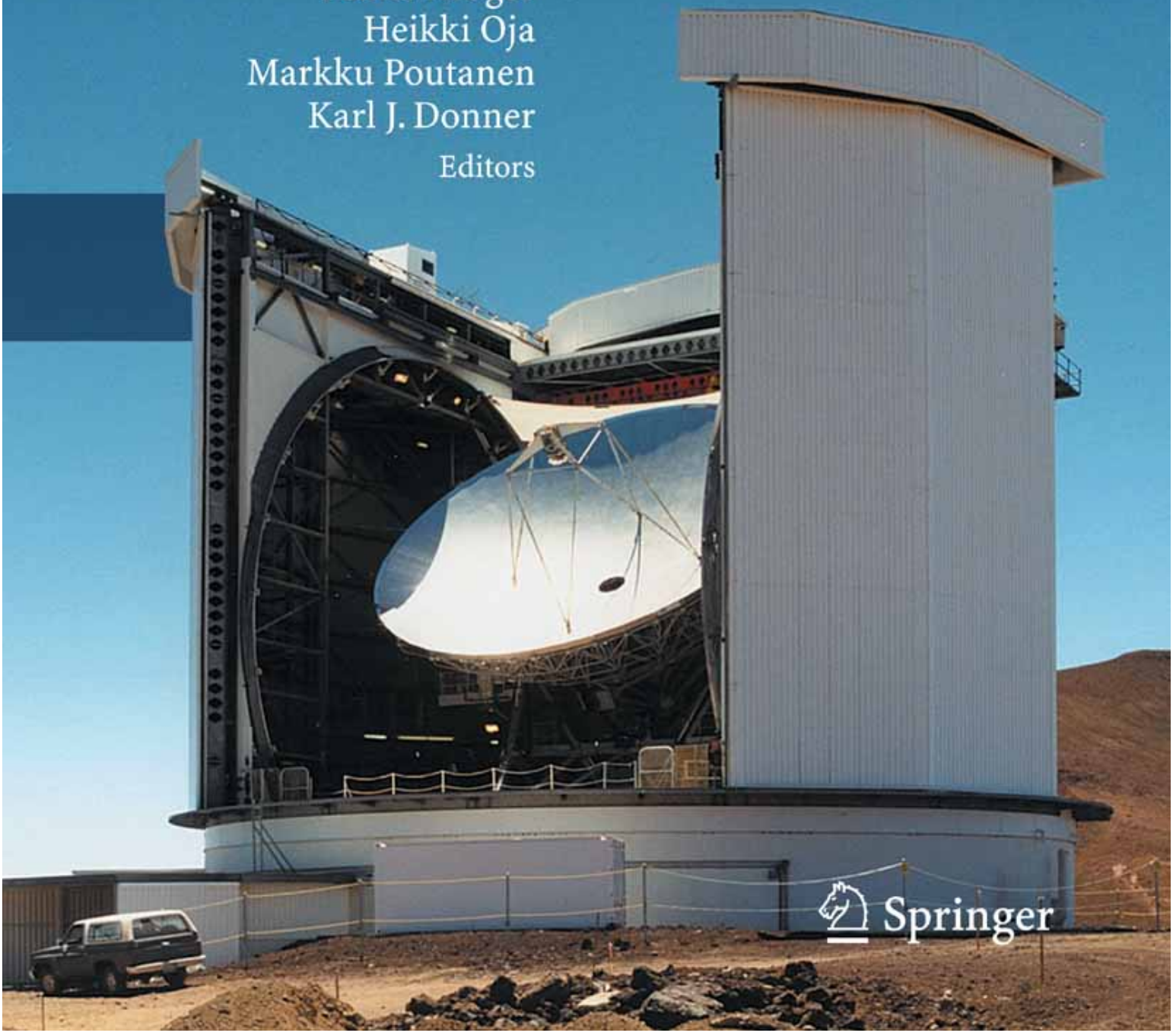
Fundamental Astronomy

Hannu Karttunen
Pekka Kröger
Heikki Oja
Markku Poutanen
Karl J. Donner
Editors

Fifth Edition



Springer

Fundamental Astronomy

H. Karttunen
P. Kröger
H. Oja
M. Poutanen
K. J. Donner (Eds.)

Fundamental Astronomy

Fifth Edition
With 449 Illustrations
Including 34 Colour Plates
and 75 Exercises with Solutions

Springer

Dr. *Hannu Karttunen*
University of Turku, Tuorla Observatory,
21500 Piikkiö, Finland
e-mail: hannu.karttunen@utu.fi

Dr. *Pekka Kröger*
Isonniitynkatu 9 C 9, 00520 Helsinki, Finland
e-mail: pekka.kroger@stadia.fi

Dr. *Heikki Oja*
Observatory, University of Helsinki,
Tähtitorninmäki (PO Box 14), 00014 Helsinki, Finland
e-mail: heikki.oja@helsinki.fi

Dr. *Markku Poutanen*
Finnish Geodetic Institute,
Dept. Geodesy and Geodynamics,
Geodeetinrinne 2, 02430 Masala, Finland
e-mail: markku.poutanen@fgi.fi

Dr. *Karl Johan Donner*
Observatory, University of Helsinki,
Tähtitorninmäki (PO Box 14), 00014 Helsinki, Finland
e-mail: donner@astro.helsinki.fi

ISBN 978-3-540-34143-7 5th Edition
Springer Berlin Heidelberg New York

ISBN 978-3-540-00179-9 4th Edition
Springer-Verlag Berlin Heidelberg New York

Library of Congress Control Number: 2007924821

Cover picture: The James Clerk Maxwell Telescope. Photo credit: Robin Phillips and Royal Observatory, Edinburgh. Image courtesy of the James Clerk Maxwell Telescope, Mauna Kea Observatory, Hawaii

Frontispiece: The Horsehead Nebula, officially called Barnard 33, in the constellation of Orion, is a dense dust cloud on the edge of a bright HII region. The photograph was taken with the 8.2 meter Kueyen telescope (VLT 2) at Paranal. (Photograph European Southern Observatory)

Title of original Finnish edition:
Tähtitieteen perusteet (Ursan julkaisuja 56)
© Tähtitieteellinen yhdistys Ursa Helsinki 1984, 1995, 2003

Sources for the illustrations are given in the captions and more fully at the end of the book. Most of the uncredited illustrations are
© Ursa Astronomical Association, Raatimiehenkatu 3A2,
00140 Helsinki, Finland

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer is a part Springer Science+Business Media

www.springer.com

© Springer-Verlag Berlin Heidelberg 1987, 1994, 1996, 2003, 2007

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

Typesetting and Production:
LE-TeX, Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover design: Erich Kirchner, Heidelberg/WMXDesign, Heidelberg
Layout: Schreiber VIS, Seeheim

Printed on acid-free paper
SPIN: 11685739 55/3180/YL 5 4 3 2 1 0

Preface to the Fifth Edition

As the title suggests, this book is about fundamental things that one might expect to remain fairly the same. Yet astronomy has evolved enormously over the last few years, and only a few chapters of this book have been left unmodified.

Cosmology has especially changed very rapidly from speculations to an exact empirical science and this process was happening when we were working with the previous edition. Therefore it is understandable that many readers wanted us to expand the chapters on extragalactic and cosmological matters. We hope that the current edition is more in this direction. There are also many revisions and additions to the chapters on the Milky Way, galaxies, and cosmology.

While we were working on the new edition, the International Astronomical Union decided on a precise definition of a planet, which meant that the chapter on the solar system had to be completely restructured and partly rewritten.

Over the last decade, many new exoplanets have also been discovered and this is one reason for the increasing interest in a new branch of science – astrobiology, which now has its own new chapter.

In addition, several other chapters contain smaller revisions and many of the previous images have been replaced with newer ones.

Helsinki *The Editors*
December 2006

Preface to the First Edition

The main purpose of this book is to serve as a university textbook for a first course in astronomy. However, we believe that the audience will also include many serious amateurs, who often find the popular texts too trivial. The lack of a good handbook for amateurs has become a problem lately, as more and more people are buying personal computers and need exact, but comprehensible, mathematical formalism for their programs. The reader of this book is assumed to have only a standard high-school knowledge of mathematics and physics (as they are taught in Finland); everything more advanced is usually derived step by step from simple basic principles. The mathematical background needed includes plane trigonometry, basic differential and integral calculus, and (only in the chapter dealing with celestial mechanics) some vector calculus. Some mathematical concepts the reader may not be familiar with are briefly explained in the appendices or can be understood by studying the numerous exercises and examples. However, most of the book can be read with very little knowledge of mathematics, and even if the reader skips the mathematically more involved sections, (s)he should get a good overview of the field of astronomy.

This book has evolved in the course of many years and through the work of several authors and editors. The first version consisted of lecture notes by one of the editors (Oja). These were later modified and augmented by the other editors and authors. Hannu Karttunen wrote the chapters on spherical astronomy and celestial mechanics; Vilppu Piirola added parts to the chapter on observational instruments, and Göran Sandell wrote the part about radio astronomy; chapters on magnitudes, radiation mechanisms and temperature were rewritten by the editors; Markku Poutanen wrote the chapter on the solar system; Juhani Kyröläinen expanded the chapter on stellar spectra; Timo Rahunen rewrote most of the chapters on stellar structure and evolution; Ilkka Tuominen revised the chapter on the Sun; Kalevi Mattila wrote the chapter on interstellar matter; Tapio Markkanen wrote the chapters on star clusters and the Milky Way; Karl Johan Donner wrote the major part of the chapter on galaxies; Mauri Valtonen wrote parts of the galaxy chapter, and, in collaboration with Pekka Teerikorpi, the chapter on cosmology. Finally, the resulting, somewhat inhomogeneous, material was made consistent by the editors.

The English text was written by the editors, who translated parts of the original Finnish text, and rewrote other parts, updating the text and correcting errors found in the original edition. The parts of text set in smaller print are less important material that may still be of interest to the reader.

For the illustrations, we received help from Veikko Sinkkonen, Mirva Vuori and several observatories and individuals mentioned in the figure captions. In the practical work, we were assisted by Arja Kyröläinen and Merja Karsma. A part of the translation was read and corrected by Brian Skiff. We want to express our warmest thanks to all of them.

Financial support was given by the Finnish Ministry of Education and Suomalaisen kirjallisuuden edistämisvarojen valtuuskunta (a foundation promoting Finnish literature), to whom we express our gratitude.

Helsinki *The Editors*
June 1987

Contents

Hubble Space Telescope photo of a galaxy pair called NGC 3314.
Through a chance alignment, a face-on spiral galaxy lies precisely in front of another larger spiral.
(Photo NASA and the Hubble Heritage Team, STScI/AURA)

1. Introduction

1.1 The Role of Astronomy

On a dark, cloudless night, at a distant location far away from the city lights, the starry sky can be seen in all its splendour (Fig. 1.1). It is easy to understand how these thousands of lights in the sky have affected people throughout the ages. After the *Sun*, necessary to all life, the *Moon*, governing the night sky and continuously changing its phases, is the most conspicuous object in the sky. The *stars* seem to stay fixed. Only some relatively bright objects, the *planets*, move with respect to the stars.

The phenomena of the sky aroused people's interest a long time ago. The *Cro Magnon people* made bone engravings 30,000 years ago, which may depict the phases of the Moon. These calendars are the oldest astronomical documents, 25,000 years older than writing.

Agriculture required a good knowledge of the seasons. Religious rituals and prognostication were based

Fig. 1.1. The North America nebula in the constellation of Cygnus. The brightest star on the right is α Cygni or Deneb. (Photo M. Poutanen and H. Virtanen)

Hannu Karttunen et al. (Eds.), *Introduction*.
In: Hannu Karttunen et al. (Eds.), Fundamental Astronomy, 5th Edition. pp. 3–9 (2007)
DOI: 11685739_1 © Springer-Verlag Berlin Heidelberg 2007

on the locations of the celestial bodies. Thus time reckoning became more and more accurate, and people learned to calculate the movements of celestial bodies in advance.

During the rapid development of seafaring, when voyages extended farther and farther from home ports, position determination presented a problem for which astronomy offered a practical solution. Solving these problems of navigation were the most important tasks of astronomy in the 17th and 18th centuries, when the first precise tables on the movements of the planets and on other celestial phenomena were published. The basis for these developments was the discovery of the laws governing the motions of the planets by *Copernicus*, *Tycho Brahe*, *Kepler*, *Galilei* and *Newton*.

Astronomical research has changed man's view of the world from geocentric, anthropocentric conceptions to the modern view of a vast universe where man and the Earth play an insignificant role. Astronomy has taught us the real scale of the nature surrounding us.

Modern astronomy is fundamental science, motivated mainly by man's curiosity, his wish to know more about Nature and the Universe. Astronomy has a central role in forming a scientific view of the world. "A scientific view of the world" means a model of the universe based on observations, thoroughly tested theories and logical reasoning. Observations are always the ultimate test of a model: if the model does not fit the observations, it has to be changed, and this process must not be limited by any philosophical, political or religious conceptions or beliefs.

Fig. 1.2. Although space probes and satellites have gathered remarkable new information, a great majority of astronomical observations is still Earth-based. The most important observatories are usually located at high altitudes far from densely populated areas. One such observatory is on Mt Paranal in Chile, which houses the European VLT telescopes. (Photo ESO)

1.2 Astronomical Objects of Research

Modern astronomy explores the whole Universe and its different forms of matter and energy. Astronomers study the contents of the Universe from the level of elementary particles and molecules (with masses of 10^{-30} kg) to the largest superclusters of galaxies (with masses of 10^{50} kg).

Astronomy can be divided into different branches in several ways. The division can be made according to either the methods or the objects of research.

The Earth (Fig. 1.3) is of interest to astronomy for many reasons. Nearly all observations must be made through the atmosphere, and the phenomena of the upper atmosphere and magnetosphere reflect the state of interplanetary space. The Earth is also the most important object of comparison for planetologists.

The Moon is still studied by astronomical methods, although spacecraft and astronauts have visited its surface and brought samples back to the Earth. To amateur astronomers, the Moon is an interesting and easy object for observations.

In the study of the planets of the solar system, the situation in the 1980's was the same as in lunar exploration 20 years earlier: the surfaces of the planets and their moons have been mapped by fly-bys of spacecraft or by orbiters, and spacecraft have soft-landed on Mars and Venus. This kind of exploration has tremendously added to our knowledge of the conditions on the planets. Continuous monitoring of the planets, however, can still only be made from the Earth, and many bodies in the solar system still await their spacecraft.

The Solar System is governed by the Sun, which produces energy in its centre by nuclear fusion. The Sun is our nearest star, and its study lends insight into conditions on other stars.

Some thousands of stars can be seen with the naked eye, but even a small telescope reveals millions of them. Stars can be classified according to their observed characteristics. A majority are like the Sun; we call them *main sequence stars*. However, some stars are much larger, *giants* or *supergiants*, and some are much smaller, *white dwarfs*. Different types of stars represent different stages of stellar evolution. Most stars are components of *binary* or *multiple systems*, many are *variable*: their brightness is not constant.

Fig. 1.3. The Earth as seen from the Moon. The picture was taken on the first Apollo flight around the Moon, Apollo 8 in 1968. (Photo NASA)

Among the newest objects studied by astronomers are the *compact stars*: *neutron stars* and *black holes*. In them, matter has been so greatly compressed and the gravitational field is so strong that Einstein's general

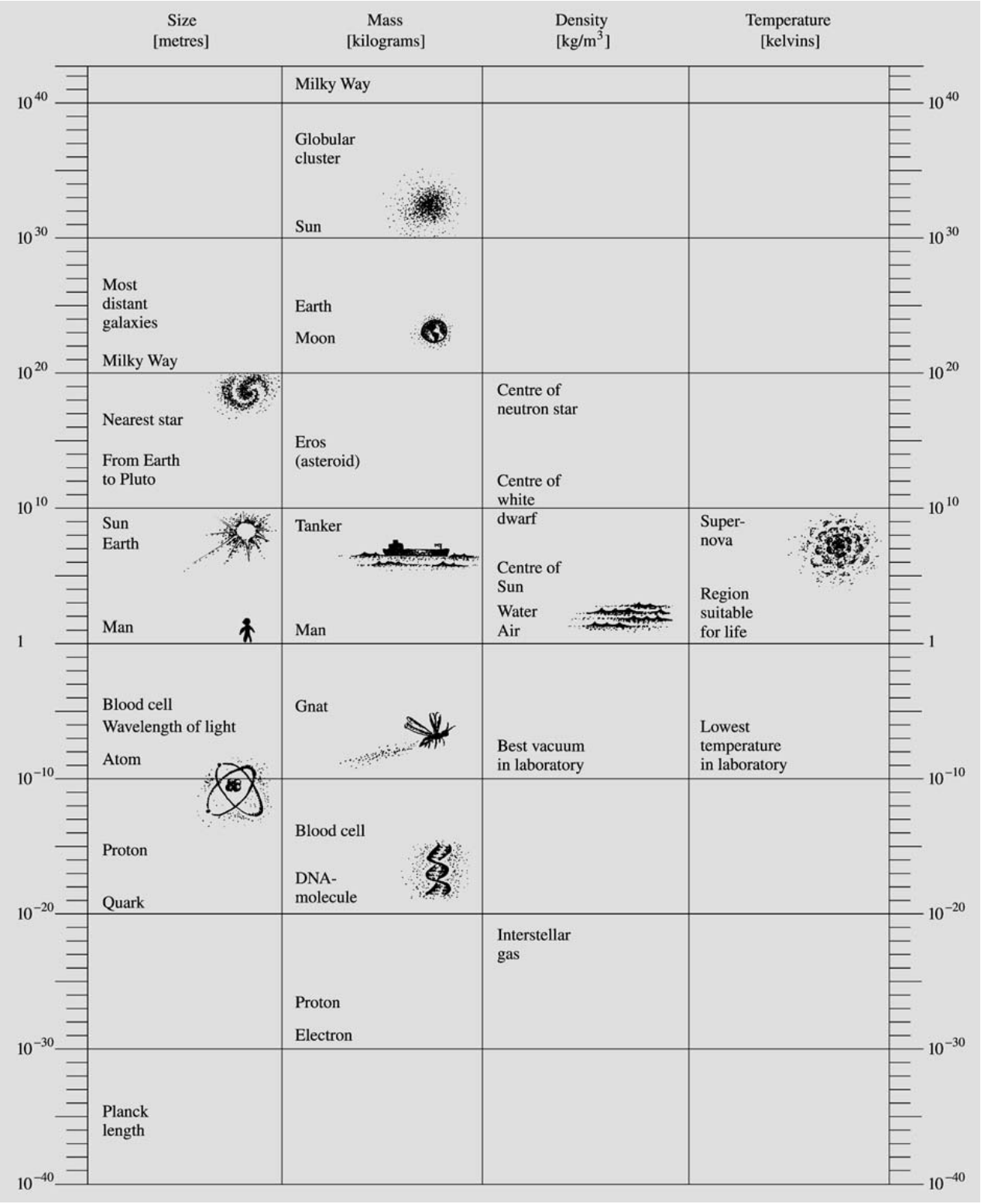


Fig. 1.4. The dimensions of the Universe

theory of relativity must be used to describe matter and space.

Stars are points of light in an otherwise seemingly empty space. Yet interstellar space is not empty, but contains large clouds of *atoms*, *molecules*, *elementary particles* and *dust*. New matter is injected into interstellar space by erupting and exploding stars; at other places, new stars are formed from contracting interstellar clouds.

Stars are not evenly distributed in space, but form concentrations, *clusters of stars*. These consist of stars born near each other, and in some cases, remaining together for billions of years.

The largest concentration of stars in the sky is the *Milky Way*. It is a massive stellar system, a *galaxy*, consisting of over 200 billion stars. All the stars visible to the naked eye belong to the Milky Way. Light travels across our galaxy in 100,000 years.

The Milky Way is not the only galaxy, but one of almost innumerable others. Galaxies often form *clusters of galaxies*, and these clusters can be clumped together into *superclusters*. Galaxies are seen at all distances as far away as our observations reach. Still further out we see *quasars* – the light of the most distant quasars we see now was emitted when the Universe was one-tenth of its present age.

The largest object studied by astronomers is the whole Universe. *Cosmology*, once the domain of theologicians and philosophers, has become the subject of physical theories and concrete astronomical observations.

Among the different branches of research, *spherical*, or positional, *astronomy* studies the coordinate systems on the celestial sphere, their changes and the apparent places of celestial bodies in the sky. *Celestial mechanics* studies the movements of bodies in the solar system, in stellar systems and among the galaxies and clusters of galaxies. *Astrophysics* is concerned with the physical properties of celestial objects; it employs methods of modern physics. It thus has a central position in almost all branches of astronomy (Table 1.1).

Astronomy can be divided into different areas according to the wavelength used in observations. We can

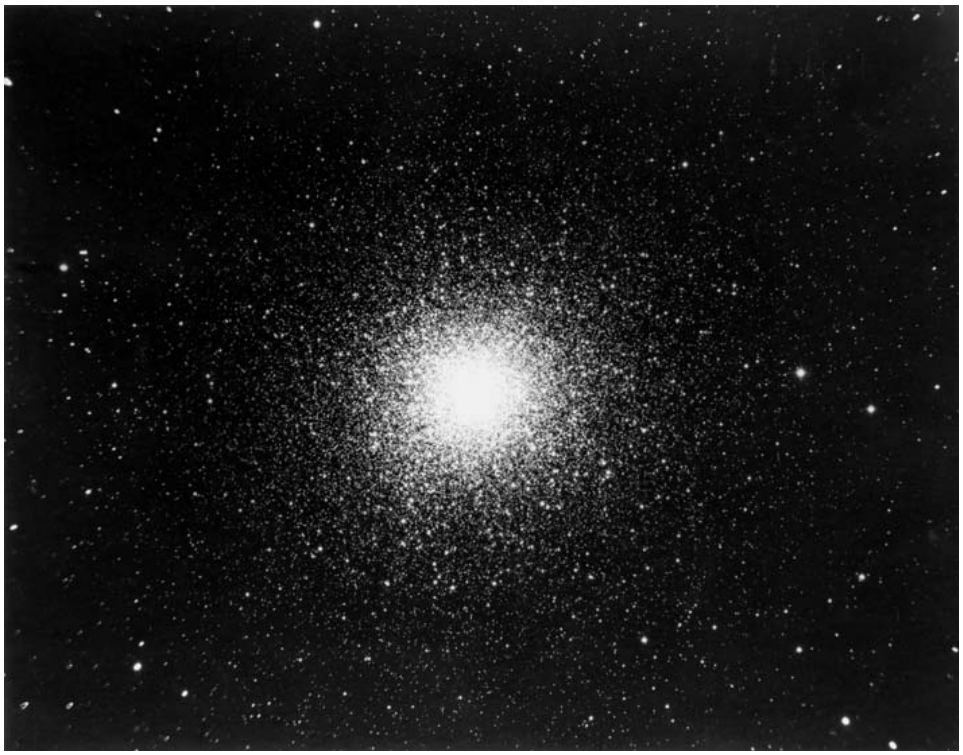


Fig. 1.5. The globular cluster M13. There are over a million stars in the cluster. (Photo Palomar Observatory)

Table 1.1. The share of different branches of astronomy in 1980, 1998 and 2005. For the first two years, the percantage of the number of publications was estimated from the printed pages of *Astronomy and Astrophysics Abstracts*, published by the Astronomische Rechen-Institut, Heidelberg. The publication of the series was discontinued in 2000, and for 2005, an estimate was made from the Smithsonian/NASA Astrophysics Data System (ADS) Abstract Service in the net. The difference between 1998 and 2005 may reflect different methods of classification, rather than actual changes in the direction of research.

