L}{L_\mathrm{edd}}\frac{M_\bullet}{t_\mathrm{gr}}\ ,$$

and the relation (5.47) is the solution of this equation.

Inserting the specific values $L = L_\mathrm{edd}$, $\epsilon = 0.1$, and $t = 10^9\,\mathrm{yr}$, we find $M_\bullet(t) = M_\bullet(0)\,\mathrm{e}^{18} \approx 6.6\times 10^7\, M_\bullet(0)$, or $M_\bullet(t) \approx 6.6\times 10^8 M_\odot$.

Solution to 5.5.

(1) We assume all clouds to be at the characteristic distance r from the SMBH; each cloud covers a solid angle of $\pi r_\mathrm{c}^2/r^2$. The covering fraction is obtained by summing the solid angle over all N_c clouds and dividing by 4π, to find $f_\mathrm{cov} = (N_\mathrm{c}/4)\ (r_\mathrm{c}/r)^2$.

(2) The volume filling factor f_V is the ratio of the total volume of the clouds, $N_\mathrm{c}(4\pi/3)r_\mathrm{c}^3$ to that of the BLR, $(4\pi/3)r^3$, $f_V = N_\mathrm{c}(r_\mathrm{c}/r)^3$.

(3) Using $f_V = 10^{-6}$, one obtains $(r_\mathrm{c}/r) = 10^{-2}N_\mathrm{c}^{-1/3}$. From $f_\mathrm{cov} = 0.1$ we get $N_\mathrm{c} = 0.4(r_\mathrm{c}/r)^{-2}$. Combining these two expressions, we obtain $N_\mathrm{c} = 6.4\times 10^{10}$ and $r_\mathrm{c} = 2.5\times 10^{-6} r = 2.5\times 10^{10}\,\mathrm{cm}$. The total gas mass of the clouds is the product of the clouds' volume times the electron density (that then yields the total number of electrons in the clouds) times the average mass per electron, which is about the proton mass (since there are about as many electrons as nucleons). Hence, $M_\mathrm{c} = n_\mathrm{e} V_\mathrm{c} m_\mathrm{p}$. The volume is $V_\mathrm{c} = N_\mathrm{c}(4\pi/3)r_\mathrm{c}^3 \approx 4\times 10^{42}\,\mathrm{cm}^3$, so that $M_\mathrm{c} = 4\times 10^{52}\ 1.67\times 10^{-24}\,\mathrm{g} \approx 6.7\times 10^{28}\,\mathrm{g} \approx 3.4\times 10^{-5} M_\odot$. Hence, the total gas mass in the BLR is very small indeed.

Solution to 5.6.

(1) According to the assumption, the optical light from the AGN is

$$L_\mathrm{AGN,opt} = 0.1L = 1.3\times 10^{37}\,\mathrm{erg/s}\,\frac{L}{L_\mathrm{edd}}\frac{M_\bullet}{M_\odot}$$

$$\approx 3400 L_\odot \frac{L}{L_{\rm edd}} \frac{M_\bullet}{M_\odot} ,$$

where we used the Solar luminosity $L_\odot = 3.85 \times 10^{38}\,{\rm erg/s}$. The optical light from the host galaxy is $L_* = M_* (M/L)^{-1} = (M_*/M_\odot) L_\odot (M/L)_\odot/(M/L)$. Using $M_\bullet = 10^{-3} f_{\rm sph} M_*$, we then find

$$\frac{L_{\rm AGN,opt}}{L_*} \approx 3.4 \frac{L}{L_{\rm edd}} f_{\rm sph} \frac{(M/L)}{(M/L)_\odot} .$$

(2) With an Eddington ratio of 0.1 and a typical mass-to-light ratio of 3 in Solar units, the light from the AGN is comparable to that of the host galaxy if the spheroidal fraction is close to unity. For late-type galaxies, where $f_{\rm sph}$ is considerably smaller, the host galaxy will typically dominate the optical emission—this is one of the reasons why a complete census of AGNs in the optical is difficult to achieve. The AGN can outshine the host galaxy only for apparently large Eddington ratios (e.g., when beaming is involved), or if the SMBH is considerably larger than the assumed scaling.