Branch of Astronomy	Percentage of publications in the year		
	1980	1998	2005
Astronomical instruments and techniques	6	6	8
Positional astronomy, celestial mechanics	4	2	5
Space research	2	1	9
Theoretical astrophysics	10	13	6
Sun	8	8	8
Earth	5	4	3
Planetary system	16	9	11
Interstellar matter, nebulae	7	6	5
Radio sources, X-ray sources, cosmic rays	9	5	12
Stellar systems, Galaxy, extragalactic objects, cosmology	14	29	22

speak of radio, infrared, optical, ultraviolet, X-ray or gamma astronomy, depending on which wavelengths of the electromagnetic spectrum are used. In the future, neutrinos and gravitational waves may also be observed.

1.3 The Scale of the Universe

The masses and sizes of astronomical objects are usually enormously large. But to understand their properties, the smallest parts of matter, molecules, atoms and elementary particles, must be studied. The densities, temperatures and magnetic fields in the Universe vary within much larger limits than can be reached in laboratories on the Earth.

The greatest natural density met on the Earth is $22{,}500\,\mathrm{kg\,m^{-3}}$ (osmium), while in neutron stars densities of the order of $10^{18}\,\mathrm{kg\,m^{-3}}$ are possible. The density in the best vacuum achieved on the Earth is only $10^{-9}\,\mathrm{kg\,m^{-3}}$, but in interstellar space the density

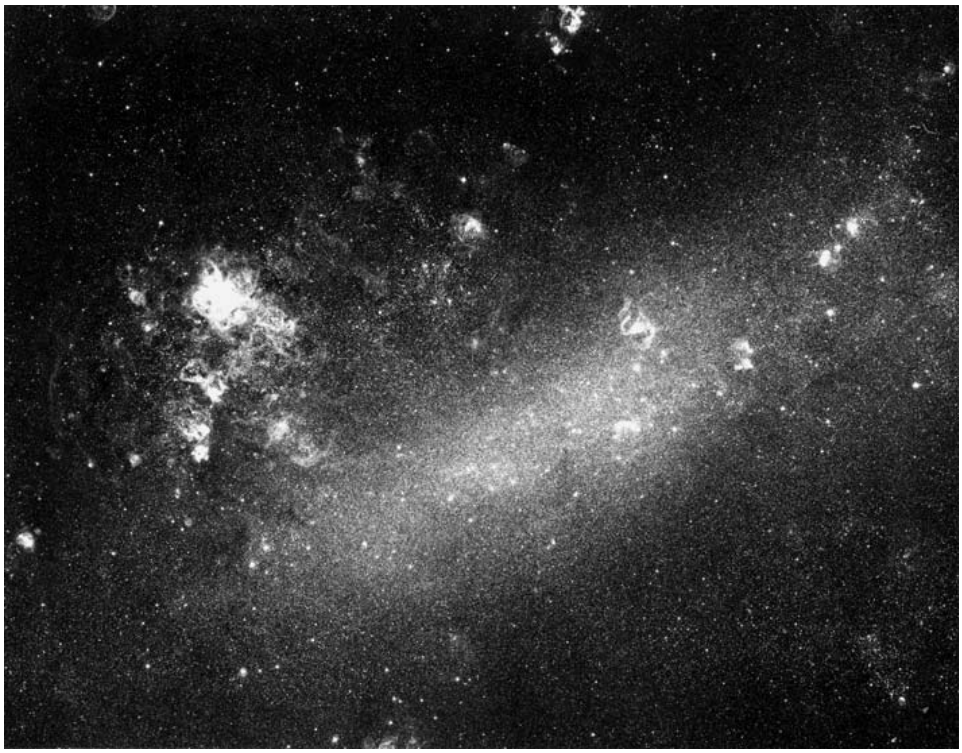


Fig. 1.6. The Large Magellanic Cloud, our nearest neighbour galaxy. (Photo National Optical Astronomy Observatories, Cerro Tololo Inter-American Observatory)

of the gas may be $10^{-21}\,\mathrm{kg\,m^{-3}}$ or even less. Modern accelerators can give particles energies of the order of 10^{12} electron volts (eV). Cosmic rays coming from the sky may have energies of over 10^{20} eV.

It has taken man a long time to grasp the vast dimensions of space. Already *Hipparchos* in the second century B.C. obtained a reasonably correct value for the distance of the Moon. The scale of the solar system was established together with the heliocentric system in the 17th century. The first measurements of stellar distances were made in the 1830's, and the distances to the galaxies were determined only in the 1920's.

We can get some kind of picture of the distances involved (Fig. 1.4) by considering the time required for light to travel from a source to the retina of the human eye. It takes 8 minutes for light to travel from the Sun, $5\frac{1}{2}$ hours from Pluto and 4 years from the nearest star. We cannot see the centre of the Milky Way, but the many globular clusters around the Milky Way are at approximately similar distances. It takes about 20,000 years for the light from the globular cluster of Fig. 1.5 to reach the Earth. It takes 150,000 years to travel the distance from the nearest galaxy, the Magellanic Cloud seen on the southern sky (Fig. 1.6). The photons that we see now started their voyage when Neanderthal Man lived on the Earth. The light coming from the *Andromeda Galaxy* in the northern sky originated 2 million years ago. Around the same time the first actual human using tools, *Homo habilis*, appeared. The most distant objects known, the quasars, are so far away that their radiation, seen on the Earth now, was emitted long before the Sun or the Earth were born.

2. Spherical Astronomy

Spherical astronomy is a science studying astronomical coordinate frames, directions and apparent motions of celestial objects, determination of position from astronomical observations, observational errors, etc. We shall concentrate mainly on astronomical coordinates, apparent motions of stars and time reckoning. Also, some of the most important star catalogues will be introduced.

For simplicity we will assume that the observer is always on the northern hemisphere. Although all definitions and equations are easily generalized for both hemispheres, this might be unnecessarily confusing. In spherical astronomy all angles are usually expressed in degrees; we will also use degrees unless otherwise mentioned.

2.1 Spherical Trigonometry

For the coordinate transformations of spherical astronomy, we need some mathematical tools, which we present now.

If a plane passes through the centre of a sphere, it will split the sphere into two identical hemispheres along a circle called a *great circle* (Fig. 2.1). A line perpendicular to the plane and passing through the centre of the sphere intersects the sphere at the *poles* P and P'. If a sphere is intersected by a plane not containing the centre, the intersection curve is a *small circle*. There is exactly one great circle passing through two given points Q and Q' on a sphere (unless these points are antipodal, in which case all circles passing through both of them are great circles). The arc QQ' of this great circle is the shortest path on the surface of the sphere between these points.

A *spherical triangle* is not just any three-cornered figure lying on a sphere; its sides must be arcs of great circles. The spherical triangle ABC in Fig. 2.2 has the arcs AB, BC and AC as its sides. If the radius of the sphere is r, the length of the arc AB is

$$|AB| = rc\,, \quad [c] = \mathrm{rad}\,,$$

where c is the angle subtended by the arc AB as seen from the centre. This angle is called the *central angle* of the side AB. Because lengths of sides and central

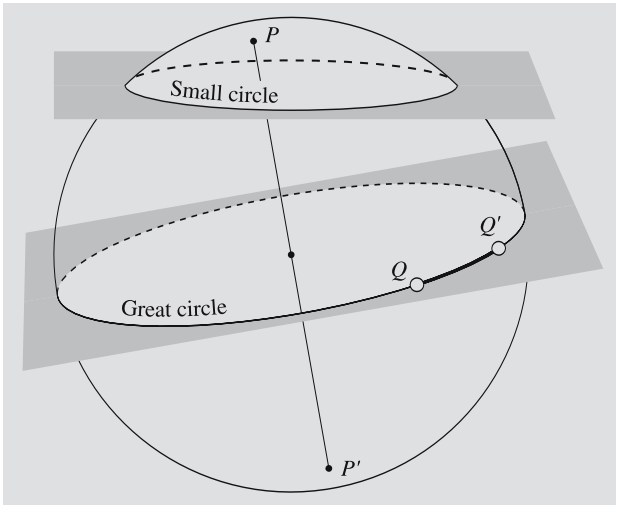


Fig. 2.1. A great circle is the intersection of a sphere and a plane passing through its centre. P and P' are the poles of the great circle. The shortest path from Q to Q' follows the great circle

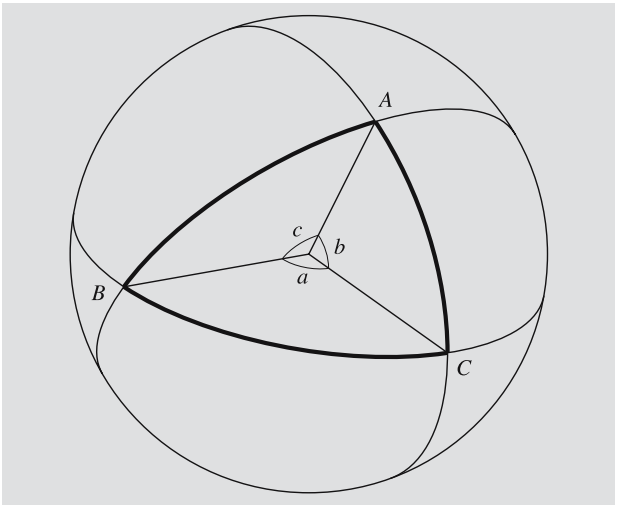


Fig. 2.2. A spherical triangle is bounded by three arcs of great circles, AB, BC and CA. The corresponding central angles are c, a, and b

Hannu Karttunen et al. (Eds.), *Spherical Astronomy*.
In: Hannu Karttunen et al. (Eds.), Fundamental Astronomy, 5th Edition. pp. 11–45 (2007)
DOI: 11685739_2 © Springer-Verlag Berlin Heidelberg 2007

angles correspond to each other in a unique way, it is customary to give the central angles instead of the sides. In this way, the radius of the sphere does not enter into the equations of spherical trigonometry. An angle of a spherical triangle can be defined as the angle between the tangents of the two sides meeting at a vertex, or as the dihedral angle between the planes intersecting the sphere along these two sides. We denote the angles of a spherical triangle by capital letters (A, B, C) and the opposing sides, or, more correctly, the corresponding central angles, by lowercase letters (a, b, c).

The sum of the angles of a spherical triangle is always greater than 180 degrees; the excess

$$E = A + B + C - 180^\circ \tag{2.1}$$

is called the *spherical excess*. It is not a constant, but depends on the triangle. Unlike in plane geometry, it is not enough to know two of the angles to determine the third one. The area of a spherical triangle is related to the spherical excess in a very simple way:

$$\text{Area} = Er^2 , \quad [E] = \text{rad} . \tag{2.2}$$

This shows that the spherical excess equals the solid angle in steradians (see Appendix A.1), subtended by the triangle as seen from the centre.

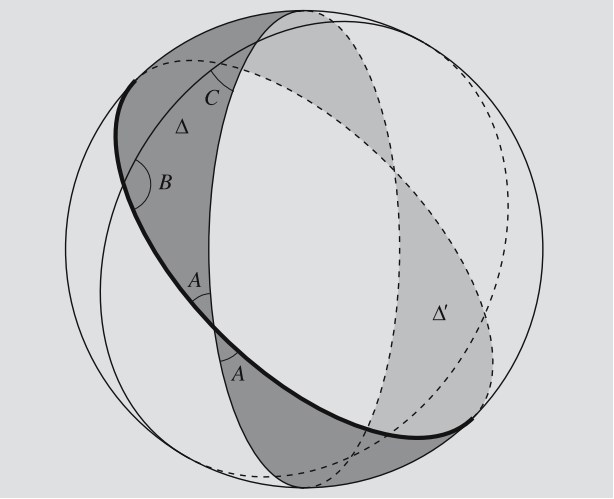


Fig. 2.3. If the sides of a spherical triangle are extended all the way around the sphere, they form another triangle Δ', antipodal and equal to the original triangle Δ. The shaded area is the slice $S(A)$

To prove (2.2), we extend all sides of the triangle Δ to great circles (Fig. 2.3). These great circles will form another triangle Δ', congruent with Δ but antipodal to it. If the angle A is expressed in radians, the area of the slice $S(A)$ bounded by the two sides of A (the shaded area in Fig. 2.3) is obviously $2A/2\pi = A/\pi$ times the area of the sphere, $4\pi r^2$. Similarly, the slices $S(B)$ and $S(C)$ cover fractions B/π and C/π of the whole sphere.

Together, the three slices cover the whole surface of the sphere, the equal triangles Δ and Δ' belonging to every slice, and each point outside the triangles, to exactly one slice. Thus the area of the slices $S(A)$, $S(B)$ and $S(C)$ equals the area of the sphere plus four times the area of Δ, $\mathcal{A}(\Delta)$:

$$\frac{A+B+C}{\pi} 4\pi r^2 = 4\pi r^2 + 4\mathcal{A}(\Delta) ,$$

whence

$$\mathcal{A}(\Delta) = (A + B + C - \pi) r^2 = Er^2 .$$

As in the case of plane triangles, we can derive relationships between the sides and angles of spherical triangles. The easiest way to do this is by inspecting certain coordinate transformations.

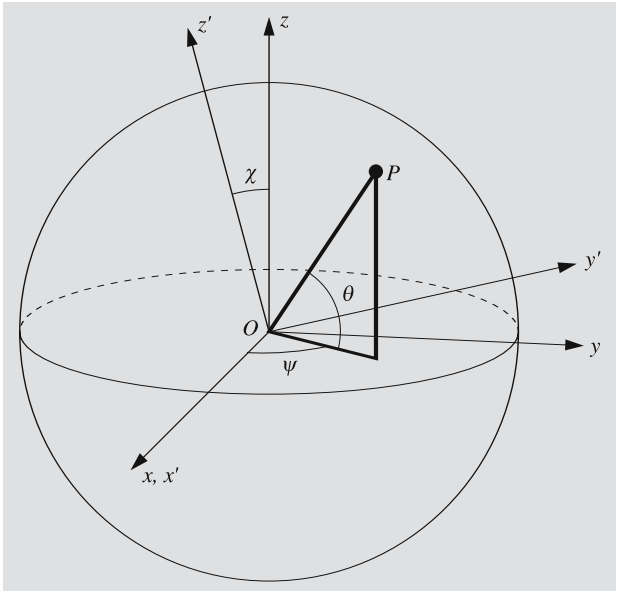


Fig. 2.4. The location of a point P on the surface of a unit sphere can be expressed by rectangular xyz coordinates or by two angles, ψ and θ. The $x'y'z'$ frame is obtained by rotating the xyz frame around its x axis by an angle χ

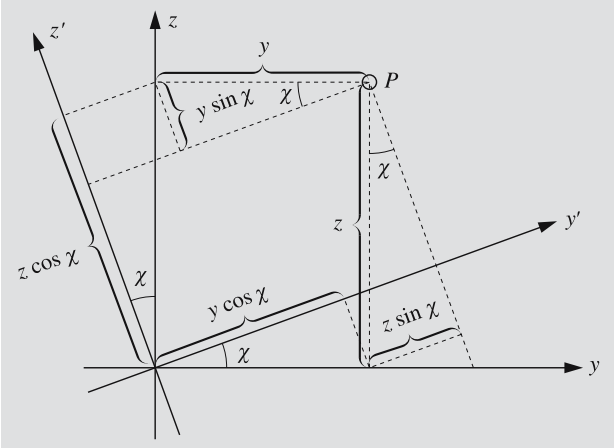


Fig. 2.5. The coordinates of the point P in the rotated frame are $x' = x$, $y' = y\cos\chi + z\sin\chi$, $z' = z\cos\chi - y\sin\chi$

Suppose we have two rectangular coordinate frames $Oxyz$ and $Ox'y'z'$ (Fig. 2.4), such that the $x'y'z'$ frame is obtained from the xyz frame by rotating it around the x axis by an angle χ.

The position of a point P on a unit sphere is uniquely determined by giving two angles. The angle ψ is measured counterclockwise from the positive x axis along the xy plane; the other angle θ tells the angular distance from the xy plane. In an analogous way, we can define the angles ψ' and θ', which give the position of the point P in the $x'y'z'$ frame. The rectangular coordinates of the point P as functions of these angles are:

$$\begin{aligned} x &= \cos\psi\cos\theta\,, & x' &= \cos\psi'\cos\theta'\,,\\ y &= \sin\psi\cos\theta\,, & y' &= \sin\psi'\cos\theta'\,,\\ z &= \sin\theta, & z' &= \sin\theta'. \end{aligned} \tag{2.3}$$

We also know that the dashed coordinates are obtained from the undashed ones by a rotation in the yz plane (Fig. 2.5):

$$\begin{aligned} x' &= x\,,\\ y' &= y\cos\chi + z\sin\chi\,,\\ z' &= -y\sin\chi + z\cos\chi\,. \end{aligned} \tag{2.4}$$

By substituting the expressions of the rectangular coordinates (2.3) into (2.4), we have

$$\begin{aligned} \cos\psi'\cos\theta' &= \cos\psi\cos\theta\,,\\ \sin\psi'\cos\theta' &= \sin\psi\cos\theta\cos\chi + \sin\theta\sin\chi\,,\\ \sin\theta' &= -\sin\psi\cos\theta\sin\chi + \sin\theta\cos\chi\,. \end{aligned} \tag{2.5}$$

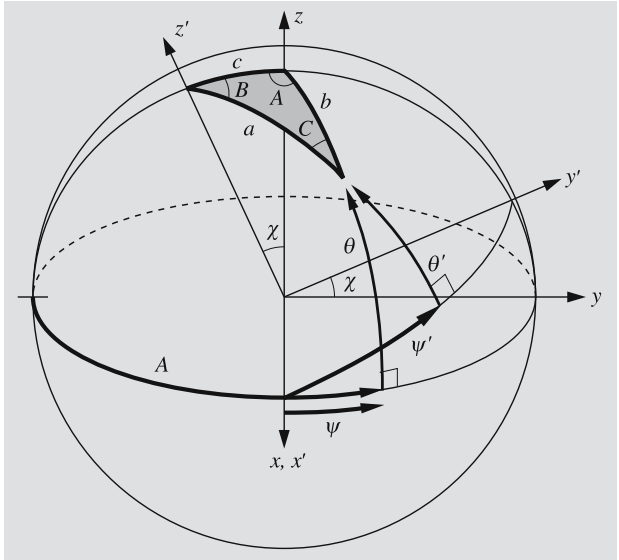


Fig. 2.6. To derive triangulation formulas for the spherical triangle ABC, the spherical coordinates ψ, θ, ψ' and θ' of the vertex C are expressed in terms of the sides and angles of the triangle

In fact, these equations are quite sufficient for all coordinate transformations we may encounter. However, we shall also derive the usual equations for spherical triangles. To do this, we set up the coordinate frames in a suitable way (Fig. 2.6). The z axis points towards the vertex A and the z' axis, towards B. Now the vertex C corresponds to the point P in Fig. 2.4. The angles ψ, θ, ψ', θ' and χ can be expressed in terms of the angles and sides of the spherical triangle:

$$\begin{aligned} &\psi = A - 90^\circ\,, \quad \theta = 90^\circ - b\,,\\ &\psi' = 90^\circ - B\,, \quad \theta' = 90^\circ - a\,, \quad \chi = c\,. \end{aligned} \tag{2.6}$$

Substitution into (2.5) gives

$$\begin{aligned} &\cos(90^\circ - B)\cos(90^\circ - a)\\ &\quad = \cos(A - 90^\circ)\cos(90^\circ - b)\,,\\ &\sin(90^\circ - B)\cos(90^\circ - a)\\ &\quad = \sin(A - 90^\circ)\cos(90^\circ - b)\cos c\\ &\qquad + \sin(90^\circ - b)\sin c\,,\\ &\sin(90^\circ - a)\\ &\quad = -\sin(A - 90^\circ)\cos(90^\circ - b)\sin c\\ &\qquad + \sin(90^\circ - b)\cos c\,, \end{aligned}$$

or

$$\begin{aligned} \sin B \sin a &= \sin A \sin b \,, \\ \cos B \sin a &= -\cos A \sin b \cos c + \cos b \sin c \,, \\ \cos a &= \cos A \sin b \sin c + \cos b \cos c \,. \end{aligned} \tag{2.7}$$

Equations for other sides and angles are obtained by cyclic permutations of the sides a, b, c and the angles A, B, C. For instance, the first equation also yields

$$\begin{aligned} \sin C \sin b &= \sin B \sin c \,, \\ \sin A \sin c &= \sin C \sin a \,. \end{aligned}$$

All these variations of the *sine formula* can be written in an easily remembered form:

$$\frac{\sin a}{\sin A} = \frac{\sin b}{\sin B} = \frac{\sin c}{\sin C} \,. \tag{2.8}$$

If we take the limit, letting the sides a, b and c shrink to zero, the spherical triangle becomes a plane triangle. If all angles are expressed in radians, we have approximately

$$\sin a \approx a \,, \quad \cos a \approx 1 - \frac{1}{2}a^2 \,.$$

Substituting these approximations into the sine formula, we get the familiar sine formula of plane geometry:

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C} \,.$$

The second equation in (2.7) is the *sine-cosine formula*, and the corresponding plane formula is a trivial one:

$$c = b \cos A + a \cos B \,.$$

This is obtained by substituting the approximations of sine and cosine into the sine-cosine formula and ignoring all quadratic and higher-order terms. In the same way we can use the third equation in (2.7), the *cosine formula*, to derive the planar cosine formula:

$$a^2 = b^2 + c^2 - 2bc \cos A \,.$$

2.2 The Earth

A position on the Earth is usually given by two spherical coordinates (although in some calculations rectangular or other coordinates may be more convenient). If necessary, also a third coordinate, e. g. the distance from the centre, can be used.

The reference plane is the *equatorial plane*, perpendicular to the rotation axis and intersecting the surface of the Earth along the *equator*. Small circles parallel to the equator are called *parallels of latitude*. Semicircles from pole to pole are *meridians*. The geographical *longitude* is the angle between the meridian and the zero meridian passing through Greenwich Observatory. We shall use positive values for longitudes east of Greenwich and negative values west of Greenwich. Sign convention, however, varies, and negative longitudes are not used in maps; so it is usually better to say explicitly whether the longitude is east or west of Greenwich.

The *latitude* is usually supposed to mean the *geographical latitude*, which is the angle between the plumb line and the equatorial plane. The latitude is positive in the northern hemisphere and negative in the southern one. The geographical latitude can be determined by astronomical observations (Fig. 2.7): the altitude of the celestial pole measured from the hori-

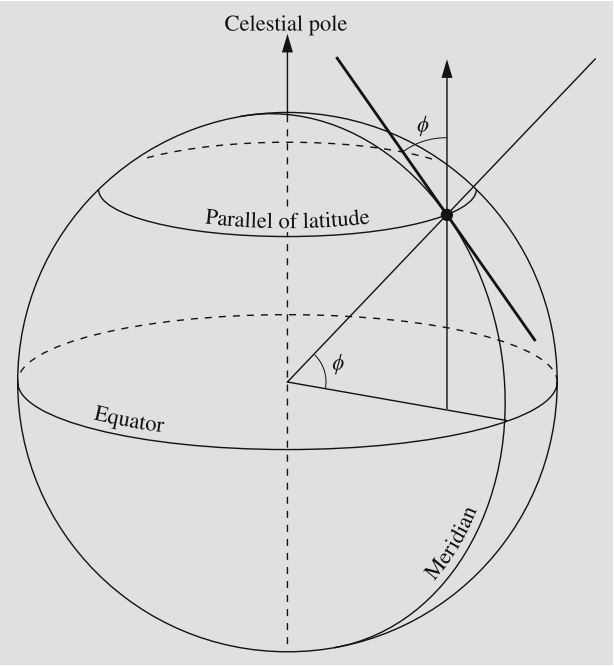


Fig. 2.7. The latitude ϕ is obtained by measuring the altitude of the celestial pole. The celestial pole can be imagined as a point at an infinite distance in the direction of the Earth's rotation axis

zon equals the geographical latitude. (The celestial pole is the intersection of the rotation axis of the Earth and the infinitely distant celestial sphere; we shall return to these concepts a little later.)

Because the Earth is rotating, it is slightly flattened. The exact shape is rather complicated, but for most purposes it can by approximated by an oblate spheroid, the short axis of which coincides with the rotation axis (Sect. 7.5). In 1979 the International Union of Geodesy and Geophysics (IUGG) adopted the Geodetic Reference System 1980 (GRS-80), which is used when global reference frames fixed to the Earth are defined. The GRS-80 reference ellipsoid has the following dimensions:

equatorial radius	$a = 6{,}378{,}137$ m,
polar radius	$b = 6{,}356{,}752$ m,
flattening	$f = (a-b)/a$
	$= 1/298.25722210.$

The shape defined by the surface of the oceans, called the *geoid*, differs from this spheroid at most by about 100 m.

The angle between the equator and the normal to the ellipsoid approximating the true Earth is called the *geodetic latitude*. Because the surface of a liquid (like an ocean) is perpendicular to the plumb line, the geodetic and geographical latitudes are practically the same.

Because of the flattening, the plumb line does not point to the centre of the Earth except at the poles and on the equator. An angle corresponding to the ordinary spherical coordinate (the angle between the equator and the line from the centre to a point on the surface), the *geocentric latitude* ϕ' is therefore a little smaller than the geographic latitude ϕ (Fig. 2.8).

We now derive an equation between the geographic latitude ϕ and geocentric latitude ϕ', assuming the Earth is an oblate spheroid and the geographic and geodesic latitudes are equal. The equation of the meridional ellipse is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \, .$$

The direction of the normal to the ellipse at a point (x, y) is given by

$$\tan\phi = -\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{a^2}{b^2}\frac{y}{x} \, .$$

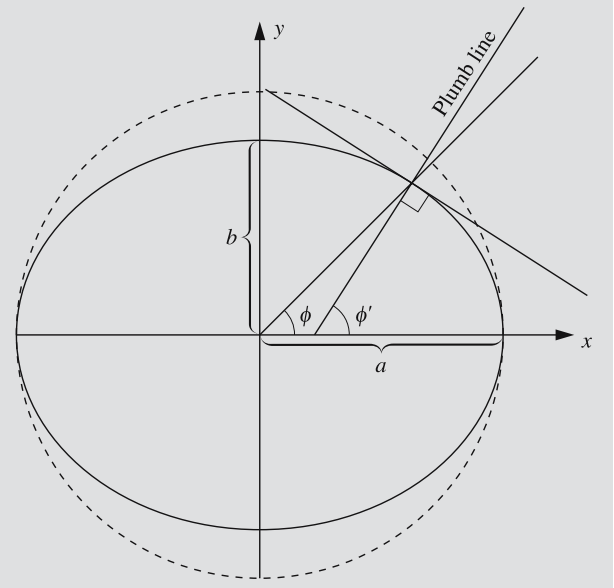


Fig. 2.8. Due to the flattening of the Earth, the geographic latitude ϕ and geocentric latitude ϕ' are different

The geocentric latitude is obtained from

$$\tan\phi' = y/x \, .$$

Hence

$$\tan\phi' = \frac{b^2}{a^2}\tan\phi = (1-e^2)\tan\phi \, , \tag{2.9}$$

where

$$e = \sqrt{1 - b^2/a^2}$$

is the eccentricity of the ellipse. The difference $\Delta\phi = \phi - \phi'$ has a maximum $11.5'$ at the latitude 45°.

Since the coordinates of celestial bodies in astronomical almanacs are given with respect to the centre of the Earth, the coordinates of nearby objects must be corrected for the difference in the position of the observer, if high accuracy is required. This means that one has to calculate the *topocentric* coordinates, centered at the observer. The easiest way to do this is to use rectangular coordinates of the object and the observer (Example 2.5).

One arc minute along a meridian is called a *nautical mile*. Since the radius of curvature varies with latitude, the length of the nautical mile so defined would depend on the latitude. Therefore one nautical mile has been

defined to be equal to one minute of arc at $\phi = 45°$, whence 1 nautical mile = 1852 m.

2.3 The Celestial Sphere

The ancient universe was confined within a finite spherical shell. The stars were fixed to this shell and thus were all equidistant from the Earth, which was at the centre of the spherical universe. This simple model is still in many ways as useful as it was in antiquity: it helps us to easily understand the diurnal and annual motions of stars, and, more important, to predict these motions in a relatively simple way. Therefore we will assume for the time being that all the stars are located on the surface of an enormous sphere and that we are at its centre. Because the radius of this celestial sphere is practically infinite, we can neglect the effects due to the changing position of the observer, caused by the rotation and orbital motion of the Earth. These effects will be considered later in Sects. 2.9 and 2.10.

Since the distances of the stars are ignored, we need only two coordinates to specify their directions. Each coordinate frame has some fixed reference plane passing through the centre of the celestial sphere and dividing the sphere into two hemispheres along a great circle. One of the coordinates indicates the angular distance from this reference plane. There is exactly one great circle going through the object and intersecting this plane perpendicularly; the second coordinate gives the angle between that point of intersection and some fixed direction.

2.4 The Horizontal System

The most natural coordinate frame from the observer's point of view is the *horizontal frame* (Fig. 2.9). Its reference plane is the tangent plane of the Earth passing through the observer; this horizontal plane intersects the celestial sphere along the *horizon*. The point just above the observer is called the *zenith* and the antipodal point below the observer is the *nadir*. (These two points are the poles corresponding to the horizon.) Great circles through the zenith are called *verticals*. All verticals intersect the horizon perpendicularly.

By observing the motion of a star over the course of a night, an observer finds out that it follows a track like one of those in Fig. 2.9. Stars rise in the east, reach their highest point, or *culminate*, on the vertical NZS, and set in the west. The vertical NZS is called the *meridian*. North and south directions are defined as the intersections of the meridian and the horizon.

One of the horizontal coordinates is the *altitude* or *elevation*, a, which is measured from the horizon along the vertical passing through the object. The altitude lies in the range $[-90°, +90°]$; it is positive for objects above the horizon and negative for the objects below the horizon. The *zenith distance*, or the angle between

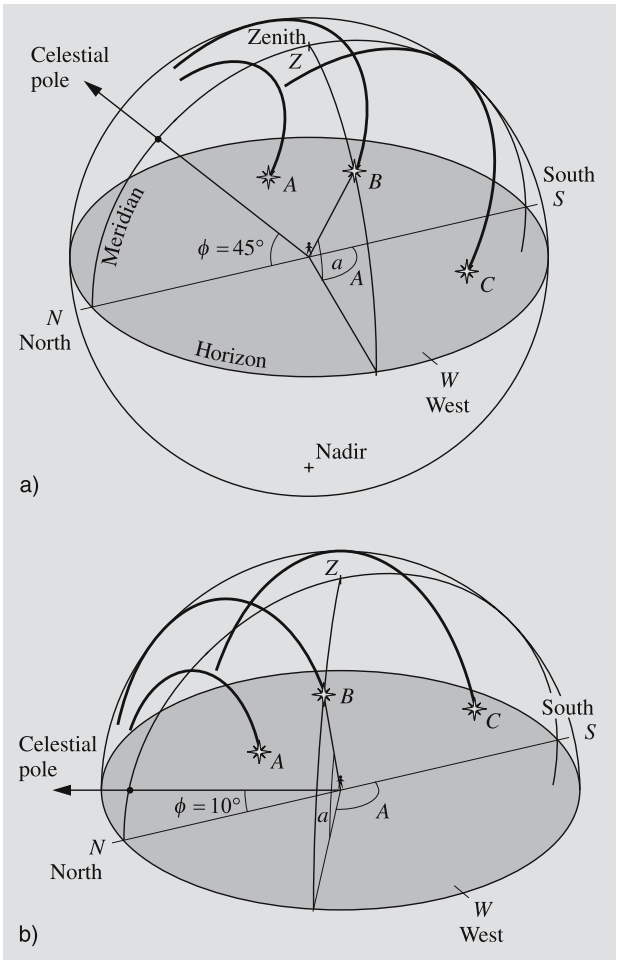


Fig. 2.9. (**a**) The apparent motions of stars during a night as seen from latitude $\phi = 45°$. (**b**) The same stars seen from latitude $\phi = 10°$

the object and the zenith, is obviously

$$z = 90^\circ - a \, . \tag{2.10}$$

The second coordinate is the *azimuth*, A; it is the angular distance of the vertical of the object from some fixed direction. Unfortunately, in different contexts, different fixed directions are used; thus it is always advisable to check which definition is employed. The azimuth is usually measured from the north or south, and though clockwise is the preferred direction, counterclockwise measurements are also occasionally made. In this book we have adopted a fairly common astronomical convention, measuring the azimuth *clockwise* from the *south*. Its values are usually normalized between 0° and 360°.

In Fig. 2.9a we can see the altitude and azimuth of a star B at some instant. As the star moves along its daily track, both of its coordinates will change. Another difficulty with this coordinate frame is its local character. In Fig. 2.9b we have the same stars, but the observer is now further south. We can see that the coordinates of the same star at the same moment are different for different observers. Since the horizontal coordinates are time and position dependent, they cannot be used, for instance, in star catalogues.

2.5 The Equatorial System

The direction of the rotation axis of the Earth remains almost constant and so does the equatorial plane perpendicular to this axis. Therefore the equatorial plane is a suitable reference plane for a coordinate frame that has to be independent of time and the position of the observer.

The intersection of the celestial sphere and the equatorial plane is a great circle, which is called the *equator of the celestial sphere*. The north pole of the celestial sphere is one of the poles corresponding to this great circle. It is also the point in the northern sky where the extension of the Earth's rotational axis meets the celestial sphere. The celestial north pole is at a distance of about one degree (which is equivalent to two full moons) from the moderately bright star Polaris. The meridian always passes through the north pole; it is divided by the pole into north and south meridians.

Fig. 2.10. At night, stars seem to revolve around the celestial pole. The altitude of the pole from the horizon equals the latitude of the observer. (Photo Pekka Parviainen)

The angular separation of a star from the equatorial plane is not affected by the rotation of the Earth. This angle is called the *declination* δ.

Stars seem to revolve around the pole once every day (Fig. 2.10). To define the second coordinate, we must again agree on a fixed direction, unaffected by the Earth's rotation. From a mathematical point of view, it does not matter which point on the equator is selected. However, for later purposes, it is more appropriate to employ a certain point with some valuable properties, which will be explained in the next section. This point is called the *vernal equinox*. Because it used to be in the constellation Aries (the Ram), it is also called the first point of Aries ant denoted by the sign of Aries, ♈. Now we can define the second coordinate as the angle from

the vernal equinox measured along the equator. This angle is the *right ascension* α (or R.A.) of the object, measured counterclockwise from ♈.

Since declination and right ascension are independent of the position of the observer and the motions of the Earth, they can be used in star maps and catalogues. As will be explained later, in many telescopes one of the axes (the hour axis) is parallel to the rotation axis of the Earth. The other axis (declination axis) is perpendicular to the hour axis. Declinations can be read immediately on the declination dial of the telescope. But the zero point of the right ascension seems to move in the sky, due to the diurnal rotation of the Earth. So we cannot use the right ascension to find an object unless we know the direction of the vernal equinox.

Since the south meridian is a well-defined line in the sky, we use it to establish a local coordinate corresponding to the right ascension. The *hour angle* is measured clockwise from the meridian. The hour angle of an object is not a constant, but grows at a steady rate, due to the Earth's rotation. The hour angle of the vernal equinox is called the *sidereal time* Θ. Figure 2.11 shows that for any object,

$$\Theta = h + \alpha \,, \tag{2.11}$$

where h is the object's hour angle and α its right ascension.

Since hour angle and sidereal time change with time at a constant rate, it is practical to express them in units of time. Also the closely related right ascension is customarily given in time units. Thus 24 hours equals 360 degrees, 1 hour = 15 degrees, 1 minute of time = 15 minutes of arc, and so on. All these quantities are in the range [0 h, 24 h).

In practice, the sidereal time can be readily determined by pointing the telescope to an easily recognisable star and reading its hour angle on the hour angle dial of the telescope. The right ascension found in a catalogue is then added to the hour angle, giving the sidereal time at the moment of observation. For any other time, the sidereal time can be evaluated by adding the time elapsed since the observation. If we want to be accurate, we have to use a sidereal clock to measure time intervals. A sidereal clock runs 3 min 56.56 s fast a day as compared with an ordinary solar time clock:

$$\begin{aligned} &24\text{ h solar time} \\ &\quad = 24\text{ h }3\text{ min }56.56\text{ s sidereal time}\,. \end{aligned} \tag{2.12}$$

The reason for this is the orbital motion of the Earth: stars seem to move faster than the Sun across the sky; hence, a sidereal clock must run faster. (This is further discussed in Sect. 2.13.)

Transformations between the horizontal and equatorial frames are easily obtained from spherical

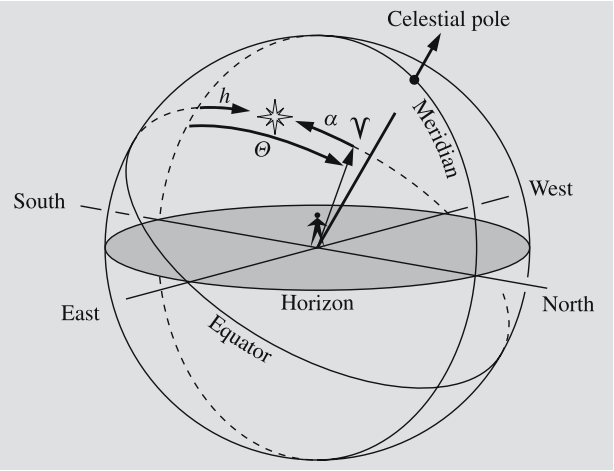


Fig. 2.11. The sidereal time Θ (the hour angle of the vernal equinox) equals the hour angle plus right ascension of any object

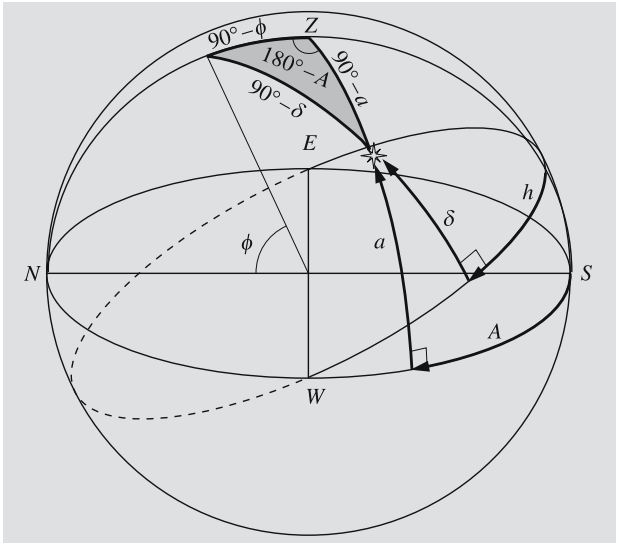


Fig. 2.12. The nautical triangle for deriving transformations between the horizontal and equatorial frames

trigonometry. Comparing Figs. 2.6 and 2.12, we find that we must make the following substitutions into (2.5):

$$\begin{aligned} &\psi = 90^\circ - A\,, \quad \theta = a\,, \\ &\psi' = 90^\circ - h\,, \quad \theta' = \delta\,, \quad \chi = 90^\circ - \phi\,. \end{aligned} \tag{2.13}$$

The angle ϕ in the last equation is the altitude of the celestial pole, or the latitude of the observer. Making the substitutions, we get

$$\begin{aligned} \sin h \cos\delta &= \sin A \cos a\,, \\ \cos h \cos\delta &= \cos A \cos a \sin\phi + \sin a \cos\phi\,, \\ \sin\delta &= -\cos A \cos a \cos\phi + \sin a \sin\phi\,. \end{aligned} \tag{2.14}$$

The inverse transformation is obtained by substituting

$$\begin{aligned} &\psi = 90^\circ - h\,, \quad \theta = \delta\,, \\ &\psi' = 90^\circ - A\,, \quad \theta' = a\,, \quad \chi = -(90^\circ - \phi)\,, \end{aligned} \tag{2.15}$$

whence

$$\begin{aligned} \sin A \cos a &= \sin h \cos\delta\,, \\ \cos A \cos a &= \cos h \cos\delta \sin\phi - \sin\delta \cos\phi\,, \\ \sin a &= \cos h \cos\delta \cos\phi + \sin\delta \sin\phi\,. \end{aligned} \tag{2.16}$$

Since the altitude and declination are in the range $[-90^\circ, +90^\circ]$, it suffices to know the sine of one of these angles to determine the other angle unambiguously. Azimuth and right ascension, however, can have any value from 0° to 360° (or from 0 h to 24 h), and to solve for them, we have to know both the sine and cosine to choose the correct quadrant.

The altitude of an object is greatest when it is on the south meridian (the great circle arc between the celestial poles containing the zenith). At that moment (called *upper culmination*, or *transit*) its hour angle is 0 h. At the *lower culmination* the hour angle is $h = 12$ h. When $h = 0$ h, we get from the last equation in (2.16)

$$\begin{aligned} \sin a &= \cos\delta \cos\phi + \sin\delta \sin\phi \\ &= \cos(\phi - \delta) = \sin(90^\circ - \phi + \delta)\,. \end{aligned}$$

Thus the altitude at the upper culmination is

$$a_{\max} = \begin{cases} 90^\circ - \phi + \delta\,, & \text{if the object culminates south of zenith}\,, \\ 90^\circ + \phi - \delta\,, & \text{if the object culminates north of zenith}\,. \end{cases} \tag{2.17}$$

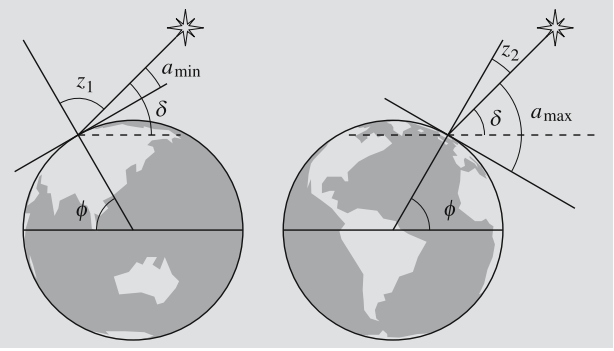


Fig. 2.13. The altitude of a circumpolar star at upper and lower culmination

The altitude is positive for objects with $\delta > \phi - 90^\circ$. Objects with declinations less than $\phi - 90^\circ$ can never be seen at the latitude ϕ. On the other hand, when $h = 12$ h we have

$$\begin{aligned} \sin a &= -\cos\delta \cos\phi + \sin\delta \sin\phi \\ &= -\cos(\delta + \phi) = \sin(\delta + \phi - 90^\circ)\,, \end{aligned}$$

and the altitude at the lower culmination is

$$a_{\min} = \delta + \phi - 90^\circ\,. \tag{2.18}$$

Stars with $\delta > 90^\circ - \phi$ will never set. For example, in Helsinki ($\phi \approx 60^\circ$), all stars with a declination higher than 30° are such *circumpolar* stars. And stars with a declination less than -30° can never be observed there.

We shall now study briefly how the (α, δ) frame can be established by observations. Suppose we observe a circumpolar star at its upper and lower culmination (Fig. 2.13). At the upper transit, its altitude is $a_{\max} = 90^\circ - \phi + \delta$ and at the lower transit, $a_{\min} = \delta + \phi - 90^\circ$. Eliminating the latitude, we get

$$\delta = \frac{1}{2}(a_{\min} + a_{\max})\,. \tag{2.19}$$

Thus we get the same value for the declination, independent of the observer's location. Therefore we can use it as one of the absolute coordinates. From the same observations, we can also determine the direction of the celestial pole as well as the latitude of the observer. After these preparations, we can find the declination of any object by measuring its distance from the pole.

The equator can be now defined as the great circle all of whose points are at a distance of 90° from the

pole. The zero point of the second coordinate (right ascension) can then be defined as the point where the Sun seems to cross the equator from south to north.

In practice the situation is more complicated, since the direction of Earth's rotation axis changes due to perturbations. Therefore the equatorial coordinate frame is nowadays defined using certain standard objects the positions of which are known very accurately. The best accuracy is achieved by using the most distant objects, *quasars* (Sect. 18.7), which remain in the same direction over very long intervals of time.

2.6 Rising and Setting Times

From the last equation (2.16), we find the hour angle h of an object at the moment its altitude is a:

$$\cos h = -\tan\delta\tan\phi + \frac{\sin a}{\cos\delta\cos\phi} \,. \tag{2.20}$$

This equation can be used for computing rising and setting times. Then $a = 0$ and the hour angles corresponding to rising and setting times are obtained from

$$\cos h = -\tan\delta\tan\phi \,. \tag{2.21}$$

If the right ascension α is known, we can use (2.11) to compute the sidereal time Θ. (Later, in Sect. 2.14, we shall study how to transform the sidereal time to ordinary time.)