Solution to 5.7. If the tidal disruption distance $R_{\rm t} = R_* (M_\bullet/M_*)^{1/3}$ is smaller than the Schwarzschild radius of the SMBH, the star will be swallowed by the black hole before it can be disrupted. Hence we require $R_{\rm t} > r_{\rm S}$. This yields $M_\bullet < R_*^{3/2} M_*^{-1/2} (2G/c^2)^{-3/2}$. Dividing the expression by the Solar mass then yields

$$\frac{M_\bullet}{M_\odot} < \left(\frac{R_*}{R_\odot}\right)^{3/2} \left(\frac{M_*}{M_\odot}\right)^{-1/2} \left(\frac{2GM_\odot}{R_\odot c^2}\right)^{-3/2} \approx 10^8 \left(\frac{R_*}{R_\odot}\right)^{3/2} \left(\frac{M_*}{M_\odot}\right)^{-1/2} ,$$

where we used the Solar radius $R_\odot \approx 7 \times 10^5\,{\rm km}$ and the Schwarzschild radius of the Sun, which is 2.95 km.

Solution to 6.1. With $t_{\rm orbit} = 2\pi r/V$ and Kepler's law $V = \sqrt{GM/r}$, we find $t_{\rm orbit} = 2\pi [GM/r^3]^{-1/2}$. Since $M = (4\pi/3) r^3 \bar\rho$, we see that

$$t_{\rm orbit} = \sqrt{\frac{3\pi}{G\bar\rho}} . \qquad \text{(E.1)}$$

Comparing this to the free-fall time yields $t_{\rm orbit}/t_{\rm ff} = \sqrt{32}$. Inserting $\bar\rho = 200 \rho_{\rm cr}(z)$ then yields $t_{\rm orbit} = (\pi/5)\, H^{-1}(z)$. Since the inverse of the expansion rate $H(z)$ is, up to factor of order unity, the age of the Universe at that epoch, we see that $t_{\rm orbit} \sim t(z)$.

Solution to 6.2. The observed bolometric flux is, according to the definition (4.50), $S = L/(4\pi D_{\rm L}^2)$, and the angular radius is $\theta = R/D_{\rm A}$. Hence, the surface brightness behaves like $I = S/(\pi\theta^2) \propto (L/R^2)(D_{\rm A}/D_{\rm L})^2 = (L/R^2)(1+z)^{-4}$. For the specific surface brightness I_ν, the redshift-dependence can be different, depending on the spectral properties of the source, according to the necessary K-correction (see Sect. 5.6.1). With $I_\nu = S_\nu/(\pi\theta^2)$ and S_ν from (5.42), we obtain $S_\nu = L_\nu/(4\pi D_{\rm L}^2)\,(1+z)^{1-\alpha}$, yielding $I_\nu \propto (1+z)^{-(3+\alpha)}$.

Solution to 7.1. If ρ is assumed to be spatially constant, the continuity equation becomes $\partial\rho/\partial t + \rho \nabla\cdot\boldsymbol{v} = 0$. Since the first term is independent of $\boldsymbol{r}$, so must be the second, which immediately implies that $\nabla\cdot\boldsymbol{v}$ can depend only on time. Inserting $\boldsymbol{v}(\boldsymbol{r},t) = H(t)\boldsymbol{r}$ then yields $\dot\rho + 3\rho H = 0$, or $\dot\rho/\rho = -3\dot a/a$. By insertion, we see that the solution with $\rho(t_0) = \rho_0$ is given by (4.11).