If higher accuracy is needed, we have to correct for the refraction of light caused by the atmosphere of the Earth (see Sect. 2.9). In that case, we must use a small negative value for a in (2.20). This value, the *horizontal refraction*, is about $-34'$.

The rising and setting times of the Sun given in almanacs refer to the time when the upper edge of the Solar disk just touches the horizon. To compute these times, we must set $a = -50'$ $(= -34' - 16')$.

Also for the Moon almanacs give rising and setting times of the upper edge of the disk. Since the distance of the Moon varies considerably, we cannot use any constant value for the radius of the Moon, but it has to be calculated separately each time. The Moon is also so close that its direction with respect to the background stars varies due to the rotation of the Earth. Thus the rising and setting times of the Moon are defined as the instants when the altitude of the Moon is $-34' - s + \pi$, where s is the apparent radius ($15.5'$ on the average) and π the parallax ($57'$ on the average). The latter quantity is explained in Sect. 2.9.

Finding the rising and setting times of the Sun, planets and especially the Moon is complicated by their motion with respect to the stars. We can use, for example, the coordinates for the noon to calculate estimates for the rising and setting times, which can then be used to interpolate more accurate coordinates for the rising and setting times. When these coordinates are used to compute new times a pretty good accuracy can be obtained. The iteration can be repeated if even higher precision is required.

2.7 The Ecliptic System

The orbital plane of the Earth, the *ecliptic*, is the reference plane of another important coordinate frame. The ecliptic can also be defined as the great circle on the celestial sphere described by the Sun in the course of one year. This frame is used mainly for planets and other bodies of the solar system. The orientation of the Earth's equatorial plane remains invariant, unaffected by annual motion. In spring, the Sun appears to move from the southern hemisphere to the northern one (Fig. 2.14). The time of this remarkable event as well as the direction to the Sun at that moment are called the *vernal equinox*. At the vernal equinox, the Sun's right ascension and declination are zero. The equatorial and ecliptic

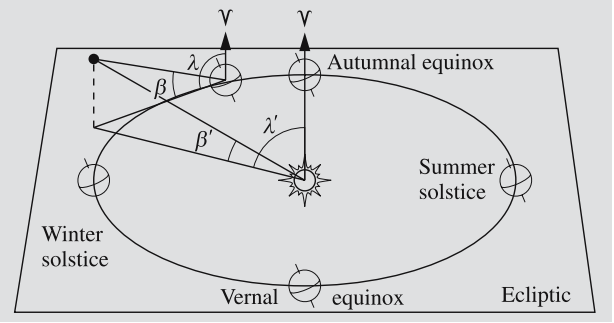


Fig. 2.14. The ecliptic geocentric (λ, β) and heliocentric (λ', β') coordinates are equal only if the object is very far away. The geocentric coordinates depend also on the Earth's position in its orbit

planes intersect along a straight line directed towards the vernal equinox. Thus we can use this direction as the zero point for both the equatorial and ecliptic coordinate frames. The point opposite the vernal equinox is the *autumnal equinox*, it is the point at which the Sun crosses the equator from north to south.

The *ecliptic latitude* β is the angular distance from the ecliptic; it is in the range $[-90°, +90°]$. The other coordinate is the *ecliptic longitude* λ, measured counterclockwise from the vernal equinox.

Transformation equations between the equatorial and ecliptic frames can be derived analogously to (2.14) and (2.16):

$$\begin{aligned} \sin\lambda\cos\beta &= \sin\delta\sin\varepsilon + \cos\delta\cos\varepsilon\sin\alpha\ , \\ \cos\lambda\cos\beta &= \cos\delta\cos\alpha\ , \\ \sin\beta &= \sin\delta\cos\varepsilon - \cos\delta\sin\varepsilon\sin\alpha\ , \end{aligned} \tag{2.22}$$

$$\begin{aligned} \sin\alpha\cos\delta &= -\sin\beta\sin\varepsilon + \cos\beta\cos\varepsilon\sin\lambda\ , \\ \cos\alpha\cos\delta &= \cos\lambda\cos\beta\ , \\ \sin\delta &= \sin\beta\cos\varepsilon + \cos\beta\sin\varepsilon\sin\lambda\ . \end{aligned} \tag{2.23}$$

The angle ε appearing in these equations is the *obliquity of the ecliptic*, or the angle between the equatorial and ecliptic planes. Its value is roughly 23°26′ (a more accurate value is given in *Reduction of Coordinates, p. 38).

Depending on the problem to be solved, we may encounter *heliocentric* (origin at the Sun), *geocentric* (origin at the centre of the Earth) or *topocentric* (origin at the observer) coordinates. For very distant objects the differences are negligible, but not for bodies of the solar system. To transform heliocentric coordinates to geocentric coordinates or vice versa, we must also know the distance of the object. This transformation is most easily accomplished by computing the rectangular coordinates of the object and the new origin, then changing the origin and finally evaluating the new latitude and longitude from the rectangular coordinates (see Examples 2.4 and 2.5).

2.8 The Galactic Coordinates

For studies of the Milky Way Galaxy, the most natural reference plane is the plane of the Milky Way (Fig. 2.15). Since the Sun lies very close to that plane, we can put the origin at the Sun. The *galactic longitude l* is measured counterclockwise (like right ascension) from the direction of the centre of the Milky Way (in Sagittarius, $\alpha = 17\,\mathrm{h}\ 45.7\,\mathrm{min}$, $\delta = -29°00'$). The *galactic latitude b* is measured from the galactic plane, positive northwards and negative southwards. This definition was officially adopted only in 1959, when the direction of the galactic centre was determined from radio observations accurately enough. The old galactic coordinates l^{I} and b^{I} had the intersection of the equator and the galactic plane as their zero point.

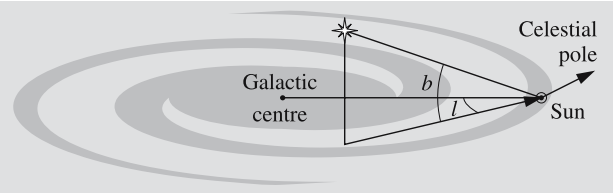


Fig. 2.15. The galactic coordinates l and b

The galactic coordinates can be obtained from the equatorial ones with the transformation equations

$$\begin{aligned} \sin(l_{\mathrm{N}} - l)\cos b &= \cos\delta\sin(\alpha - \alpha_{\mathrm{P}})\ , \\ \cos(l_{\mathrm{N}} - l)\cos b &= -\cos\delta\sin\delta_{\mathrm{P}}\cos(\alpha - \alpha_{\mathrm{P}}) \\ &\quad + \sin\delta\cos\delta_{\mathrm{P}}\ , \\ \sin b &= \cos\delta\cos\delta_{\mathrm{P}}\cos(\alpha - \alpha_{\mathrm{P}}) \\ &\quad + \sin\delta\sin\delta_{\mathrm{P}}\ , \end{aligned} \tag{2.24}$$

where the direction of the Galactic north pole is $\alpha_{\mathrm{P}} = 12\,\mathrm{h}\ 51.4\,\mathrm{min}$, $\delta_{\mathrm{P}} = 27°08'$, and the galactic longitude of the celestial pole, $l_{\mathrm{N}} = 123.0°$.

2.9 Perturbations of Coordinates

Even if a star remains fixed with respect to the Sun, its coordinates can change, due to several disturbing effects. Naturally its altitude and azimuth change constantly because of the rotation of the Earth, but even its right ascension and declination are not quite free from perturbations.

Precession. Since most of the members of the solar system orbit close to the ecliptic, they tend to pull the equatorial bulge of the Earth towards it. Most of this "flattening" torque is caused by the Moon and the Sun.

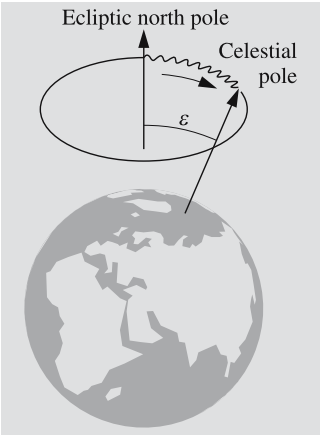


Fig. 2.16. Due to precession the rotation axis of the Earth turns around the ecliptic north pole. Nutation is the small wobble disturbing the smooth precessional motion. In this figure the magnitude of the nutation is highly exaggerated

But the Earth is rotating and therefore the torque cannot change the inclination of the equator relative to the ecliptic. Instead, the rotation axis turns in a direction perpendicular to the axis and the torque, thus describing a cone once in roughly 26,000 years. This slow turning of the rotation axis is called *precession* (Fig. 2.16). Because of precession, the vernal equinox moves along the ecliptic clockwise about 50 seconds of arc every year, thus increasing the ecliptic longitudes of all objects at the same rate. At present the rotation axis points about one degree away from Polaris, but after 12,000 years, the celestial pole will be roughly in the direction of Vega. The changing ecliptic longitudes also affect the right ascension and declination. Thus we have to know the instant of time, or *epoch*, for which the coordinates are given.

Currently most maps and catalogues use the epoch J2000.0, which means the beginning of the year 2000, or, to be exact, the noon of January 1, 2000, or the Julian date 2,451,545.0 (see Sect. 2.15).

Let us now derive expressions for the changes in right ascension and declination. Taking the last transformation equation in (2.23),

$$\sin\delta = \cos\varepsilon\sin\beta + \sin\varepsilon\cos\beta\sin\lambda\ ,$$

and differentiating, we get

$$\cos\delta\,\mathrm{d}\delta = \sin\varepsilon\cos\beta\cos\lambda\,\mathrm{d}\lambda\ .$$

Applying the second equation in (2.22) to the right-hand side, we have, for the change in declination,

$$\mathrm{d}\delta = \mathrm{d}\lambda\sin\varepsilon\cos\alpha\ . \tag{2.25}$$

By differentiating the equation

$$\cos\alpha\cos\delta = \cos\beta\cos\lambda\ ,$$

we get

$$-\sin\alpha\cos\delta\,\mathrm{d}\alpha - \cos\alpha\sin\delta\,\mathrm{d}\delta = -\cos\beta\sin\lambda\,\mathrm{d}\lambda\ ;$$

and, by substituting the previously obtained expression for $\mathrm{d}\delta$ and applying the first equation (2.22), we have

$$\begin{aligned}\sin\alpha\cos\delta\,\mathrm{d}\alpha &= \mathrm{d}\lambda(\cos\beta\sin\lambda - \sin\varepsilon\cos^2\alpha\sin\delta)\\ &= \mathrm{d}\lambda(\sin\delta\sin\varepsilon + \cos\delta\cos\varepsilon\sin\alpha\\ &\quad - \sin\varepsilon\cos^2\alpha\sin\delta)\ .\end{aligned}$$

Simplifying this, we get

$$\mathrm{d}\alpha = \mathrm{d}\lambda(\sin\alpha\sin\varepsilon\tan\delta + \cos\varepsilon)\ . \tag{2.26}$$

If $\mathrm{d}\lambda$ is the annual increment of the ecliptic longitude (about $50''$), the precessional changes in right ascension and declination in one year are thus

$$\begin{aligned}\mathrm{d}\delta &= \mathrm{d}\lambda\sin\varepsilon\cos\alpha\ ,\\ \mathrm{d}\alpha &= \mathrm{d}\lambda(\sin\varepsilon\sin\alpha\tan\delta + \cos\varepsilon)\ .\end{aligned} \tag{2.27}$$

These expressions are usually written in the form

$$\begin{aligned}\mathrm{d}\delta &= n\cos\alpha\ ,\\ \mathrm{d}\alpha &= m + n\sin\alpha\tan\delta\ ,\end{aligned} \tag{2.28}$$

where

$$\begin{aligned}m &= \mathrm{d}\lambda\cos\varepsilon\ ,\\ n &= \mathrm{d}\lambda\sin\varepsilon\end{aligned} \tag{2.29}$$

are the *precession constants*. Since the obliquity of the ecliptic is not exactly a constant but changes with time, m and n also vary slowly with time. However, this variation is so slow that usually we can regard m and n as constants unless the time interval is very long. The values of these constants for some epochs are given in

Table 2.1. Precession constants m and n. Here, "a" means a tropical year

Epoch	m	n	
1800	3.07048 s/a	1.33703 s/a =	$20.0554''$/a
1850	3.07141	1.33674	20.0511
1900	3.07234	1.33646	20.0468
1950	3.07327	1.33617	20.0426
2000	3.07419	1.33589	20.0383

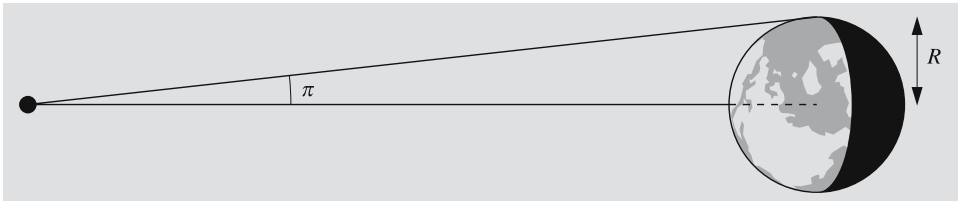


Fig. 2.17. The horizontal parallax π of an object is the angle subtended by the Earth's equatorial radius as seen from the object

Table 2.1. For intervals longer than a few decades a more rigorous method should be used. Its derivation exceeds the level of this book, but the necessary formulas are given in *Reduction of Coordinates (p. 38).

Nutation. The Moon's orbit is inclined with respect to the ecliptic, resulting in precession of its orbital plane. One revolution takes 18.6 years, producing perturbations with the same period in the precession of the Earth. This effect, *nutation*, changes ecliptic longitudes as well as the obliquity of the ecliptic (Fig. 2.16). Calculations are now much more complicated, but fortunately nutational perturbations are relatively small, only fractions of an arc minute.

Parallax. If we observe an object from different points, we see it in different directions. The difference of the observed directions is called the *parallax*. Since the amount of parallax depends on the distance of the observer from the object, we can utilize the parallax to measure distances. Human stereoscopic vision is based (at least to some extent) on this effect. For astronomical purposes we need much longer baselines than the distance between our eyes (about 7 cm). Appropriately large and convenient baselines are the radius of the Earth and the radius of its orbit.

Distances to the nearest stars can be determined from the *annual parallax*, which is the angle subtended by the radius of the Earth's orbit (called the *astronomical unit*, AU) as seen from the star. (We shall discuss this further in Sect. 2.10.)

By *diurnal parallax* we mean the change of direction due to the daily rotation of the Earth. In addition to the distance of the object, the diurnal parallax also depends on the latitude of the observer. If we talk about the parallax of a body in our solar system, we always mean the angle subtended by the Earth's equatorial radius (6378 km) as seen from the object (Fig. 2.17). This equals the apparent shift of the object with respect to the background stars seen by an observer at the equator if (s)he observes the object moving from the horizon to the zenith. The parallax of the Moon, for example, is about $57'$, and that of the Sun $8.79''$.

In astronomy parallax may also refer to distance in general, even if it is not measured using the shift in the observed direction.

Aberration. Because of the finite speed of light, an observer in motion sees an object shifted in the direction of her/his motion (Figs. 2.18 and 2.19). This change of apparent direction is called *aberration*. To derive

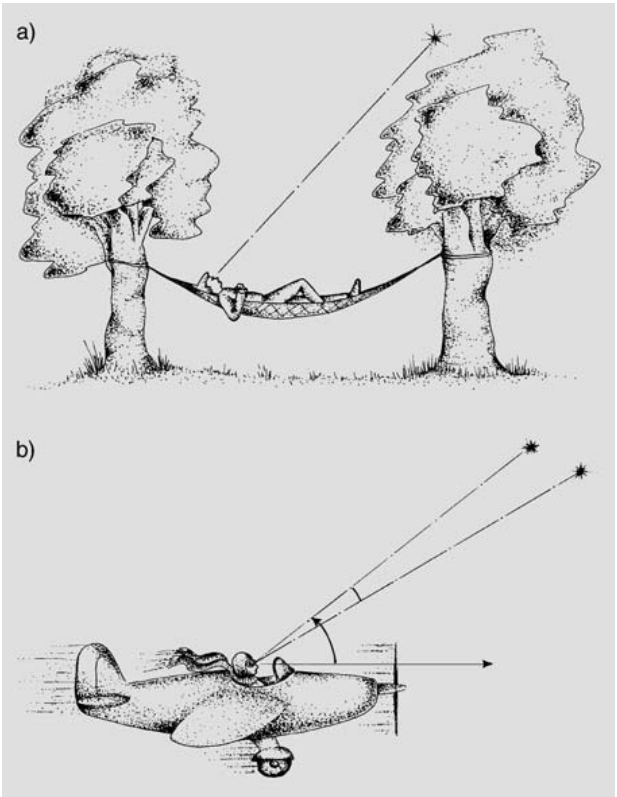


Fig. 2.18a,b. The effect of aberration on the apparent direction of an object. (**a**) Observer at rest. (**b**) Observer in motion

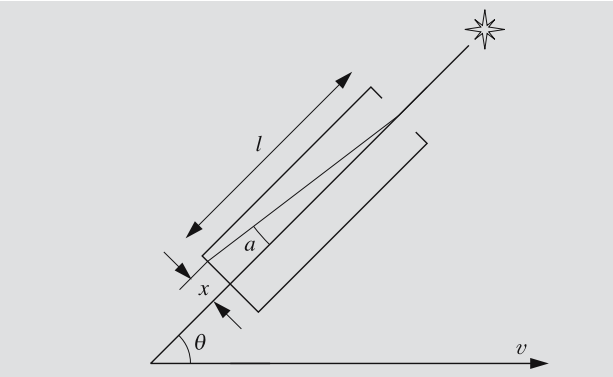


Fig. 2.19. A telescope is pointed in the true direction of a star. It takes a time $t = l/c$ for the light to travel the length of the telescope. The telescope is moving with velocity v, which has a component $v \sin\theta$, perpendicular to the direction of the light beam. The beam will hit the bottom of the telescope displaced from the optical axis by a distance $x = tv \sin\theta = l(v/c) \sin\theta$. Thus the change of direction in radians is $a = x/l = (v/c) \sin\theta$

the exact value we have to use the special theory of relativity, but for practical purposes it suffices to use the approximate value

$$a = \frac{v}{c} \sin\theta \,, \quad [a] = \text{rad} \,, \tag{2.30}$$

where v is the velocity of the observer, c is the speed of light and θ is the angle between the true direction of the object and the velocity vector of the observer. The greatest possible value of the aberration due to the orbital motion of the Earth, v/c, called the *aberration constant*, is $21''$. The maximal shift due to the Earth's rotation, the diurnal aberration constant, is much smaller, about $0.3''$.

Refraction. Since light is refracted by the atmosphere, the direction of an object differs from the true direction by an amount depending on the atmospheric conditions along the line of sight. Since this refraction varies with atmospheric pressure and temperature, it is very difficult to predict it accurately. However, an approximation good enough for most practical purposes is easily derived. If the object is not too far from the zenith, the atmosphere between the object and the observer can be approximated by a stack of parallel planar layers, each of which has a certain index of refraction n_i (Fig. 2.20). Outside the atmosphere, we have $n = 1$.

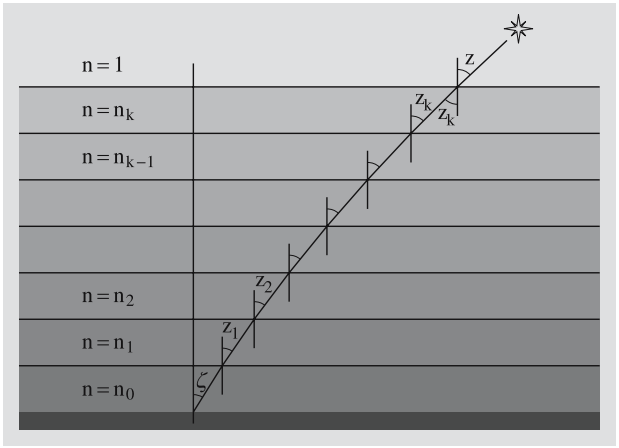


Fig. 2.20. Refraction of a light ray travelling through the atmosphere

Let the true zenith distance be z and the apparent one, ζ. Using the notations of Fig. 2.20, we obtain the following equations for the boundaries of the successive layers:

$$\begin{aligned} \sin z &= n_k \sin z_k \,, \\ &\vdots \\ n_2 \sin z_2 &= n_1 \sin z_1 \,, \\ n_1 \sin z_1 &= n_0 \sin\zeta \,, \end{aligned}$$

or

$$\sin z = n_0 \sin\zeta \,. \tag{2.31}$$

When the *refraction angle* $R = z - \zeta$ is small and is expressed in radians, we have

$$\begin{aligned} n_0 \sin\zeta &= \sin z = \sin(R+\zeta) \\ &= \sin R \cos\zeta + \cos R \sin\zeta \\ &\approx R \cos\zeta + \sin\zeta \,. \end{aligned}$$

Thus we get

$$R = (n_0 - 1) \tan\zeta \,, \quad [R] = \text{rad} \,. \tag{2.32}$$

The index of refraction depends on the density of the air, which further depends on the pressure and temperature. When the altitude is over 15°, we can use an approximate formula

$$R = \frac{P}{273+T}\, 0.00452^\circ \tan(90^\circ - a) \,, \tag{2.33}$$

where a is the altitude in degrees, T temperature in degrees Celsius, and P the atmospheric pressure in hectopascals (or, equivalently, in millibars). At lower altitudes the curvature of the atmosphere must be taken into account. An approximate formula for the refraction is then

$$R = \frac{P}{273+T}\,\frac{0.1594+0.0196a+0.00002a^2}{1+0.505a+0.0845a^2}\,. \tag{2.34}$$

These formulas are widely used, although they are against the rules of dimensional analysis. To get correct values, all quantities must be expressed in correct units. Figure 2.21 shows the refraction under different conditions evaluated from these formulas.

Altitude is always (except very close to zenith) increased by refraction. On the horizon the change is about $34'$, which is slightly more than the diameter of the Sun. When the lower limb of the Sun just touches the horizon, the Sun has in reality already set.

Light coming from the zenith is not refracted at all if the boundaries between the layers are horizontal. Under some climatic conditions, a boundary (e. g. between cold and warm layers) can be slanted, and in this case, there can be a small zenith refraction, which is of the order of a few arc seconds.

Stellar positions given in star catalogues are *mean places*, from which the effects of parallax, aberration and nutation have been removed. The mean place of the date (i. e. at the observing time) is obtained by correcting the mean place for the proper motion of the star (Sect. 2.10) and precession. The *apparent place* is obtained by correcting this place further for nutation, parallax and aberration. There is a catalogue published annually that gives the apparent places of certain references stars at intervals of a few days. These positions have been corrected for precession, nutation, parallax and annual aberration. The effects of diurnal aberration and refraction are not included because they depend on the location of the observer.

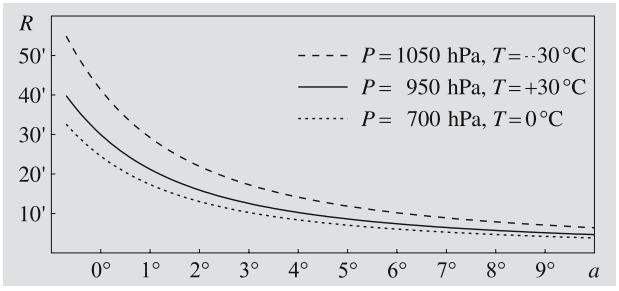
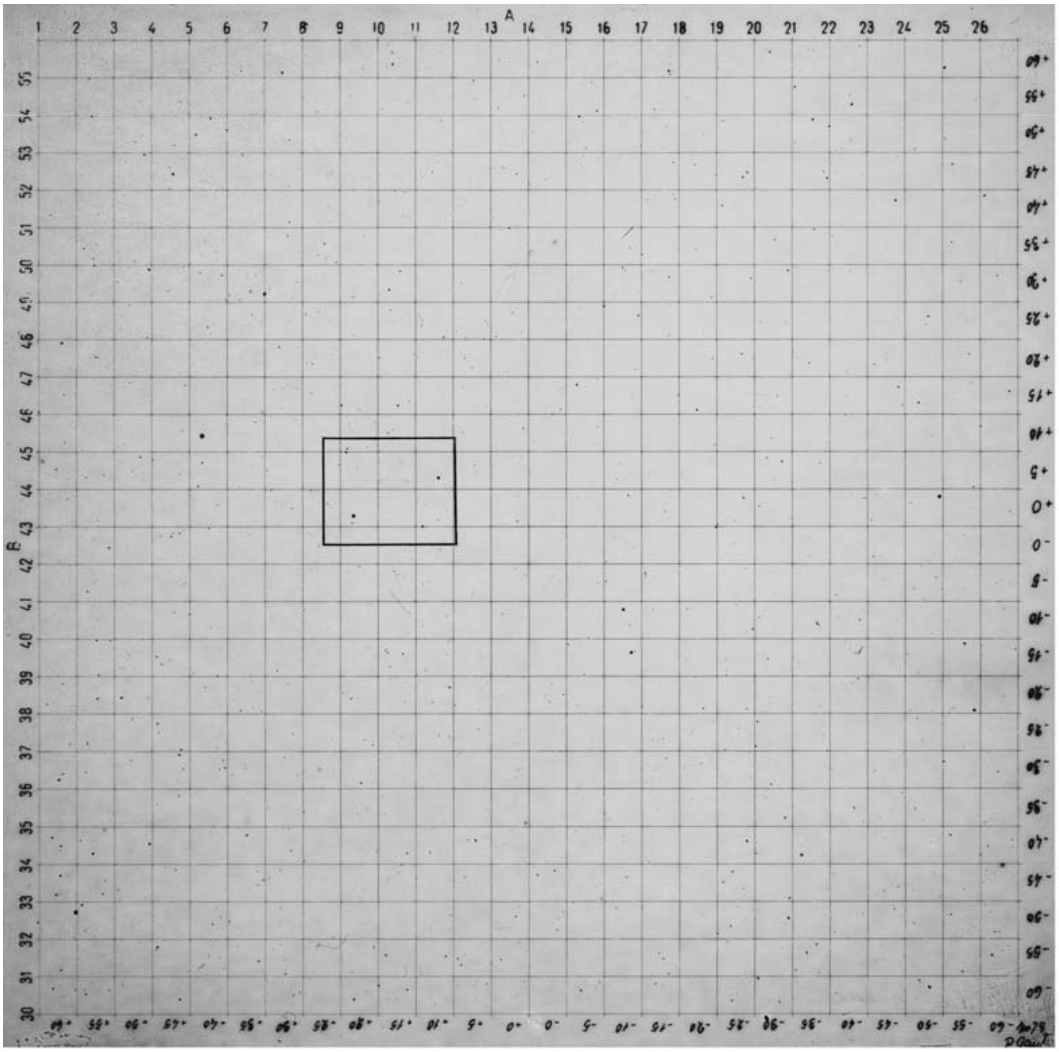


Fig. 2.21. Refraction at different altitudes. The refraction angle R tells how much higher the object seems to be compared with its true altitude a. Refraction depends on the density and thus on the pressure and temperature of the air. The *upper curves* give the refraction at sea level during rather extreme weather conditions. At the altitude of 2.5 kilometers the average pressure is only 700 hPa, and thus the effect of refraction smaller (*lowest curve*)

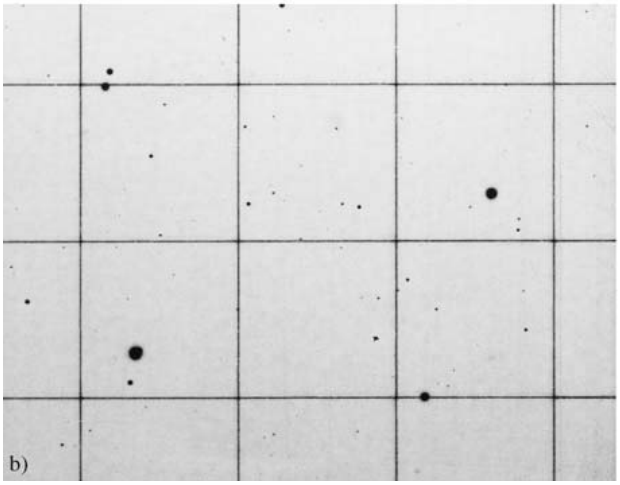
2.10 Positional Astronomy

The position of a star can be measured either with respect to some reference stars (relative astrometry) or with respect to a fixed coordinate frame (absolute astrometry).

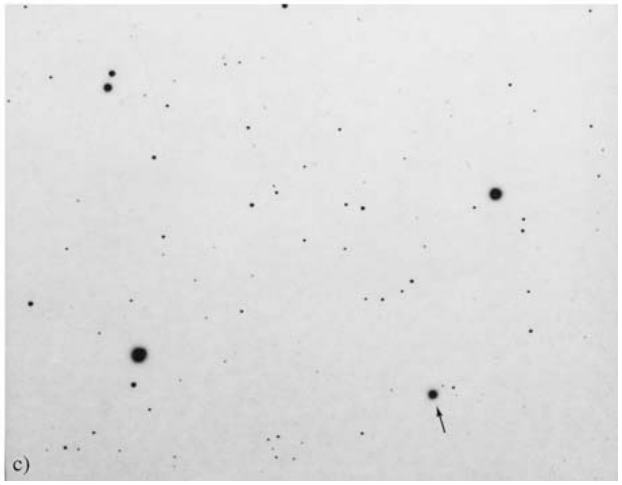
Fig. 2.22. Astronomers discussing observations with the transit circle of Helsinki Observatory in 1904



a)



b)



c)

Absolute coordinates are usually determined using a *meridian circle*, which is a telescope that can be turned only in the meridional plane (Fig. 2.22). It has only one axis, which is aligned exactly in the east-west direction. Since all stars cross the meridian in the course of a day, they all come to the field of the meridian circle at some time or other. When a star culminates, its altitude and the time of the transit are recorded. If the time is determined with a sidereal clock, the sidereal time immediately gives the right ascension of the star, since the hour angle is $h = 0\,\mathrm{h}$. The other coordinate, the declination δ, is obtained from the altitude:

$$\delta = a - (90^\circ - \phi)\ ,$$

where a is the observed altitude and ϕ is the geographic latitude of the observatory.

Relative coordinates are measured on photographic plates (Fig. 2.23) or CCD images containing some known reference stars. The scale of the plate as well as the orientation of the coordinate frame can be determined from the reference stars. After this has been done, the right ascension and declination of any object in the image can be calculated if its coordinates in the image are measured.

All stars in a small field are almost equally affected by the dominant perturbations, precession, nutation, and aberration. The much smaller effect of parallax, on the other hand, changes the relative positions of the stars.

The shift in the direction of a star with respect to distant background stars due to the annual motion of the Earth is called the *trigonometric parallax* of the star. It gives the distance of the star: the smaller the parallax, the farther away the star is. Trigonometric parallax is, in fact, the only direct method we currently have of measuring distances to stars. Later we shall be introduced to some other, indirect methods, which require certain assumptions on the motions or structure of stars. The same method of triangulation is employed to measure distances of earthly objects. To measure distances to stars, we have to use the longest baseline available, the diameter of the orbit of the Earth.

◀ **Fig. 2.23.** (**a**) A plate photographed for the Carte du Ciel project in Helsinki on November 21, 1902. The centre of the field is at $\alpha = 18\,\mathrm{h}\ 40\,\mathrm{min}$, $\delta = 46^\circ$, and the area is $2^\circ \times 2^\circ$. Distance between coordinate lines (exposed separately on the plate) is 5 minutes of arc. (**b**) The framed region on the same plate. (**c**) The same area on a plate taken on November 7, 1948. The bright star in the lower right corner (SAO 47747) has moved about 12 seconds of arc. The brighter, slightly drop-shaped star to the left is a binary star (SAO 47767); the separation between its components is $8''$

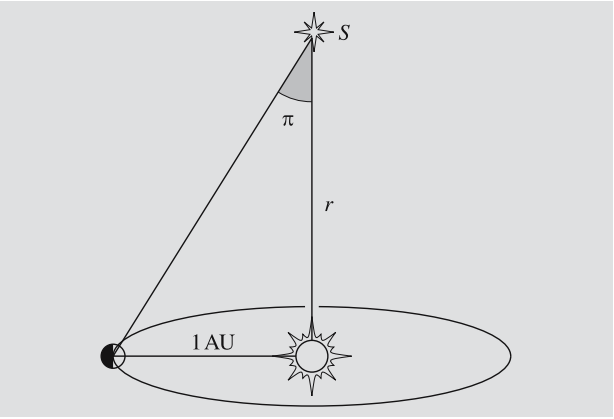


Fig. 2.24. The trigonometric parallax π of a star S is the angle subtended by the radius of the orbit of the Earth, or one astronomical unit, as seen from the star

During the course of one year, a star will appear to describe a circle if it is at the pole of the ecliptic, a segment of line if it is in the ecliptic, or an ellipse otherwise. The semimajor axis of this ellipse is called the parallax of the star. It is usually denoted by π. It equals the angle subtended by the radius of the Earth's orbit as seen from the star (Fig. 2.24).

The unit of distance used in astronomy is *parsec* (pc). At a distance of one parsec, one astronomical unit subtends an angle of one arc second. Since one radian is about $206{,}265''$, 1 pc equals 206,265 AU. Furthermore, because $1\,\mathrm{AU} = 1.496 \times 10^{11}\,\mathrm{m}$, $1\,\mathrm{pc} \approx 3.086 \times 10^{16}\,\mathrm{m}$. If the parallax is given in arc seconds, the distance is simply

$$r = 1/\pi\ , \quad [r] = \mathrm{pc}\ , \quad [\pi] = {}''\ . \tag{2.35}$$

In popular astronomical texts, distances are usually given in *light-years*, one light-year being the distance light travels in one year, or $9.5 \times 10^{15}\,\mathrm{m}$. Thus one parsec is about 3.26 light-years.

The first parallax measurement was accomplished by *Friedrich Wilhelm Bessel* (1784–1846) in 1838. He found the parallax of 61 Cygni to be $0.3''$. The nearest star Proxima Centauri has a parallax of $0.762''$ and thus a distance of 1.31 pc.

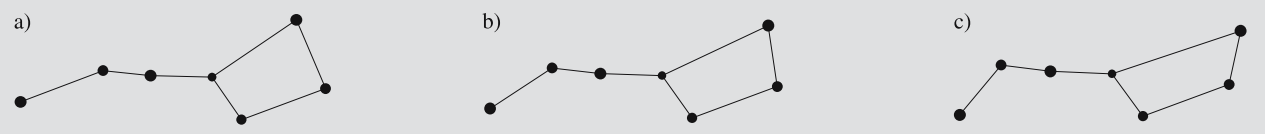


Fig. 2.25a–c. Proper motions of stars slowly change the appearance of constellations. (**a**) The Big Dipper during the last ice age 30,000 years ago, (**b**) nowadays, and (**c**) after 30,000 years

In addition to the motion due to the annual parallax, many stars seem to move slowly in a direction that does not change with time. This effect is caused by the relative motion of the Sun and the stars through space; it is called the *proper motion*. The appearance of the sky and the shapes of the constellations are constantly, although extremely slowly, changed by the proper motions of the stars (Fig. 2.25).

The velocity of a star with respect to the Sun can be divided into two components (Fig. 2.26), one of which is directed along the line of sight (the radial component or the *radial velocity*), and the other perpendicular to it (the tangential component). The tangential velocity results in the proper motion, which can be measured by taking plates at intervals of several years or decades. The proper motion μ has two components, one giving the change in declination μ_δ and the other, in right ascension, $\mu_\alpha \cos\delta$. The coefficient $\cos\delta$ is used to correct the scale of right ascension: hour circles (the great circles with $\alpha =$ constant) approach each other towards the poles, so the coordinate difference must be multiplied by $\cos\delta$ to obtain the true angular separation. The total proper motion is

$$\mu = \sqrt{\mu_\alpha^2 \cos^2\delta + \mu_\delta^2}\,. \tag{2.36}$$

The greatest known proper motion belongs to Barnard's Star, which moves across the sky at the enormous speed of 10.3 arc seconds per year. It needs less than 200 years to travel the diameter of a full moon.

In order to measure proper motions, we must observe stars for decades. The radial component, on the other hand, is readily obtained from a single observation, thanks to the *Doppler effect*. By the Doppler effect we mean the change in frequency and wavelength of radiation due to the radial velocity of the radiation source. The same effect can be observed, for example, in the sound of an ambulance, the pitch being higher when the ambulance is approaching and lower when it is receding.

The formula for the Doppler effect for small velocities can be derived as in Fig. 2.27. The source of radiation transmits electromagnetic waves, the period of one cycle being T. In time T, the radiation approaches the observer by a distance $s = cT$, where c is the speed of propagation. During the same time, the source moves with respect to the observer a distance $s' = vT$, where v is the speed of the source, positive for a receding source and negative for an approaching one. We find that the length of one cycle, the wavelength λ, equals

$$\lambda = s + s' = cT + vT\,.$$

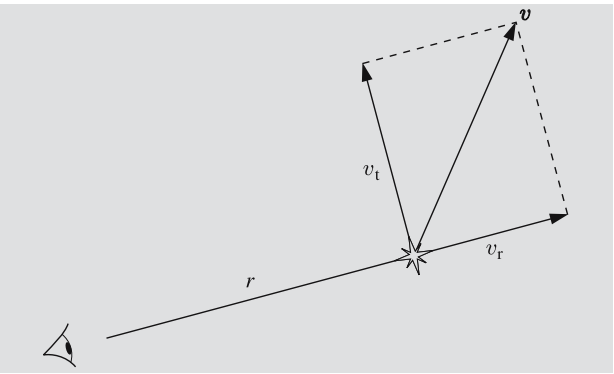


Fig. 2.26. The radial and tangential components, v_r and v_t of the velocity $\boldsymbol{v}$ of a star. The latter component is observed as proper motion

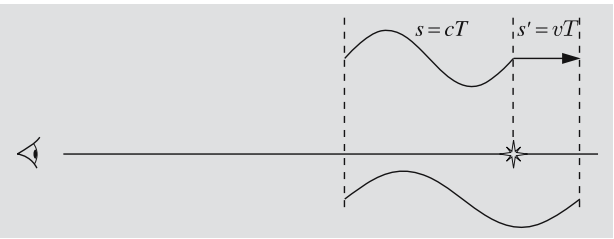


Fig. 2.27. The wavelength of radiation increases if the source is receding

If the source were at rest, the wavelength of its radiation would be $\lambda_0 = cT$. The motion of the source changes the wavelength by an amount

$$\Delta\lambda = \lambda - \lambda_0 = cT + vT - cT = vT \,,$$

and the relative change $\Delta\lambda$ of the wavelength is

$$\frac{\Delta\lambda}{\lambda_0} = \frac{v}{c} \,. \tag{2.37}$$

This is valid only when $v \ll c$. For very high velocities, we must use the relativistic formula

$$\frac{\Delta\lambda}{\lambda_0} = \sqrt{\frac{1+v/c}{1-v/c}} - 1 \,. \tag{2.38}$$

In astronomy the Doppler effect can be seen in stellar spectra, in which the spectral lines are often displaced towards the blue (shorter wavelengths) or red (longer wavelengths) end of the spectrum. A *blueshift* means that the star is approaching, while a *redshift* indicates that it is receding.

The displacements due to the Doppler effect are usually very small. In order to measure them, a reference spectrum is exposed on the plate next to the stellar spectrum. Now that CCD-cameras have replaced photographic plates, separate calibration exposures of reference spectra are taken to determine the wavelength scale. The lines in the reference spectrum are produced by a light source at rest in the laboratory. If the reference spectrum contains some lines found also in the stellar spectrum, the displacements can be measured.

Displacements of spectral lines give the radial velocity v_r of the star, and the proper motion μ can be measured from photographic plates or CCD images. To find the tangential velocity v_t, we have to know the distance r, obtainable from e.g. parallax measurements. Tangential velocity and proper motion are related by

$$v_\mathrm{t} = \mu r \,. \tag{2.39}$$

If μ is given in arc seconds per year and r in parsecs we have to make the following unit transformations to get v_t in km/s:

$$\begin{aligned} &1 \text{ rad} = 206{,}265'' \,, \quad 1 \text{ year} = 3.156 \times 10^7 \text{ s} \,, \\ &1 \text{ pc} = 3.086 \times 10^{13} \text{ km} \,. \end{aligned}$$

Hence

$$\begin{aligned} &v_\mathrm{t} = 4.74\,\mu r \,, \quad [v_\mathrm{t}] = \text{km/s} \,, \\ &[\mu] = ''/\text{a} \,, \quad [r] = \text{pc} \,. \end{aligned} \tag{2.40}$$

The total velocity v of the star is then

$$v = \sqrt{v_\mathrm{r}^2 + v_\mathrm{t}^2} \,. \tag{2.41}$$

2.11 Constellations

At any one time, about 1000–1500 stars can be seen in the sky (above the horizon). Under ideal conditions, the number of stars visible to the naked eye can be as high as 3000 on a hemisphere, or 6000 altogether. Some stars seem to form figures vaguely resembling something; they have been ascribed to various mythological and other animals. This grouping of stars into constellations is a product of human imagination without any physical basis. Different cultures have different constellations, depending on their mythology, history and environment.

About half of the shapes and names of the constellations we are familiar with date back to Mediterranean antiquity. But the names and boundaries were far from unambiguous as late as the 19th century. Therefore the International Astronomical Union (IAU) confirmed fixed boundaries at its 1928 meeting.

The official boundaries of the constellations were established along lines of constant right ascension and declination for the epoch 1875. During the time elapsed since then, precession has noticeably turned the equatorial frame. However, the boundaries remain fixed with respect to the stars. So a star belonging to a constellation will belong to it forever (unless it is moved across the boundary by its proper motion).

The names of the 88 constellations confirmed by the IAU are given in Table C.21 at the end of the book. The table also gives the abbreviation of the Latin name, its genitive (needed for names of stars) and the English name.

In his star atlas *Uranometria* (1603) *Johannes Bayer* started the current practice to denote the brightest stars of each constellation by Greek letters. The brightest star is usually α (alpha), e. g. Deneb in the constellation Cygnus is α Cygni, which is abbreviated as α Cyg. The second brightest star is β (beta), the next one γ (gamma) and so on. There are, however, several exceptions to this rule; for example, the stars of the Big Dipper are named in the order they appear in the constellation. After the Greek alphabet has been exhausted, Latin letters can be employed. Another method is to use

numbers, which are assigned in the order of increasing right ascension; e. g. 30 Tau is a bright binary star in the constellation Taurus. Moreover, variable stars have their special identifiers (Sect. 13.1). About two hundred bright stars have a proper name; e. g. the bright α Aur is called also Capella.

As telescopes evolved, more and more stars were seen and catalogued. It soon became impractical to continue this method of naming. Thus most of the stars are known only by their catalogue index numbers. One star may have many different numbers; e. g. the abovementioned Capella (α Aur) is number BD+45° 1077 in the Bonner Durchmusterung and HD 34029 in the Henry Draper catalogue.

2.12 Star Catalogues and Maps

The first actual star catalogue was published by *Ptolemy* in the second century; this catalogue appeared in the book to be known later as *Almagest* (which is a Latin corruption of the name of the Arabic translation, *Almijisti*). It had 1025 entries; the positions of these bright stars had been measured by *Hipparchos* 250 years earlier. Ptolemy's catalogue was the only widely used one prior to the 17th century.

The first catalogues still being used by astronomers were prepared under the direction of *Friedrich Wilhelm August Argelander* (1799–1875). Argelander worked in Turku and later served as professor of astronomy in Helsinki, but he made his major contributions in Bonn. Using a 72 mm telescope, he and his assistants measured the positions and estimated the magnitudes of 320,000 stars. The catalogue, *Bonner Durchmusterung*, contains nearly all stars brighter than magnitude 9.5 between the north pole and declination −2°. (Magnitudes are further discussed in Chap. 4.) Argelander's work was later used as a model for two other large catalogues covering the whole sky. The total number of stars in these catalogues is close to one million.

The purpose of these *Durchmusterungen* or general catalogues was to systematically list a great number of stars. In the *zone catalogues*, the main goal is to give the positions of stars as exactly as possible. A typical zone catalogue is the German *Katalog der Astronomischen Gesellschaft* (AGK). Twelve observatories, each measuring a certain region in the sky, contributed to this catalogue. The work was begun in the 1870's and completed at the turn of the century.

General and zone catalogues were based on visual observations with a telescope. The evolution of photography made this kind of work unnecessary at the end of the 19th century. Photographic plates could be stored for future purposes, and measuring the positions of stars became easier and faster, making it possible to measure many more stars.