Since $\nabla^2|\boldsymbol{r}|^2 = 6$, a solution of the Poisson equation is

$$\Phi(\boldsymbol{r},t) = \left(\frac{2\pi}{3} G\rho(t) - \frac{\Lambda}{6}\right) \boldsymbol{r}^2 ,$$

so that $\nabla\Phi = [(4\pi/3)G\rho - \Lambda/3]\,\boldsymbol{r}$. For the terms on the l.h.s. of the Euler equation, we find

$$\frac{\partial\boldsymbol{v}}{\partial t} = \dot H \boldsymbol{r} = \left(\frac{\ddot a}{a} - \frac{\dot a^2}{a^2}\right)\boldsymbol{r} ,$$
$$(\boldsymbol{v}\cdot\nabla)\boldsymbol{v} = H^2 (\boldsymbol{r}\cdot\nabla)\boldsymbol{r} = H^2 \boldsymbol{r} .$$

Combining these terms, we see that the Euler equation yields (4.19), with $P = 0$.

Solution to 7.2.

(1) From its definition, $H = \dot a/a$, we get with (4.19), setting $P = 0$,

$$\dot H = \frac{\ddot a}{a} - \left(\frac{\dot a}{a}\right)^2 = -\frac{4\pi G}{3}\bar\rho + \frac{\Lambda}{3} - H^2 .$$

Using $\dot{\bar\rho} = -3H\bar\rho$, which follows from $\bar\rho \propto a^{-3}$, we then find by differentiation that

$$\ddot H = 4\pi G \bar\rho H - 2H\dot H ,$$

which is the same equation as (7.15) with $D = H$. Hence, the Hubble function satisfies the growth equation.

(2) We show next that $D_+(a) = CH(a)I(a)$, with

$$I(a) = \int_0^a \frac{{\rm d}a'}{[a'\,H(a')]^3} ,$$

is a solution of (7.15). We first note that

$$\dot{I} = \frac{\mathrm{d}a}{\mathrm{d}t}\frac{\mathrm{d}I}{\mathrm{d}a} = aH\,\frac{1}{a^3H^3} = \frac{1}{a^2H^2}\,,$$

so that

$$\dot{D}_+ = C\left(\dot{H}I + \frac{H}{a^2H^2}\right) = C\left(\dot{H}I + \frac{1}{a^2H}\right)\,.$$

Furthermore,

$$\ddot{D}_+ = C\left(\ddot{H}I + \frac{\dot{H}}{a^2H^2} - \frac{2\dot{a}}{a^3H} - \frac{\dot{H}}{a^2H^2}\right)$$
$$= C\left(\ddot{H}I - \frac{2}{a^2}\right).$$

Collecting terms,

$$\ddot{D}_+ + 2H\dot{D}_+ - 4\pi G\bar{\rho}D_+$$
$$= CI\left(\ddot{H} + 2H\dot{H} - 4\pi G\bar{\rho}H\right) = 0\,,$$

where the final equality was shown in the first part of this problem.

(3) For the EdS model, we have $H = 2/(3t)$, $\dot{H} = -2/(3t^2)$ and $\ddot{H} = 4/(3t^3)$. Inserting these expression as $D = H$ into (7.15), using $H_0 = 2/(3t_0)$, shows that it is satisfied, i.e., H is a solution of the growth equation. Furthermore, specializing (7.17) to the EdS parameters yields

$$D_+ \propto \left(\frac{t_0}{t}\right)\int_0^a \mathrm{d}a'\,a'^{3/2} \propto \left(\frac{t_0}{t}\right)a^{5/2} \propto \left(\frac{t}{t_0}\right)^{2/3}\,.$$

Inserting $D = (t/t_0)^{2/3}$ into (7.15) shows that it indeed solves the growth equation.

Solution to 7.3.

(1) The relation between virial mass and virial radius is given by (7.56), so that

$$r_{200} = \left(\frac{GM}{100\,H^2(z)}\right)^{1/3}\,.$$

Using the Hubble function for a flat universe, and transforming the product GM into the Schwarzschild radius yields

$$r_{200} = \left(\frac{2GM_\odot}{c^2}\,\frac{M}{M_\odot}\,\frac{c^2}{200\,H_0^2\,[\Omega_\mathrm{m}(1+z)^3+\Omega_\Lambda]}\right)^{1/3}\,.$$