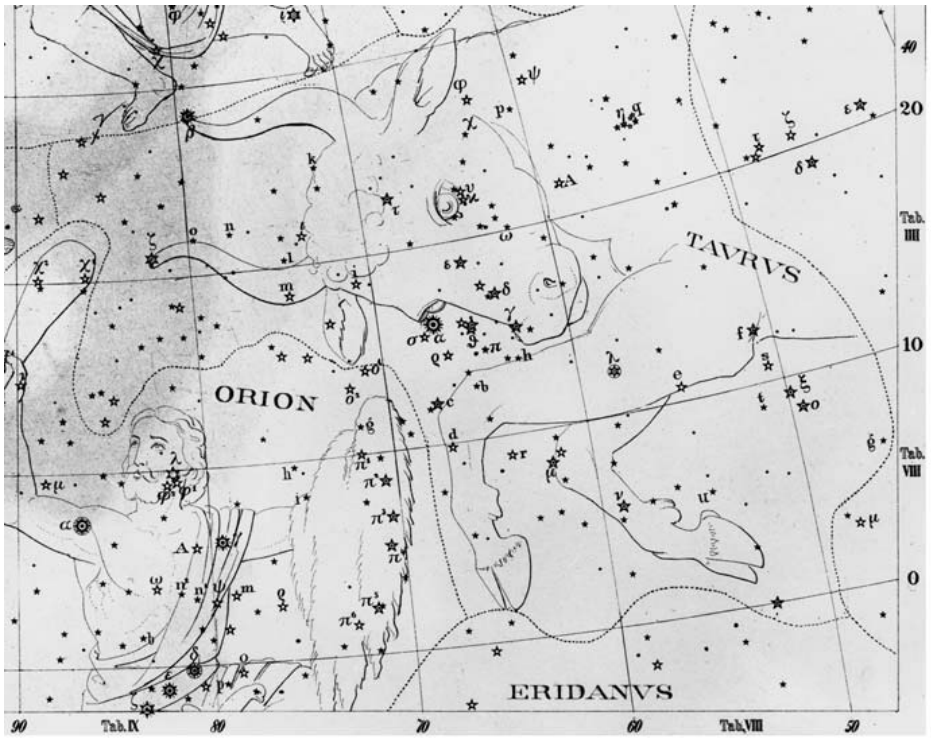
A great international program was started at the end of the 19th century in order to photograph the entire sky. Eighteen observatories participated in this *Carte du Ciel* project, all using similar instruments and plates. The positions of stars were first measured with respect to a rectangular grid exposed on each plate (Fig. 2.23a). These coordinates could then be converted into declination and right ascension.

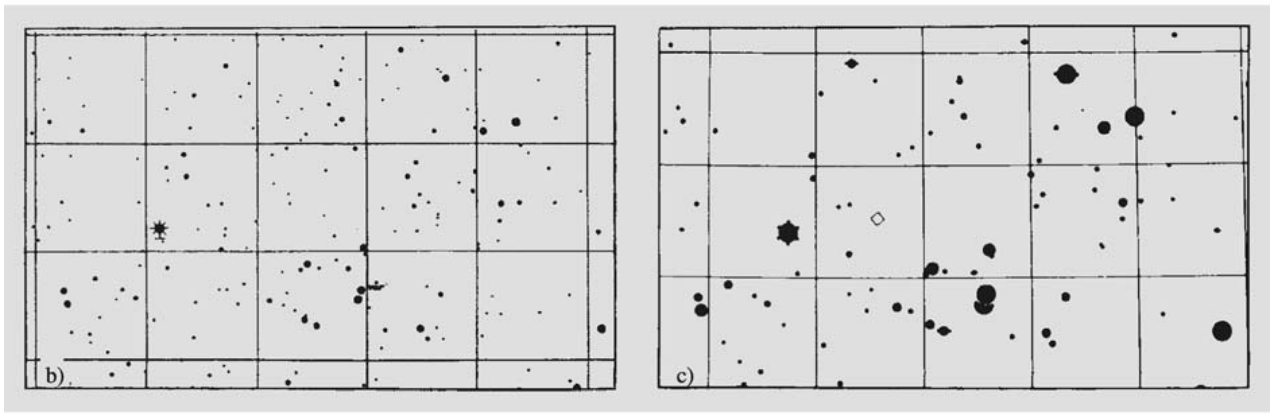
Positions of stars in catalogues are measured with respect to certain comparison stars, the coordinates of which are known with high accuracy. The coordinates of these reference stars are published in fundamental catalogues. The first such catalogue was needed for the AGK catalogue; it was published in Germany in 1879. This *Fundamental Katalog* (FK 1) gives the positions of over 500 stars.

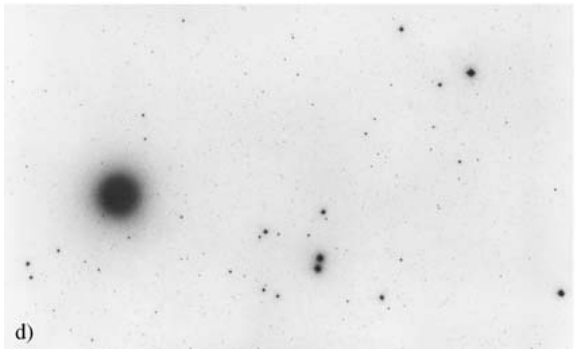
The fundamental catalogues are usually revised every few decades. The fifth edition, the FK 5, appeared in 1984. At the same time, a new system of astronomical constants was adopted. The catalogue contains 1535 fundamental and 3117 additional stars.

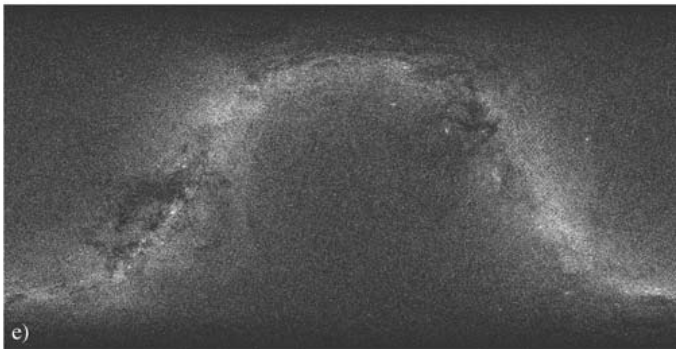
A widely used catalogue is the SAO catalogue, published by the Smithsonian Astrophysical Observatory in the 1960's. It contains the exact positions, magnitudes, proper motions, spectral classifications, etc. of 258,997 stars brighter than magnitude 9. The catalogue was accompanied by a star map containing all the stars in the catalogue.

Fig. 2.28. The representations in four atlases of the Hyades cluster in the constellation Taurus. (**a**) Heis: Atlas Coelestis, published in 1872. (**b**) Bonner Durchmusterung. (**c**) SAO, (**d**) Palomar Sky Atlas, red plate. The big blob is the brightest star of Taurus, or α Tauri alias Aldebaran. (**e**) All the stars in the Tycho Catalog, numbering over one million, are marked on an all-sky chart. The bright lane is the Milky Way. (Picture David Seal, NASA/JPL/Caltech) ►


ORION
TAVRVS
ERIDANVS
Tab. IX
Tab. VIII
Tab. III
Tab. VIII
a)


b)
c)


d)


e)

In the 1990's a large astrometric catalogue, *PPM* (Positions and Proper Motions), was published to replace the AGK and SAO catalogues. It contained all stars brighter than 7.5 magnitudes, and was almost complete to magnitude 8.5. Altogether, the four volumes of the catalogue contained information on 378,910 stars.

The PPM was effectively replaced by the Tycho catalogue from Hipparcos satellite. Hipparcos was the first astrometric satellite, and was launched by the European Space Agency (ESA) in 1989. Although Hipparcos didn't reach the planned geosynchronous orbit, it gave exact positions of over a hundred thousand stars. The *Hipparcos catalogue*, based on the measurements of the satellite, contains astrometric and photometric data of 118,000 stars. The coordinates are precise to a couple of milliarcseconds. The less precise *Tycho catalogue* contains the data of about one million stars.

In 1999 and 2000, the sixth version of the Fundamental Katalog, the FK6, was published. It combined the Hipparcos data and FK5 for 4150 fundamental stars. The typical mean error in proper motion was 0.35 milliarcseconds per year for the basic stars. With the advance of the Internet, the printed versions of star catalogues were discontinued in the first years of the new millennium, and the catalogues were moved to compact discs and the net.

With the new media, the size of the star catalogues exploded. The first Hubble Guide Star Catalog from the early 1990's contained 18 million stars and the second Guide Star Catalog from the year 2001, nearly 500 million stars. It was surpassed by the U.S. Naval Observatory USNO-B1.0 Catalog, which contains entries for 1,024,618,261 stars and galaxies from digitized images of several photographic sky surveys. The catalogue presents right ascension and declination, proper motion and magnitude estimates.

The next step in the accuracy of astrometry will be achieved in the 2010's with a new European astrometric satellite. The Gaia satellite, planned to be launched in about 2014, is expected to improve the accurary to about 10^{-5} seconds of arc.

Star maps have been published since ancient times, but the earliest maps were globes showing the celestial sphere as seen from the outside. At the beginning of the 17th century, a German, Johannes Bayer, published the first map showing the stars as seen from inside the celestial sphere, as we see them in the sky. Constellations were usually decorated with drawings of mythological figures. The *Uranometria Nova* (1843) by Argelander represents a transition towards modern maps: mythological figures are beginning to fade away. The map accompanying the *Bonner Durchmusterung* carried this evolution to its extreme. The sheets contain nothing but stars and coordinate lines.

Most maps are based on star catalogues. Photography made it possible to produce star maps without the cataloguing stage. The most important of such maps is a photographic atlas the full name of which is *The National Geographic Society – Palomar Observatory Sky Atlas*. The plates for this atlas were taken with the 1.2 m Schmidt camera on Mount Palomar. The Palomar Sky Atlas was completed in the 1950's. It consists of 935 pairs of photographs: each region has been photographed in red and blue light. The size of each plate is about 35 cm × 35 cm, covering an area of 6.6° × 6.6°. The prints are negatives (black stars on a light background), because in this way, fainter objects are visible. The limiting magnitude is about 19 in blue and 20 in red.

The Palomar atlas covers the sky down to −30°. Work to map the rest of the sky was carried out later at two observatories in the southern hemisphere, at Siding Spring Observatory in Australia, and at the European Southern Observatory (ESO) in Chile. The instruments and the scale on the plates are similar to those used earlier for the Palomar plates, but the atlas is distributed on film transparencies instead of paper prints.

For amateurs there are several star maps of various kinds. Some of them are mentioned in the references.

2.13 Sidereal and Solar Time

Time measurements can be based on the rotation of the Earth, orbital motion around the Sun, or on atomic clocks. The last-mentioned will be discussed in the next section. Here we consider the sidereal and solar times related to the rotation of the Earth.

We defined the sidereal time as the hour angle of the vernal equinox. A good basic unit is a *sidereal day*, which is the time between two successive upper culminations of the vernal equinox. After one sidereal day the celestial sphere with all its stars has returned to its original position with respect to the observer. The flow of sidereal time is as constant as the rotation of the

Earth. The rotation rate is slowly decreasing, and thus the length of the sidereal day is increasing. In addition to the smooth slowing down irregular variations of the order of one millisecond have been observed.

Unfortunately, also the sidereal time comes in two varieties, apparent and mean. The *apparent sidereal time* is determined by the true vernal equinox, and so it is obtained directly from observations.

Because of the precession the ecliptic longitude of the vernal equinox increases by about $50''$ a year. This motion is very smooth. Nutation causes more complicated wobbling. The *mean equinox* is the point where the vernal equinox would be if there were no nutation. The *mean sidereal time* is the hour angle of this mean equinox.

The difference of the apparent and mean sidereal time is called the *equation of equinoxes*:

$$\Theta_\mathrm{a} - \Theta_\mathrm{M} = \Delta\psi \cos\varepsilon \,, \tag{2.42}$$

where ε is the obliquity of the ecliptic at the instant of the observation, and $\Delta\psi$, the nutation in longitude. This value is tabulated for each day e. g. in the *Astronomical Almanac*. It can also be computed from the formulae given in *Reduction of Coordinates. It is at most about one second, so it has to be taken into account only in the most precise calculations.

Figure 2.29 shows the Sun and the Earth at vernal equinox. When the Earth is at the point A, the Sun culminates and, at the same time, a new sidereal day begins in the city with the huge black arrow standing in its central square. After one sidereal day, the Earth has moved along its orbit almost one degree of arc to the point B. Therefore the Earth has to turn almost a degree further before the Sun will culminate. The *solar* or *synodic day* is therefore 3 min 56.56 s (sidereal time) longer than the sidereal day. This means that the beginning of the sidereal day will move around the clock during the course of one year. After one year, sidereal and solar time will again be in phase. The number of sidereal days in one year is one higher than the number of solar days.

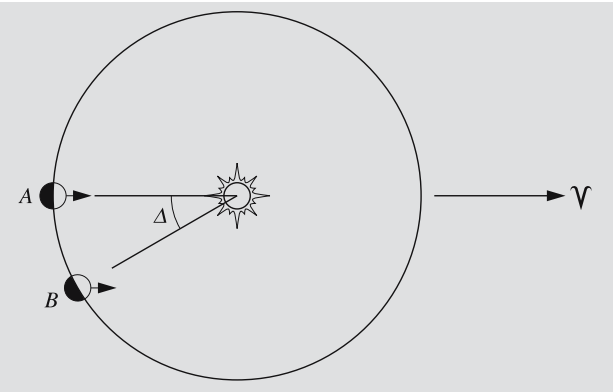


Fig. 2.29. One sidereal day is the time between two successive transits or upper culminations of the vernal equinox. By the time the Earth has moved from A to B, one sidereal day has elapsed. The angle Δ is greatly exaggerated; in reality, it is slightly less than one degree

When we talk about rotational periods of planets, we usually mean sidereal periods. The length of day, on the other hand, means the rotation period with respect to the Sun. If the orbital period around the Sun is P, sidereal rotation period τ_* and synodic day τ, we now know that the number of sidereal days in time P, P/τ_*, is one higher than the number of synodic days, P/τ:

$$\frac{P}{\tau_*} - \frac{P}{\tau} = 1 \,,$$

or

$$\frac{1}{\tau} = \frac{1}{\tau_*} - \frac{1}{P} \,. \tag{2.43}$$

This holds for a planet rotating in the direction of its orbital motion (counterclockwise). If the sense of rotation is opposite, or *retrograde*, the number of sidereal days in one orbital period is one less than the number of synodic days, and the equation becomes

$$\frac{1}{\tau} = \frac{1}{\tau_*} + \frac{1}{P} \,. \tag{2.44}$$

For the Earth, we have $P = 365.2564$ d, and $\tau = 1$ d, whence (2.43) gives $\tau_* = 0.99727\ \mathrm{d} = 23\ \mathrm{h}\ 56\ \mathrm{min}\ 4\ \mathrm{s}$, solar time.

Since our everyday life follows the alternation of day and night, it is more convenient to base our timekeeping on the apparent motion of the Sun rather than that of the stars. Unfortunately, solar time does not flow at a constant rate. There are two reasons for this. First, the orbit of the Earth is not exactly circular, but an ellipse, which means that the velocity of the Earth along its orbit is not constant. Second, the Sun moves along the ecliptic, not the equator. Thus its right ascension does not increase at a constant rate. The change is fastest at the end of December (4 min 27 s per day) and slowest in

mid-September (3 min 35 s per day). As a consequence, the hour angle of the Sun (which determines the solar time) also grows at an uneven rate.

To find a solar time flowing at a constant rate, we define a fictitious *mean sun*, which moves along the celestial equator with constant angular velocity, making a complete revolution in one year. By year we mean here the *tropical year*, which is the time it takes for the Sun to move from one vernal equinox to the next. In one tropical year, the right ascension of the Sun increases exactly 24 hours. The length of the tropical year is 365 d 5 h 48 min 46 s = 365.2422 d. Since the direction of the vernal equinox moves due to precession, the tropical year differs from the sidereal year, during which the Sun makes one revolution with respect to the background stars. One sidereal year is 365.2564 d.

Using our artificial mean sun, we now define an evenly flowing solar time, the *mean solar time* (or simply *mean time*) T_{M}, which is equal to the hour angle h_{M} of the centre of the mean sun plus 12 hours (so that the date will change at midnight, to annoy astronomers):

$$T_{\mathrm{M}} = h_{\mathrm{M}} + 12\,\mathrm{h}\,. \tag{2.45}$$

The difference between the true solar time T and the mean time T_{M} is called the *equation of time*:

$$\mathrm{E.T.} = T - T_{\mathrm{M}}\,. \tag{2.46}$$

(In spite of the identical abbreviation, this has nothing to do with a certain species of little green men.) The greatest positive value of E.T. is about 16 minutes and the greatest negative value about −14 minutes (see Fig. 2.30). This is also the difference between the true noon (the meridian transit of the Sun) and the mean noon.

Both the true solar time and mean time are *local times*, depending on the hour angle of the Sun, real or artificial. If one observes the true solar time by direct measurement and computes the mean time from (2.46), a digital watch will probably be found to disagree with both of them. The reason for this is that we do not use local time in our everyday life; instead we use the *zonal time* of the nearest *time zone*.

In the past, each city had its own local time. When travelling became faster and more popular, the great variety of local times became an inconvenience. At the end of the 19th century, the Earth was divided into 24 zones, the time of each zone differing from the neighboring ones by one hour. On the surface of the Earth, one hour

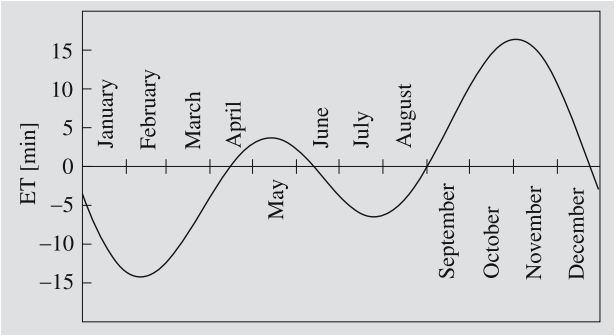


Fig. 2.30. Equation of time. A sundial always shows (if correctly installed) true local solar time. To find the local mean time the equation of time must be subtracted from the local solar time

in time corresponds to 15° in longitude; the time of each zone is determined by the local mean time at one of the longitudes 0°, 15°, . . . , 345°.

The time of the zero meridian going through Greenwich is used as an international reference, Universal Time. In most European countries, time is one hour ahead of this (Fig. 2.31).

In summer, many countries switch to *daylight saving time*, during which time is one hour ahead of the ordinary time. The purpose of this is to make the time when people are awake coincide with daytime in order to save electricity, particularly in the evening, when people go to bed one hour earlier. During daylight saving time, the difference between the true solar time and the official time can grow even larger.

In the EU countries the daylight saving time begins on the last Sunday of March, at 1 o'clock UTC in the morning, when the clocks are moved forward to read 2 AM, and ends on the last Sunday of October at 1 o'clock.

2.14 Astronomical Time Systems

Time can be defined using several different phenomena:

1. The solar and sidereal times are based on the rotation of the Earth.
2. The standard unit of time in the current SI system, the second, is based on quantum mechanical atomary phenomena.
3. Equations of physics like the ones describing the motions of celestial bodies involve a time variable

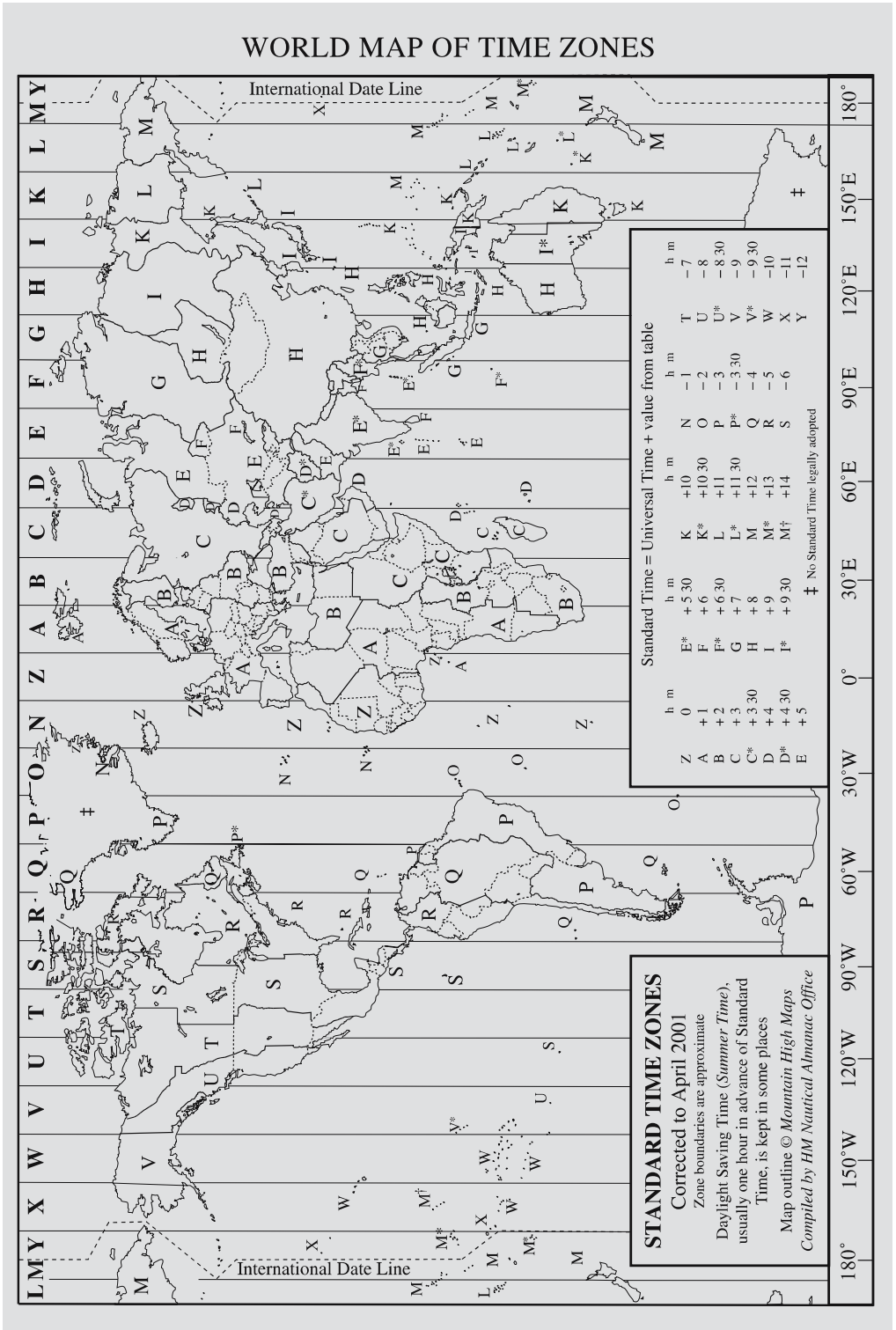


Fig. 2.31. The time zones. The map gives the difference of the local zonal time from the Greenwich mean time (UT). During daylight saving time, one hour must be added to the given figures. When travelling across the date line westward, the date must be incremented by one day, and decremented if going eastward. For example, a traveller taking a flight from Honolulu to Tokyo on Monday morning will arrive on Tuesday, even though (s)he does not see a single night en route. (Drawing U.S. Nval Observatory)

corresponding to an ideal time running at a constant pace. The ephemeris time and dynamical time discussed a little later are such times.

Observations give directly the apparent sidereal time as the hour angle of the true vernal equinox. From the apparent sidereal time the mean sidereal time can be calculated.