Inserting the values for the Schwarzschild radius of the Sun and c/H_0, we find

$$r_{200} \approx \left(\frac{M}{M_\odot}\right)^{1/3}\left[\Omega_\mathrm{m}(1+z)^3+\Omega_\Lambda\right]^{-1/3}h^{-2/3}$$
$$\times\left(\frac{3\times10^5\;\;81\times10^{54}}{200}\right)^{1/3}\mathrm{cm}\,.$$

With the estimates $3\times81/200 \approx 1.2$, the number in the last parenthesis yields approximately $1.2\times10^{59} = 120\times10^{57}$. Since $5^3 = 125$, the third root of 120 is very close to 5, so that

$$r_{200} \approx 5\times10^{19}\mathrm{cm}\left(\frac{M}{M_\odot}\right)^{1/3}\left[\Omega_\mathrm{m}(1+z)^3+\Omega_\Lambda\right]^{-1/3}h^{-2/3}$$
$$\approx 16\mathrm{pc}\left(\frac{M}{M_\odot}\right)^{1/3}\left[\Omega_\mathrm{m}(1+z)^3+\Omega_\Lambda\right]^{-1/3}h^{-2/3}. \qquad \text{(E.2)}$$

Thus, at $z = 0$, we obtain $r_{200} = 160h^{-1}\,\mathrm{kpc}$ and $r_{200} = 1.6h^{-1}\,\mathrm{Mpc}$ for the galaxy and cluster mass halo, respectively.

At $z = 2$, we see that the matter term in the bracket dominates, since $\Omega_\mathrm{m}(1+z)^3 = 0.3\times27 = 8.1 \gg \Omega_\Lambda$, so that $\left[\Omega_\mathrm{m}(1+z)^3+\Omega_\Lambda\right]^{-1/3} \approx 1/2$, since $2^3 = 8$. Thus, the virial radius at $z = 2$ is about half the size of that today, for fixed mass. Hence, halos of a given mass are smaller at higher redshift, as expected, since the definition of a halo—'mean density equals 200 times critical density'—together with the increase of the critical density with redshift implies that higher-redshift halos have a larger mean density, i.e., they are more compact.

For V_{200}, we start from (7.58) and use the preceding result:

$$V_{200} = 10\,H(z)\,r_{200}$$
$$= 10\,h\,100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}\sqrt{\Omega_\mathrm{m}(1+z)^3+\Omega_\Lambda}$$
$$\times\,16\mathrm{pc}\left(\frac{M}{M_\odot}\right)^{1/3}\left[\Omega_\mathrm{m}(1+z)^3+\Omega_\Lambda\right]^{-1/3}h^{-2/3} \qquad \text{(E.3)}$$
$$= 16\times10^{-3}\frac{\mathrm{km}}{\mathrm{s}}\left(\frac{M}{M_\odot}\right)^{1/3}\left[\Omega_\mathrm{m}(1+z)^3+\Omega_\Lambda\right]^{1/6}h^{1/3}.$$

Hence, our two halos have virial velocities of 160 km/s and 1600 km/s, respectively, at redshift zero, independent of h. At $z = 2$, the Ω-dependent term is $\approx \sqrt{2}$ (recall the earlier estimate), hence the corresponding virial velocities are 225 km/s and 2250 km/s, respectively. That they are higher by a factor $\sqrt{2}$ is already

clear from Kepler's law, since the radius is smaller by a factor 2, for fixed mass.

(2) The mass M contained in a proper volume $V_{\rm prop}$ at redshift z in a homogeneous Universe is $M = \bar\rho_{\rm m0}(1+z)^3 V_{\rm prop}$. Correspondingly, the mass in the comoving volume $V_{\rm com} = (1+z)^3 V_{\rm prop}$ is $M = \bar\rho_{\rm m0} V_{\rm com}$, independent of redshift. Writing $V_{\rm com} = (4\pi/3)R^3$ and using the definition of the virial radius, specialized to the current epoch, we obtain

$$\frac{4\pi}{3} R^3\, \Omega_{\rm m}\rho_{\rm cr} = \frac{4\pi}{3} r_{200}^3 200\rho_{\rm cr}\ ,$$

where r_{200} is the virial radius of the halo of mass M today. Thus,

$$R = \left(\frac{200}{\Omega_{\rm m}}\right)^{1/3} r_{200} \sim 8.7 r_{200}\ .$$

Thus, the mass of our two halos were assembled from a volume corresponding to a sphere of comoving radius of $1.4h^{-1}$ Mpc and $14h^{-1}$ Mpc, respectively.

(3) According to (7.38), the scale factor a at which a perturbation of length scale L enters the horizon is given as

$$\begin{aligned} L &= \frac{c}{H_0\sqrt{\Omega_{\rm m}}} \int_0^a \frac{{\rm d}a'}{\sqrt{a'+a_{\rm eq}}} \\ &= \frac{2c}{H_0\sqrt{\Omega_{\rm m}}} \left(\sqrt{a+a_{\rm eq}} - \sqrt{a_{\rm eq}}\right) \\ &= \frac{2c}{H_0}\sqrt{\frac{a_{\rm eq}}{\Omega_{\rm m}}} \left(\sqrt{1+\frac{a}{a_{\rm eq}}} - 1\right) \\ &\approx \frac{c}{H_0}\frac{a}{\sqrt{\Omega_{\rm m} a_{\rm eq}}} = \frac{c}{H_0}\frac{a}{\sqrt{\Omega_{\rm r}}}\ , \end{aligned} \qquad \text{(E.4)}$$

where we made a first-order Taylor expansion, assuming $a \ll a_{\rm eq}$, and used the definition of $a_{\rm eq}$. Note that we have just rederived (4.76). Using (4.28), this becomes

$$L = 3000h^{-1}{\rm Mpc}\frac{a}{\sqrt{42\times 10^{-6}h^{-2}}} \approx 4.6\times 10^5\, a\,{\rm Mpc}\ .$$

Thus, the perturbation that eventually led to the formation of our galaxy-mass halo entered the horizon at $a \approx 3\times 10^{-6}h^{-1}$, that corresponding to our cluster-mass halo at ten times this scale factor. Note that both entered the horizon in the radiation-dominated epoch, $a \ll a_{\rm eq}$, so that indeed (4.76) applies.

Solution to 7.4.

(1) The growth factor (7.17) for $a \ll 1$ is obtained from the product of $H(a) \approx H_0\sqrt{\Omega_{\rm m}}a^{-3/2}$ and the integral I in (7.17), in which for $a \ll 1$ (and thus $a' \ll 1$) the matter term dominates, i.e., $I \approx (2/5)\Omega_{\rm m}^{-3/2}a^{5/2}$. Together, $D_+ \propto a$ for sufficiently small a.

(2) To obtain the next-order term, we write for a flat universe

$$\begin{aligned} I(a) &:= \int_0^a \frac{{\rm d}a'}{\left(\Omega_{\rm m}/a' + \Omega_\Lambda a'^2\right)^{3/2}} \\ &= \frac{1}{\Omega_{\rm m}^{3/2}} \int_0^a \frac{{\rm d}a'\, a'^{3/2}}{\left[1+(\Omega_\Lambda/\Omega_{\rm m})a'^3\right]^{3/2}} \\ &\approx \frac{1}{\Omega_{\rm m}^{3/2}} \int_0^a {\rm d}a'\, a'^{3/2}\left(1 - \frac{3}{2}\frac{\Omega_\Lambda}{\Omega_{\rm m}}a'^3\right) \\ &= \frac{1}{\Omega_{\rm m}^{3/2}}\frac{2}{5}a^{5/2}\left(1 - \frac{15}{22}\frac{\Omega_\Lambda}{\Omega_{\rm m}}a^3\right)\ , \end{aligned}$$

where we performed a Taylor expansion of the integrand in the second step. A similar Taylor expansion of the Hubble function yields

$$\frac{H(a)}{H_0} \approx \frac{\sqrt{\Omega_{\rm m}}}{a^{3/2}}\left(1 + \frac{1}{2}\frac{\Omega_\Lambda}{\Omega_{\rm m}}a^3\right)\ .$$