The *universal time* UT is defined by the equation

$$\begin{aligned}\mathrm{GMST}(0\,\mathrm{UT}) = {} & 24{,}110.54841\ \mathrm{s} \\ & + T \times 8{,}640{,}184.812866\ \mathrm{s} \\ & + T^2 \times 0.093104\ \mathrm{s} \\ & - T^3 \times 0.0000062\ \mathrm{s}\ ,\end{aligned} \tag{2.47}$$

where GMST is the Greenwich mean sidereal time and T the Julian century. The latter is obtained from the Julian date J, which is a running number of the day (Sect. 2.15 and *Julian date, p. 41):

$$T = \frac{J - 2{,}451{,}545.0}{36{,}525}\ . \tag{2.48}$$

This gives the time elapsed since January 1, 2000, in Julian centuries.

Sidereal time and hence also UT are related to the rotation of the Earth, and thus contain perturbations due to the irregular variations, mainly slowing down, of the rotation.

In (2.47) the constant 8,640,184.812866 s tells how much sidereal time runs fast compared to the UT in a Julian century. As the rotation of the Earth is slowing down the solar day becomes longer. Since the Julian century T contains a fixed number of days, it will also become longer. This gives rise to the small correction terms in (2.47).

Strictly speaking this universal time is the time denoted by UT1. Observations give UT0, which contains a small perturbation due to the wandering of the geographical pole, or *polar variation*. The direction of the axis with respect to the solid surface varies by about $0.1''$ (a few metres on the surface) with a period of about 430 days (*Chandler period*). In addition to this, the polar motion contains a slow nonperiodic part.

The z axis of the astronomical coordinates is aligned with the angular momentum vector of the Earth, but the terrestrial coordinates refer to the axis at the epoch 1903.5. In the most accurate calculations this has to be taken into account.

Nowadays the SI unit of time, the second, is defined in a way that has nothing to do with celestial phenomena. Periods of quantum mechanical phenomena remain more stable than the motions of celestial bodies involving complicated perturbations.

In 1967, one second was defined as 9,192,631,770 times the period of the light emitted by cesium 133 isotope in its ground state, transiting from hyperfine level $F = 4$ to $F = 3$. Later, this definition was revised to include small relativistic effects produced by gravitational fields. The relative accuracy of this atomic time is about 10^{-12}.

The *international atomic time*, TAI, was adopted as the basis of time signals in 1972. The time is maintained by the *Bureau International des Poids et Mesures* in Paris, and it is the average of several accurate atomic clocks.

Even before atomic clocks there was a need for an ideal time proceeding at a perfectly constant rate, corresponding to the time variable in the equations of Newtonian mechanics. The *ephemeris time* was such a time. It was used e. g. for tabulating ephemerides. The unit of ephemeris time was the *ephemeris second*, which is the length of the tropical year 1900 divided by 31,556,925.9747. Ephemeris time was not known in advance. Only afterwards was it possible to determine the difference of ET and UT from observational data.

In 1984 ephemeris time was replaced by *dynamical time*. It comes in two varieties.

The *terrestrial dynamical time* (TDT) corresponds to the proper time of an observer moving with the Earth. The time scale is affected by the relativistic time dilation due to the orbital speed of the Earth. The rotation velocity depends on the latitude, and thus in TDT it is assumed that the observer is not rotating with the Earth. The zero point of TDT was chosen so that the old ET changed without a jump to TDT.

In 1991 a new standard time, the *terrestrial time* (TT), was adopted. Practically it is equivalent to TDT.

TT (or TDT) is the time currently used for tabulating ephemerides of planets and other celestial bodies. For example, the *Astronomical Almanac* gives the coordinates of the planets for each day at 0 TT.

The *Astronomical Almanac* also gives the difference

$$\Delta T = \mathrm{TDT} - \mathrm{UT} \tag{2.49}$$

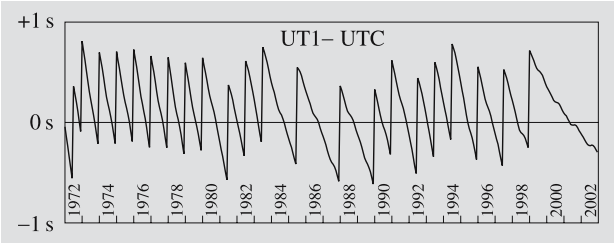


Fig. 2.32. The difference between the universal time UT1, based on the rotation of the Earth, and the coordinated universal time UTC during 1972–2002. Because the rotation of the Earth is slowing down, the UT1 will run slow of the UTC by about 0.8 seconds a year. Leap seconds are added to the UTC when necessary to keep the times approximately equal. In the graph these leap seconds are seen as one second jumps upward

for earlier years. For the present year and some future years a prediction extrapolated from the earlier years is given. Its accuracy is about 0.1 s. At the beginning of 1990 the difference was 56.7 s; it increases every year by an amount that is usually a little less than one second.

The terrestrial time differs from the atomic time by a constant offset

$$\mathrm{TT} = \mathrm{TAI} + 32.184\ \mathrm{s}\ . \tag{2.50}$$

TT is well suited for ephemerides of phenomena as seen from the Earth. The equations of motion of the solar system, however, are solved in a frame the origin of which is the centre of mass or *barycentre* of the solar system. The coordinate time of this frame is called the *barycentric dynamical time*, TDB. The unit of TDB is defined so that, on the average, it runs at the same rate as TT, the difference containing only periodic terms depending on the orbital motion of the Earth. The difference can usually be neglected, since it is at most about 0.002 seconds.

Which of these many times should we use in our alarm-clocks? None of them. Yet another time is needed for that purpose. This official wall-clock time is called the *coordinated universal time*, UTC. The zonal time follows UTC but differs from it usually by an integral number of hours.

UTC is defined so that it proceeds at the same rate as TAI, but differs from it by an integral number of seconds. These *leap seconds* are used to adjust UTC so that the difference from UT1 never exceeds 0.9 seconds. A leap second is added either at the beginning of a year or the night between June and July.

The difference

$$\Delta\mathrm{AT} = \mathrm{TAI} - \mathrm{UTC} \tag{2.51}$$

is also tabulated in the *Astronomical Almanac*. According to the definition of UTC the difference in seconds is always an integer. The difference cannot be predicted very far to the future.

From (2.50) and (2.51) we get

$$\mathrm{TT} = \mathrm{UTC} + 32.184\ \mathrm{s} + \Delta\mathrm{AT}\ , \tag{2.52}$$

which gives the terrestrial time TT corresponding to a given UTC. Table 2.2 gives this correction. The table is easy to extend to the future. When it is told in the news that a leap second will be added the difference will increase by one second. In case the number of leap seconds is not known, it can be approximated that a leap second will be added every 1.25 years.

The unit of the coordinated universal time UTC, atomic time TAI and terrestrial time TT is the same

Table 2.2. Differences of the atomic time and UTC (ΔAT) and the terrestrial time TT and UTC. The terrestrial time TT used in ephemerides is obtained by adding $\Delta\mathrm{AT} + 32.184$ s to the ordinary time UTC

	ΔAT	TT – UTC
1.1.1972– 30.6.1972	10 s	42.184 s
1.7.1972–31.12.1972	11 s	43.184 s
1.1.1973–31.12.1973	12 s	44.184 s
1.1.1974–31.12.1974	13 s	45.184 s
1.1.1975–31.12.1975	14 s	46.184 s
1.1.1976–31.12.1976	15 s	47.184 s
1.1.1977–31.12.1977	16 s	48.184 s
1.1.1978–31.12.1978	17 s	49.184 s
1.1.1979–31.12.1979	18 s	50.184 s
1.1.1980– 30.6.1981	19 s	51.184 s
1.7.1981– 30.6.1982	20 s	52.184 s
1.7.1982– 30.6.1983	21 s	53.184 s
1.7.1983– 30.6.1985	22 s	54.184 s
1.7.1985–31.12.1987	23 s	55.184 s
1.1.1988–31.12.1989	24 s	56.184 s
1.1.1990–31.12.1990	25 s	57.184 s
1.1.1991– 30.6.1992	26 s	58.184 s
1.7.1992– 30.6.1993	27 s	59.184 s
1.7.1993– 30.6.1994	28 s	60.184 s
1.7.1994–31.12.1995	29 s	61.184 s
1.1.1996– 31.6.1997	30 s	62.184 s
1.7.1997–31.12.1998	31 s	63.184 s
1.1.1999–31.12.2005	32 s	64.184 s
1.1.2006–	33 s	65.184 s

second of the SI system. Hence all these times proceed at the same rate, the only difference being in their zero points. The difference of the TAI and TT is always the same, but due to the leap seconds the UTC will fall behind in a slightly irregular way.

Culminations and rising and setting times of celestial bodies are related to the rotation of the Earth. Thus the sidereal time and hence the UT of such an event can be calculated precisely. The corresponding UTC cannot differ from the UT by more than 0.9 seconds, but the exact value is not known in advance. The future coordinates of the Sun, Moon and planets can be calculated as functions of the TT, but the corresponding UTC can only be estimated.

2.15 Calendars

Our calendar is a result of long evolution. The main problem it must contend with is the incommensurability of the basic units, day, month and year: the numbers of days and months in a year are not integers. This makes it rather complicated to develop a calendar that takes correctly into account the alternation of seasons, day and night, and perhaps also the lunar phases.

Our calendar has its origin in the Roman calendar, which, in its earliest form, was based on the phases of the Moon. From around 700 B.C. on, the length of the year has followed the apparent motion of the Sun; thus originated the division of the year into twelve months. One month, however, still had a length roughly equal to the lunar cycle. Hence one year was only 354 days long. To keep the year synchronised with the seasons, a leap month had to be added to every other year.

Eventually the Roman calendar got mixed up. The mess was cleared by Julius Caesar in about 46 B.C., when the *Julian calendar* was developed upon his orders. The year had 365 days and a leap day was added to every fourth year.

In the Julian calendar, the average length of one year is 365 d 6 h, but the tropical year is 11 min 14 s shorter. After 128 years, the Julian year begins almost one day too late. The difference was already 10 days in 1582, when a calendar reform was carried out by Pope Gregory XIII. In the *Gregorian calendar*, every fourth year is a leap year, the years divisible by 100 being exceptions. Of these, only the years divisible by 400 are leap years. Thus 1900 was not a leap year, but 2000 was. The Gregorian calendar was adopted slowly, at different times in different countries. The transition period did not end before the 20th century.

Even the Gregorian calendar is not perfect. The differences from the tropical year will accumulate to one day in about 3300 years.

Since years and months of variable length make it difficult to compute time differences, especially astronomers have employed various methods to give each day a running number. The most widely used numbers are the *Julian dates*. In spite of their name, they are not related to the Julian calendar. The only connection is the length of a *Julian century* of 36,525 days, a quantity appearing in many formulas involving Julian dates. The Julian day number 0 dawned about 4700 B.C. The day number changes always at 12 : 00 UT. For example, the Julian day 2,451,545 began at noon in January 1, 2000. The Julian date can be computed using the formulas given in *Julian Date (p. 41).

Julian dates are uncomfortably big numbers, and therefore *modified Julian dates* are often used. The zero point can be e. g. January 1, 2000. Sometimes 0.5 is subtracted from the date to make it to coincide with the date corresponding to the UTC. When using such dates, the zero point should always be mentioned.

* Reduction of Coordinates

Star catalogues give coordinates for some standard epoch. In the following we give the formulas needed to reduce the coordinates to a given date and time. The full reduction is rather laborious, but the following simplified version is sufficient for most practical purposes.

We assume that the coordinates are given for the epoch J2000.0.

1. First correct the place for proper motion unless it is negligible.
2. Precess the coordinates to the time of the observation. First we use the coordinates of the standard epoch (α_0, δ_0) to find a unit vector pointing in the direction of the star:

$$\boldsymbol{p}_0 = \begin{pmatrix} \cos\delta_0 \cos\alpha_0 \\ \cos\delta_0 \sin\alpha_0 \\ \sin\delta_0 \end{pmatrix} .$$

Precession changes the ecliptic longitude of the object. The effect on right ascension and declination can be calculated as three rotations, given by three rotation matrices. By multiplying these matrices we get the combined precession matrix that maps the previous unit vector to its precessed equivalent. A similar matrix can be derived for the nutation. The transformations and constants given here are based on the system standardized by the IAU in 1976.
The precession and nutation matrices contain several quantities depending on time. The time variables appearing in their expressions are

$$t = J - 2{,}451{,}545.0 \,,$$
$$T = \frac{J - 2{,}451{,}545.0}{36{,}525} \,.$$

Here J is the Julian date of the observation, t the number of days since the epoch J2000.0 (i. e. noon of January 1, 2000), and T the same interval of time in Julian centuries.
The following three angles are needed for the precession matrix

$$\zeta = 2306.2181'' T + 0.30188'' T^2 + 0.017998'' T^3 \,,$$
$$z = 2306.2181'' T + 1.09468'' T^2 + 0.018203'' T^3 \,,$$
$$\theta = 2004.3109'' T - 0.42665'' T^2 - 0.041833'' T^3 \,.$$

The precession matrix is now

$$\boldsymbol{P} = \begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix} \,.$$

The elements of this matrix in terms of the abovementioned angles are

$$P_{11} = \cos z \cos\theta \cos\zeta - \sin z \sin\zeta \,,$$
$$P_{12} = -\cos z \cos\theta \sin\zeta - \sin z \cos\zeta \,,$$
$$P_{13} = -\cos z \sin\theta \,,$$
$$P_{21} = \sin z \cos\theta \cos\zeta + \cos z \sin\zeta \,,$$
$$P_{22} = -\sin z \cos\theta \sin\zeta + \cos z \cos\zeta \,,$$
$$P_{23} = -\sin z \sin\theta \,,$$
$$P_{31} = \sin\theta \cos\zeta \,,$$
$$P_{32} = -\sin\theta \sin\zeta \,,$$
$$P_{33} = \cos\theta \,.$$

The new coordinates are now obtained by multiplying the coordinates of the standard epoch by the precession matrix:

$$\boldsymbol{p}_1 = \boldsymbol{P}\boldsymbol{p}_0 \,.$$

This is the mean place at the given time and date.
If the standard epoch is not J2000.0, it is probably easiest to first transform the given coordinates to the epoch J2000.0. This can be done by computing the precession matrix for the given epoch and multiplying the coordinates by the inverse of this matrix. Inverting the precession matrix is easy: we just transpose it, i. e. interchange its rows and columns. Thus coordinates given for some epoch can be precessed to J2000.0 by multiplying them by

$$\boldsymbol{P}^{-1} = \begin{pmatrix} P_{11} & P_{21} & P_{31} \\ P_{12} & P_{22} & P_{32} \\ P_{13} & P_{23} & P_{33} \end{pmatrix} \,.$$

In case the required accuracy is higher than about one minute of arc, we have to do the following further corrections.

3. The full nutation correction is rather complicated. The nutation used in astronomical almanacs involves series expansions containing over a hundred terms. Very often, though, the following simple form is sufficient. We begin by finding the mean obliquity of the ecliptic at the observation time:

$$\varepsilon_0 = 23^\circ\ 26'\ 21.448'' - 46.8150'' T - 0.00059'' T^2 + 0.001813'' T^3 \,.$$

The mean obliquity means that periodic perturbations have been omitted. The formula is valid a few centuries before and after the year 2000.
The true obliquity of the ecliptic, ε, is obtained by adding the nutation correction to the mean obliquity:

$$\varepsilon = \varepsilon_0 + \Delta\varepsilon \,.$$

The effect of the nutation on the ecliptic longitude (denoted usually by $\Delta\psi$) and the obliquity of the ecliptic can be found from

$$C_1 = 125^\circ - 0.05295^\circ t \,,$$
$$C_2 = 200.9^\circ + 1.97129^\circ t \,,$$
$$\Delta\psi = -0.0048^\circ \sin C_1 - 0.0004^\circ \sin C_2 \,,$$
$$\Delta\varepsilon = 0.0026^\circ \cos C_1 + 0.0002^\circ \cos C_2 \,.$$

Since $\Delta\psi$ and $\Delta\varepsilon$ are very small angles, we have, for example, $\sin\Delta\psi \approx \Delta\psi$ and $\cos\Delta\psi \approx 1$, when the angles are expressed in radians. Thus we get the nutation matrix

$$\boldsymbol{N} = \begin{pmatrix} 1 & -\Delta\psi\cos\varepsilon & -\Delta\psi\sin\varepsilon \\ \Delta\psi\cos\varepsilon & 1 & -\Delta\varepsilon \\ \Delta\psi\sin\varepsilon & \Delta\varepsilon & 1 \end{pmatrix} .$$

This is a linearized version of the full transformation. The angles here must be in radians. The place in the coordinate frame of the observing time is now

$$\boldsymbol{p}_2 = \boldsymbol{N}\boldsymbol{p}_1 .$$

4. The annual aberration can affect the place about as much as the nutation. Approximate corrections are obtained from

$$\begin{aligned} \Delta\alpha\cos\delta = &-20.5''\sin\alpha\sin\lambda \\ &-18.8''\cos\alpha\cos\lambda , \\ \Delta\delta = &\ 20.5''\cos\alpha\sin\delta\sin\lambda \\ &+18.8''\sin\alpha\sin\delta\cos\lambda - 8.1''\cos\delta\cos\lambda , \end{aligned}$$

where λ is the ecliptic longitude of the Sun. Sufficiently accurate value for this purpose is given by

$$\begin{aligned} G &= 357.528^\circ + 0.985600^\circ t , \\ \lambda &= 280.460^\circ + 0.985647^\circ t \\ &\quad + 1.915^\circ\sin G + 0.020^\circ\sin 2G . \end{aligned}$$

These reductions give the apparent place of the date with an accuracy of a few seconds of arc. The effects of parallax and diurnal aberration are even smaller.

Example. The coordinates of Regulus (α Leo) for the epoch J2000.0 are

$$\begin{aligned} \alpha &= 10\,\mathrm{h}\ 8\,\mathrm{min}\ 22.2\,\mathrm{s} = 10.139500\,\mathrm{h} , \\ \delta &= 11^\circ\ 58'\ 02'' = 11.967222^\circ . \end{aligned}$$

Find the apparent place of Regulus on March 12, 1995.

We start by finding the unit vector corresponding to the catalogued place:

$$\boldsymbol{p}_0 = \begin{pmatrix} -0.86449829 \\ 0.45787318 \\ 0.20735204 \end{pmatrix} .$$

The Julian date is $J = 2{,}449{,}789.0$, and thus $t = -1756$ and $T = -0.04807666$. The angles of the precession matrix are $\zeta = -0.03079849^\circ$, $z = -0.03079798^\circ$ and $\theta = -0.02676709^\circ$. The precession matrix is then

$$\boldsymbol{P} = \begin{pmatrix} 0.99999931 & 0.00107506 & 0.00046717 \\ -0.00107506 & 0.99999942 & -0.00000025 \\ -0.00046717 & -0.00000025 & 0.99999989 \end{pmatrix} .$$

The precessed unit vector is

$$\boldsymbol{p}_1 = \begin{pmatrix} -0.86390858 \\ 0.45880225 \\ 0.20775577 \end{pmatrix} .$$

The angles needed for the nutation are $\Delta\psi = 0.00309516^\circ$, $\Delta\varepsilon = -0.00186227^\circ$, $\varepsilon = 23.43805403^\circ$, which give the nutation matrix

$$\boldsymbol{N} = \begin{pmatrix} 1 & -0.00004956 & -0.00002149 \\ 0.00004956 & 1 & 0.00003250 \\ 0.00002149 & -0.00003250 & 1 \end{pmatrix} .$$

The place in the frame of the date is

$$\boldsymbol{p}_2 = \begin{pmatrix} -0.86393578 \\ 0.45876618 \\ 0.20772230 \end{pmatrix} ,$$

whence

$$\begin{aligned} \alpha &= 10.135390\,\mathrm{h} , \\ \delta &= 11.988906^\circ . \end{aligned}$$

To correct for the aberration we first find the longitude of the Sun: $G = -1373.2^\circ = 66.8^\circ$, $\lambda = -8.6^\circ$. The correction terms are then

$$\begin{aligned} \Delta\alpha &= 18.25'' = 0.0050^\circ \\ \Delta\delta &= -5.46'' = -0.0015^\circ . \end{aligned}$$

Adding these to the previously obtained coordinates we get the apparent place of Regulus on March 12, 1995:

$$\begin{aligned} \alpha &= 10.1357\,\mathrm{h} = 10\,\mathrm{h}\ 8\,\mathrm{min}\ 8.5\,\mathrm{s}, \\ \delta &= 11.9874^\circ = 11^\circ\ 59'\ 15'' . \end{aligned}$$

Comparison with the places given in the catalogue *Apparent Places of Fundamental Stars* shows that we are within about $3''$ of the correct place, which is a satisfactory result.

*** Julian Date**

There are several methods for finding the Julian date. The following one, developed by Fliegel and Van Flandern in 1968, is well adapted for computer programs. Let y be the year (with all four digits), m the month and d the day. The Julian date J at noon is then

$$\begin{aligned} J = {} & 367y - \{7[y + (m+9)/12]\}/4 \\ & - \big(3\{[y + (m-9)/7]/100 + 1\}\big)/4 \\ & + 275m/9 + d + 1721029 \,. \end{aligned}$$

The division here means an integer division, the decimal part being truncated: e. g. $7/3 = 2$ and $-7/3 = -2$.

Example. Find the Julian date on January 1, 1990.
Now $y = 1990$, $m = 1$ and $d = 1$.

$$\begin{aligned} J &= 367 \times 1990 - 7 \times [1990 + (1+9)/12]/4 \\ &\quad - 3 \times \{[1990 + (1-9)/7]/100 + 1\}/4 \\ &\quad + 275 \times 1/9 + 1 + 1{,}721{,}029 \\ &= 730{,}330 - 3482 - 15 + 30 + 1 + 1{,}721{,}029 \\ &= 2{,}447{,}893 \,. \end{aligned}$$

Astronomical tables usually give the Julian date at 0 UT. In this case that would be 2,447,892.5.

The inverse procedure is a little more complicated. In the following J is the Julian date at noon (so that it will be an integer):

$$\begin{aligned} a &= J + 68{,}569 \,, \\ b &= (4a)/146{,}097 \,, \\ c &= a - (146{,}097b + 3)/4 \,, \\ d &= [4000(c+1)]/1{,}461{,}001 \,, \\ e &= c - (1461d)/4 + 31 \,, \\ f &= (80e)/2447 \,, \\ \text{day} &= e - (2447f)/80 \,, \\ g &= f/11 \,, \\ \text{month} &= f + 2 - 12g \,, \\ \text{year} &= 100(b-49) + d + g \,. \end{aligned}$$

Example. In the previous example we got $J = 2{,}447{,}893$. Let's check this by calculating the corresponding calendar date:

$$\begin{aligned} a &= 2{,}447{,}893 + 68{,}569 = 2{,}516{,}462 \,, \\ b &= (4 \times 2{,}516{,}462)/146{,}097 = 68 \,, \\ c &= 2{,}516{,}462 - (146{,}097 \times 68 + 3)/4 = 32{,}813 \,, \\ d &= [4000(32{,}813+1)]/1{,}461{,}001 = 89 \,, \\ e &= 32{,}813 - (1461 \times 89)/4 + 31 = 337 \,, \\ f &= (80 \times 337)/2447 = 11 \,, \\ \text{day} &= 337 - (2447 \times 11)/80 = 1 \,, \\ g &= 11/11 = 1 \,, \\ \text{month} &= 11 + 2 - 12 \times 1 = 1 \,, \\ \text{year} &= 100(68-49) + 89 + 1 = 1990 \,. \end{aligned}$$

Thus we arrived back to the original date.