Multiplying these two results then yields

$$D_+ \propto a\left(1 - \frac{2}{11}\frac{\Omega_\Lambda}{\Omega_{\rm m}}a^3 + \mathcal{O}(a^6)\right)\ .$$

If we consider a significant deviation from the linear behavior to occur when the second term in the parenthesis becomes of order 0.1, then we request $a^3 \lesssim 0.1(11/2)(\Omega_{\rm m}/\Omega_\Lambda)$. Taken the parameters which apply to our Universe, $\Omega_{\rm m}/\Omega_\Lambda \sim 3/7$, this becomes $a \lesssim 0.6$. This estimate is in concordance with the behavior of the dashed curve in Fig. 7.3.

Solution to 8.1.

(1) The mean number density of observed galaxies is, according to (8.50),

$$\bar n = \int {\rm d}x_3\ \nu(x_3)\,\bar n_3(x_3)\ ,$$

so that the probability $p_x(x_3)\,{\rm d}x_3$ for an observed galaxy to have distance within ${\rm d}x_3$ of x_3 is proportional to $\nu(x_3)\,\bar n_3(x_3)\,{\rm d}x_3$. Normalizing $p_x(x_3)$ to unity yields (8.51).

(2) Substituting $\nu(x_3)$ by $p_x(x_3)$ in (8.50) yields

$$\begin{aligned} n(\boldsymbol{\theta}) &= \bar n \int {\rm d}x_3\ p_x(x_3)\,\frac{n_3(f_k(x_3)\boldsymbol{\theta}, x_3)}{\bar n_3(x_3)} \\ &= \bar n\left(1 + \int {\rm d}x_3\ p_x(x_3)\,\delta_{\rm g}(f_k(x_3)\boldsymbol{\theta}, x_3)\right)\ , \end{aligned}$$

where we used that $\delta_g(\boldsymbol{x}) = [n_3(\boldsymbol{x}) - \bar{n}_3(x_3)]/\bar{n}_3(x_3)$.

(3) We first define the number density contrast

$$\delta_n(\boldsymbol{\theta}) = \frac{n(\boldsymbol{\theta}) - \bar{n}}{\bar{n}} = \int \mathrm{d}x_3 \; p_x(x_3)\, \delta_g(f_k(x_3)\boldsymbol{\theta}, x_3) \; .$$

In analogy to (7.27), the angular correlation function w is defined as

$$\langle \delta_n(\boldsymbol{\phi})\, \delta_n(\boldsymbol{\phi} + \boldsymbol{\theta}) \rangle = w(|\boldsymbol{\theta}|) \; .$$

We next define the redshift probability distribution $p(z)$ of the galaxies, given in terms of $p_x(x_3)$ by $p(z)\,\mathrm{d}z = p_x(x_3)\,\mathrm{d}x_3$, where the function $x_3(z)$ is given in (4.53); however, we do not need to use this relation explicitly. Then,

$$\begin{aligned} \langle \delta_n(\boldsymbol{\phi})\, \delta_n(\boldsymbol{\phi} + \boldsymbol{\theta}) \rangle &= \int \mathrm{d}z_1 \; p(z_1) \int \mathrm{d}z_2 \; p(z_2) \left\langle \delta_g \, \delta_g \right\rangle \\ &= \int \mathrm{d}z_1 \; p(z_1) \int \mathrm{d}z_2 \; p(z_2)\, \xi_g \; , \end{aligned}$$

where we made use of the definition (7.27) of the correlation function ξ_g, and where the argument of ξ_g depends on the separation of the two points characterized by the directions $\boldsymbol{\phi}$ and $\boldsymbol{\phi} + \boldsymbol{\theta}$ and redshifts z_i. For this, we can either use the comoving separation (as was done during most of this and the previous chapter), or use the proper separation—they just differ by a factor $(1 + z)$. The correlation ξ_g is assumed to be zero unless the two galaxies are close in space. In particular, for a non-zero correlation the two galaxies need to have a similar redshift. We thus define $z_1 = z + \Delta z/2$, $z_2 = z - \Delta z/2$, where $z = (z_1 + z_2)/2$ is the mean redshift and we assume that ξ_g vanishes unless $|\Delta z| = |z_2 - z_1| \ll z$. Then approximating $p(z_1) \approx p(z) \approx p(z_2)$, and replacing the integration over z_1 and z_2 by one over z and Δz, we obtain (8.17), where the separation in the argument of ξ_g is valid provided the angular separation $\theta \ll 1$, so that we can use $\tan\theta \approx \theta$.

Solution to 8.2. We have $L_X = 4\pi D_L^2 S_X = 4\pi(1 + z_{cl})^4 D_A^2 S_X$, where we used (4.52). Since X-ray emission is a two-body process, we have $L_X \propto n_e^2 V$, where $V \propto R^3$ is the volume of the region considered, and $R = D_A\theta$. The temperature of the gas can be determined independent of the distance, and just merely enters in the constant of proportionality. Combining the two expression for L_X, we find

$$D_A^2 \propto L_X \propto n_e^2 D_A^3 \; ,$$

or $n_e \propto D_A^{-1/2}$. The estimated gas mass is $M_{gas} \propto n_e V \propto D_A^{5/2}$. The total mass within θ as determined from (6.37) can be written as

$$M \propto T\, R \frac{\mathrm{d}\ln(n_e T)}{\mathrm{d}\ln R} \; .$$

Varying D_A adds a constant to $\ln R$ and thus does not change the derivative; hence, $M \propto D_A$. Together with the previous result, we find $f_{gas} \propto D_A^{3/2}$.

Solution to 8.3. We start with the Friedmann equation written as (4.14),

$$\dot{a}^2 = (8\pi G/3)\, \rho(a)\, a^2 - Kc^2 \; ,$$

and subtract from it the same equation specialized to the current epoch ($a = 1$) to obtain

$$\dot{a}^2 - (8\pi G/3)\, \rho(a)\, a^2 = H_0^2 - (8\pi G/3)\, \rho_0 \; ,$$

so that the constant K is eliminated. With $\rho_{cr}(a) = 3H^2(a)/(8\pi G)$ and $\Omega_0(a) = \rho(a)/\rho_{cr}(a)$, this can be rewritten as

$$a^2\, H^2(a)\, [1 - \Omega_0(a)] = H_0^2\, [1 - \Omega_0(a = 1)] \; ,$$

which reproduces (4.81), with F given by (4.82), or

$$F(a) = \left(\frac{\Omega_r}{a^2} + \frac{\Omega_m}{a} + [1 - \Omega_0(a = 1)] + \frac{\Omega_{DE}}{a^{(1+3w)}} \right)^{-1} \; ,$$

where we used (8.49) but included the curvature term. Since $w < -1/3$, the final term cannot increase as $a \to 0$. Hence, as in the case of a cosmological constant, for early times the radiation term dominates and the argument from Sect. 4.5.2 remains unchanged.

Solution to 8.4. Writing $\xi(r) = (r/r_0)^{-\gamma}$, Limber's equation (8.17) yields

$$w(\theta) = \int \mathrm{d}z p^2(z) \int_{-\infty}^{\infty} \mathrm{d}(\Delta z) \left(\frac{D_A^2(z)\theta^2 + D'^2(\Delta z)^2}{r_0^2} \right)^{-\gamma/2} \; ,$$

where $D' = \mathrm{d}D/\mathrm{d}z$. Substituting $\Delta z = \theta\, y$ and replacing the inner integral by one over y with $\mathrm{d}(\Delta z) = \theta\, \mathrm{d}y$ immediately yields the scaling $w(\theta) \propto \theta^{-(\gamma-1)}$. With the same method applied to (8.13) it is readily shown that $w_p(r_p) \propto r_p^{-(\gamma-1)}$.

Index

P. Schneider, *Extragalactic Astronomy and Cosmology*, DOI 10.1007/978-3-642-54083-7,