Since the days of the week repeat in seven day cycles, the remainder of the division $J/7$ unambiguously determines the day of the week. If J is the Julian date at noon, the remainder of $J/7$ tells the day of the week in the following way:

$$\begin{aligned} 0 &= \text{Monday} \,, \\ &\vdots \\ 5 &= \text{Saturday} \,, \\ 6 &= \text{Sunday} \,. \end{aligned}$$

Example. The Julian date corresponding to January 1, 1990 was 2,447,893. Since $2{,}447{,}893 = 7 \times 349{,}699$, the remainder is zero, and the day was Monday.

2.16 Examples

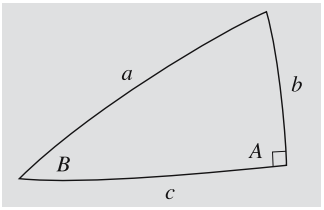
Example 2.1 *Trigonometric Functions in a Rectangular Spherical Triangle*

Let the angle A be a right angle. When the figure is a plane triangle, the trigonometric functions of the angle B would be:

$$\sin B = b/a \,, \quad \cos B = c/a \,, \quad \tan B = b/c \,.$$

For the spherical triangle we have to use the equations in (2.7), which are now simply:

$$\sin B \sin a = \sin b \,,$$
$$\cos B \sin a = \cos b \sin c \,,$$
$$\cos a = \cos b \cos c \,.$$



The first equation gives the sine of B:

$$\sin B = \sin b / \sin a \,.$$

Dividing the second equation by the third one, we get the cosine of B:

$$\cos B = \tan c / \tan a \,.$$

And the tangent is obtained by dividing the first equation by the second one:

$$\tan B = \tan b / \sin c \,.$$

The third equation is the equivalent of the Pythagorean theorem for rectangular triangles.

Example 2.2 *The Coordinates of New York City*

The geographic coordinates are 41° north and 74° west of Greenwich, or $\phi = +41°$, $\lambda = -74°$. In time units, the longitude would be $74/15\,\mathrm{h} = 4\,\mathrm{h}\ 56\,\mathrm{min}$ west of Greenwich. The geocentric latitude is obtained from

$$\tan\phi' = \frac{b^2}{a^2}\tan\phi = \left(\frac{6{,}356{,}752}{6{,}378{,}137}\right)^2 \tan 41° = 0.86347 \quad \Rightarrow \quad \phi' = 40°\ 48'\ 34'' \,.$$

The geocentric latitude is $11'\ 26''$ less than the geographic latitude.

Example 2.3 The angular separation of two objects in the sky is quite different from their coordinate difference.

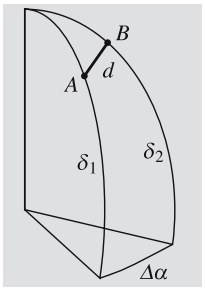
Suppose the coordinates of a star A are $\alpha_1 = 10\,\mathrm{h}$, $\delta_1 = 70°$ and those of another star B, $\alpha_2 = 11\,\mathrm{h}$, $\delta_2 = 80°$.

Using the Pythagorean theorem for plane triangles, we would get

$$d = \sqrt{(15°)^2 + (10°)^2} = 18° \,.$$

But if we use the third equation in (2.7), we get

$$\begin{aligned}\cos d &= \cos(\alpha_1 - \alpha_2) \\ &\quad \times \sin(90° - \delta_1)\sin(90° - \delta_2) \\ &\quad + \cos(90° - \delta_1)\cos(90° - \delta_2) \\ &= \cos(\alpha_1 - \alpha_2)\cos\delta_1\cos\delta_2 \\ &\quad + \sin\delta_1 \sin\delta_2 \\ &= \cos 15° \cos 70° \cos 80° \\ &\quad + \sin 70° \sin 80° \\ &= 0.983 \,,\end{aligned}$$



which yields $d = 10.6°$. The figure shows why the result obtained from the Pythagorean theorem is so far from being correct: hour circles (circles with $\alpha = \text{constant}$) approach each other towards the poles and their angular separation becomes smaller, though the coordinate difference remains the same.

Example 2.4 Find the altitude and azimuth of the Moon in Helsinki at midnight at the beginning of 1996.

The right ascension is $\alpha = 2\,\mathrm{h}\ 55\,\mathrm{min}\ 7\,\mathrm{s} = 2.9186\,\mathrm{h}$ and declination $\delta = 14°\ 42' = 14.70°$, the sidereal time is $\Theta = 6\,\mathrm{h}\ 19\,\mathrm{min}\ 26\,\mathrm{s} = 6.3239\,\mathrm{h}$ and latitude $\phi = 60.16°$.

The hour angle is $h = \Theta - \alpha = 3.4053\,\mathrm{h} = 51.08°$. Next we apply the equations in (2.16):

$$\begin{aligned}\sin A \cos a &= \sin 51.08° \cos 14.70° = 0.7526 \,, \\ \cos A \cos a &= \cos 51.08° \cos 14.70° \sin 60.16° \\ &\quad - \sin 14.70° \cos 60.16° \\ &= 0.4008 \,, \\ \sin a &= \cos 51.08° \cos 14.70° \cos 60.16° \\ &\quad + \sin 14.70° \sin 60.16° \\ &= 0.5225.\end{aligned}$$

Thus the altitude is $a = 31.5°$. To find the azimuth we have to compute its sine and cosine:

$$\sin A = 0.8827 \,, \quad \cos A = 0.4701 \,.$$

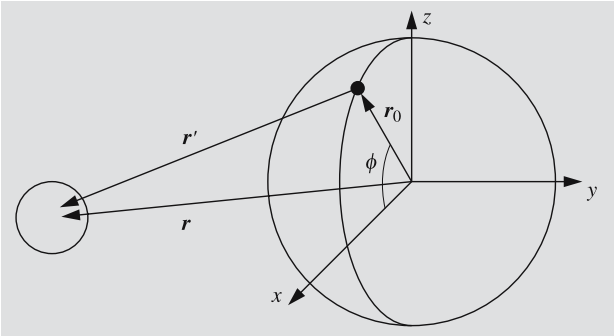
Hence the azimuth is $A = 62.0°$. The Moon is in the southwest, 31.5 degrees above the horizon. Actually, this would be the direction if the Moon were infinitely distant.

Example 2.5 Find the topocentric place of the Moon in the case of the previous example.

The geocentric distance of the Moon at that time is $R = 62.58$ equatorial radii of the Earth. For simplicity, we can assume that the Earth is spherical.

We set up a rectangular coordinate frame in such a way that the z axis points towards the celestial pole and the observing site is in the xz plane. When the radius of the Earth is used as the unit of distance, the radius vector of the observing site is

$$\boldsymbol{r}_0 = \begin{pmatrix} \cos\phi \\ 0 \\ \sin\phi \end{pmatrix} = \begin{pmatrix} 0.4976 \\ 0 \\ 0.8674 \end{pmatrix} .$$



The radius vector of the Moon is

$$\boldsymbol{r} = R \begin{pmatrix} \cos\delta\cos h \\ -\cos\delta\sin h \\ \sin\delta \end{pmatrix} = 62.58 \begin{pmatrix} 0.6077 \\ -0.7526 \\ 0.2538 \end{pmatrix} .$$

The topocentric place of the Moon is

$$\boldsymbol{r}' = \boldsymbol{r} - \boldsymbol{r}_0 = \begin{pmatrix} 37.53 \\ -47.10 \\ 15.02 \end{pmatrix} .$$

We divide this vector by its length 62.07 to get the unit vector $\boldsymbol{e}$ pointing to the direction of the Moon. This can be expressed in terms of the topocentric coordinates δ' and h':

$$\boldsymbol{e} = \begin{pmatrix} 0.6047 \\ -0.7588 \\ 0.2420 \end{pmatrix} = \begin{pmatrix} \cos\delta'\cos h' \\ -\cos\delta'\sin h' \\ \sin\delta' \end{pmatrix} ,$$

which gives $\delta' = 14.00°$ and $h' = 51.45°$. Next we can calculate the altitude and azimuth as in the previous example, and we get $a = 30.7°$, $A = 61.9°$.

Another way to find the altitude is to take the scalar product of the vectors $\boldsymbol{e}$ and $\boldsymbol{r}_0$, which gives the cosine of the zenith distance:

$$\begin{aligned} \cos z &= \boldsymbol{e}\cdot\boldsymbol{r}_0 = 0.6047\times 0.4976 + 0.2420\times 0.8674 \\ &= 0.5108 , \end{aligned}$$

whence $z = 59.3°$ and $a = 90° - z = 30.7°$. We see that this is 0.8° less than the geocentric altitude; i.e. the difference is more than the apparent diameter of the Moon.

Example 2.6 The coordinates of Arcturus are $\alpha = 14\text{ h }15.7\text{ min}$, $\delta = 19°\ 1'1'$. Find the sidereal time at the moment Arcturus rises or sets in Boston ($\phi = 42°\ 19'$).

Neglecting refraction, we get

$$\begin{aligned} \cos h &= -\tan 19°\ 11'\tan 42°\ 19' \\ &= -0.348\times 0.910 = -0.317 . \end{aligned}$$

Hence, $h = \pm 108.47° = 7\text{ h }14\text{ min}$. The more accurate result is

$$\begin{aligned} \cos h &= -\tan 19°\ 11'\tan 42°\ 19' \\ &\quad - \frac{\sin 35'}{\cos 19°\ 11'\cos 42°19'} \\ &= -0.331 , \end{aligned}$$

whence $h = \pm 109.35° = 7\text{ h }17\text{ min}$. The plus and minus signs correspond to setting ant rising, respectively. When Arcturus rises, the sidereal time is

$$\begin{aligned} \Theta &= \alpha + h = 14\text{ h }16\text{ min} - 7\text{ h }17\text{ min} \\ &= 6\text{ h }59\text{ min} \end{aligned}$$

and when it sets, the sidereal time is

$$\begin{aligned} \Theta &= 14\text{ h }16\text{ min} + 7\text{ h }17\text{ min} \\ &= 21\text{ h }33\text{ min} . \end{aligned}$$

Note that the result is independent of the date: a star rises and sets at the same sidereal time every day.

Example 2.7 The proper motion of Aldebaran is $\mu = 0.20''/\text{a}$ and parallax $\pi = 0.048''$. The spectral line of iron at $\lambda = 440.5\text{ nm}$ is displaced 0.079 nm towards the

red. What are the radial and tangential velocities and the total velocity?

The radial velocity is found from

$$\frac{\Delta\lambda}{\lambda} = \frac{v_\mathrm{r}}{c}$$
$$\Rightarrow\ v_\mathrm{r} = \frac{0.079}{440.5}\cdot 3\times 10^8\ \mathrm{m/s} = 5.4\times 10^4\ \mathrm{m/s}$$
$$= 54\ \mathrm{km/s}\ .$$

The tangential velocity is now given by (2.40), since μ and π are in correct units:

$$v_\mathrm{t} = 4.74\mu r = 4.74\mu/\pi = \frac{4.74\times 0.20}{0.048} = 20\ \mathrm{km/s}\ .$$

The total velocity is

$$v = \sqrt{v_\mathrm{r}^2 + v_\mathrm{t}^2} = \sqrt{54^2 + 20^2}\ \mathrm{km/s} = 58\ \mathrm{km/s}\ .$$

Example 2.8 Find the local time in Paris (longitude $\lambda = 2^\circ$) at 12:00.

Local time coincides with the zonal time along the meridian 15° east of Greenwich. Longitude difference $15^\circ - 2^\circ = 13^\circ$ equals $(13^\circ/15^\circ)\times 60\ \mathrm{min} = 52$ minutes. The local time is 52 minutes less than the official time, or 11:08. This is mean solar time. To find the true solar time, we must add the equation of time. In early February, E.T. $= -14$ min and the true solar time is $11{:}08 - 14\ \mathrm{min} = 10{:}54$. At the beginning of November, ET $= +16$ min and the solar time would be 11:24. Since -14 min and $+16$ min are the extreme values of E.T., the true solar time is in the range 10:54–11:24, the exact time depending on the day of the year. During daylight saving time, we must still subtract one hour from these times.

Example 2.9 *Estimating Sidereal Time*

Since the sidereal time is the hour angle of the vernal equinox ♈, it is 0 h when ♈ culminates or transits the south meridian. At the moment of the vernal equinox, the Sun is in the direction of ♈ and thus culminates at the same time as ♈. So the sidereal time at 12:00 local solar time is 0:00, and at the time of the vernal equinox, we have

$$\Theta = T + 12\ \mathrm{h}\ ,$$

where T is the local solar time. This is accurate within a couple of minutes. Since the sidereal time runs about 4 minutes fast a day, the sidereal time, n days after the vernal equinox, is

$$\Theta \approx T + 12\ \mathrm{h} + n\times 4\ \mathrm{min}\ .$$

At autumnal equinox ♈ culminates at 0:00 local time, and sidereal and solar times are equal.

Let us try to find the sidereal time in Paris on April 15 at 22:00, Central European standard time (= 23:00 daylight saving time). The vernal equinox occurs on the average on March 21; thus the time elapsed since the equinox is $10 + 15 = 25$ days. Neglecting the equation of time, the local time T is 52 minutes less than the zonal time. Hence

$$\Theta = T + 12\ \mathrm{h} + n\times 4\ \mathrm{min}$$
$$= 21\ \mathrm{h}\ 8\ \mathrm{min} + 12\ \mathrm{h} + 25\times 4\ \mathrm{min}$$
$$= 34\ \mathrm{h}\ 48\ \mathrm{min} = 10\ \mathrm{h}\ 48\ \mathrm{min}\ .$$

The time of the vernal equinox can vary about one day in either direction from the average. Therefore the accuracy of the result is roughly 5 min.

Example 2.10 Find the rising time of Arcturus in Boston on January 10.

In Example 2.6 we found the sidereal time of this event, $\Theta = 6\ \mathrm{h}\ 59\ \mathrm{min}$. Since we do not know the year, we use the rough method of Example 2.9. The time between January 1 and vernal equinox (March 21) is about 70 days. Thus the sidereal time on January 1 is

$$\Theta \approx T + 12\ \mathrm{h} - 70\times 4\ \mathrm{min} = T + 7\ \mathrm{h}\ 20\ \mathrm{min}\ ,$$

from which

$$T = \Theta - 7\ \mathrm{h}\ 20\ \mathrm{min} = 6\ \mathrm{h}\ 59\ \mathrm{min} - 7\ \mathrm{h}\ 20\ \mathrm{min}$$
$$= 30\ \mathrm{h}\ 59\ \mathrm{min} - 7\ \mathrm{h}\ 20\ \mathrm{min} = 23\ \mathrm{h}\ 39\ \mathrm{min}\ .$$

The longitude of Boston is 71° W, and the Eastern standard time is $(4^\circ/15^\circ)\times 60\ \mathrm{min} = 16$ minutes less, or 23:23.

Example 2.11 Find the sidereal time in Helsinki on April 15, 1982 at 20:00 UT.

The Julian date is $J = 2{,}445{,}074.5$ and

$$T = \frac{2{,}445{,}074.5 - 2{,}451{,}545.0}{36{,}525} = -0.1771526\,.$$

Next, we use (2.47) to find the sidereal time at 0 UT:

$$\Theta_0 = -1{,}506{,}521.0\,\mathrm{s} = -418\,\mathrm{h}\ 28\,\mathrm{min}\ 41\,\mathrm{s} = 13\,\mathrm{h}\ 31\,\mathrm{min}\ 19\,\mathrm{s}\,.$$

Since the sidereal time runs 3 min 57 s fast a day as compared to the solar time, the difference in 20 hours will be

$$\frac{20}{24} \times 3\,\mathrm{min}\ 57\,\mathrm{s} = 3\,\mathrm{min}\ 17\,\mathrm{s}\,,$$

and the sidereal time at 20 UT will be 13 h 31 min 19 s + 20 h 3 min 17 s = 33 h 34 min 36 s = 9 h 34 min 36 s.

At the same time (at 22:00 Finnish time, 23:00 daylight saving time) in Helsinki the sidereal time is ahead of this by the amount corresponding to the longitude of Helsinki, 25°, i. e. 1 h 40 min 00 s. Thus the sidereal time is 11 h 14 min 36 s.

2.17 Exercises

Exercise 2.1 Find the distance between Helsinki and Seattle along the shortest route. Where is the northernmost point of the route, and what is its distance from the North Pole? The longitude of Helsinki is 25°E and latitude 60°; the longitude of Seattle is 122°W and latitude 48°. Assume that the radius of the Earth is 6370 km.

Exercise 2.2 A star crosses the south meridian at an altitude of 85°, and the north meridian at 45°. Find the declination of the star and the latitude of the observer.

Exercise 2.3 Where are the following statements true?

a) Castor (α Gem, declination $\delta = 31°53'$) is circumpolar.
b) Betelgeuze (α Ori, $\delta = 7°\ 24'$) culminates at zenith.
c) α Cen ($\delta = -60°\ 50'$) rises to an altitude of 30°.

Exercise 2.4 In his *Old Man and the Sea* Hemingway wrote:

> *It was dark now as it becomes dark quickly after the Sun sets in September. He lay against the worn wood of the bow and rested all that he could. The first stars were out. He did not know the name of Rigel but he saw it and knew soon they would all be out and he would have all his distant friends.*

How was Hemingway's astronomy?

Exercise 2.5 The right ascension of the Sun on June 1, 1983, was 4 h 35 min and declination 22° 00′. Find the ecliptic longitude and latitude of the Sun and the Earth.

Exercise 2.6 Show that on the Arctic Circle the Sun

a) rises at the same sidereal time Θ_0 between December 22 and June 22,
b) sets at the same sidereal time Θ_0 between June 22 and December 22.

What is Θ_0?

Exercise 2.7 Derive the equations (2.24), which give the galactic coordinates as functions of the ecliptic coordinates.

Exercise 2.8 The coordinates of Sirius for the epoch 1900.0 were $\alpha = 6\,\mathrm{h}\ 40\,\mathrm{min}\ 45\,\mathrm{s}$, $\delta = -16°\ 35'$, and the components of its proper motion were $\mu_\alpha = -0.037\,\mathrm{s/a}$, $\mu_\delta = -1.12''\mathrm{a}^{-1}$. Find the coordinates of Sirius for 2000.0. The precession must also be taken into account.

Exercise 2.9 The parallax of Sirius is 0.375″ and radial velocity −8 km/s.

a) What are the tangential and total velocities of Sirius? (See also the previous exercise.)
b) When will Sirius be closest to the Sun?
c) What will its proper motion and parallax be then?

3. Observations and Instruments

Up to the end of the Middle Ages, the most important means of observation in astronomy was the human eye. It was aided by various mechanical devices to measure the positions of celestial bodies in the sky. The telescope was invented in Holland at the beginning of the 17th century, and in 1609 Galileo Galilei made his first astronomical observations with this new instrument. Astronomical photography was introduced at the end of the 19th century, and during the last few decades many kinds of electronic detectors have been adopted for the study of electromagnetic radiation from space. The electromagnetic spectrum from the shortest gamma rays to long radio waves can now be used for astronomical observations.

3.1 Observing Through the Atmosphere

With satellites and spacecraft, observations can be made outside the atmosphere. Yet, the great majority of astronomical observations are carried out from the surface of the Earth. In the preceding chapter, we discussed refraction, which changes the apparent altitudes of objects. The atmosphere affects observations in many other ways as well. The air is never quite steady, and there are layers with different temperatures and densities; this causes convection and turbulence. When the light from a star passes through the unsteady air, rapid changes in refraction in different directions result. Thus, the amount of light reaching a detector, e. g. the human eye, constantly varies; the star is said to *scintillate* (Fig. 3.1). Planets shine more steadily, since they are not point sources like the stars.

A telescope collects light over a larger area, which evens out rapid changes and diminishes scintillation. Instead, differences in refraction along different paths of light through the atmosphere smear the image and point sources are seen in telescopes as vibrating speckles. This phenomenon is called *seeing*, and the size of the seeing disc may vary from less than an arc second to several tens of arc seconds. If the size of the seeing disc is small, we speak of good seeing. Seeing and scintillation both tend to blot out small details when one looks through a telescope, for example, at a planet.

Some wavelength regions in the electromagnetic spectrum are strongly absorbed by the atmosphere. The most important transparent interval is the *optical window* from about 300 to 800 nm. This interval coincides with the region of sensitivity of the human eye (about 400–700 nm).

At wavelengths under 300 nm absorption by atmospheric ozone prevents radiation from reaching the ground. The ozone is concentrated in a thin layer at a height of about 20–30 km, and this layer protects the Earth from harmful ultraviolet radiation. At still shorter wavelengths, the main absorbers are O_2, N_2 and free atoms. Nearly all of the radiation under 300 nm is absorbed by the upper parts of the atmosphere.

At wavelengths longer than visible light, in the near-infrared region, the atmosphere is fairly transparent up

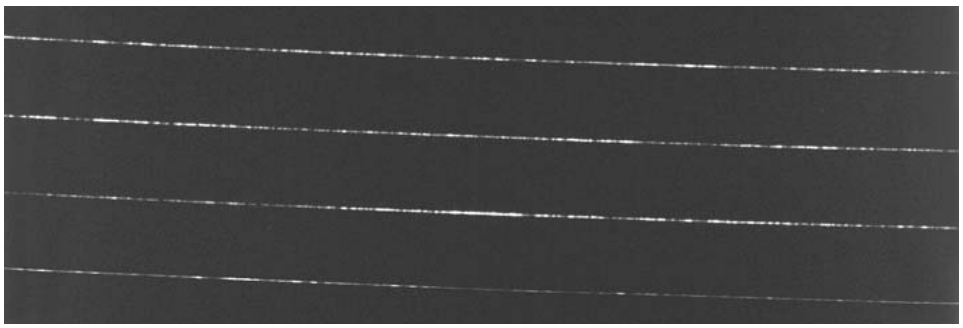


Fig. 3.1. Scintillation of Sirius during four passes across the field of view. The star was very low on the horizon. (Photo by Pekka Parviainen)

Hannu Karttunen et al. (Eds.), *Observations and Instruments*.
In: Hannu Karttunen et al. (Eds.), Fundamental Astronomy, 5th Edition. pp. 47–82 (2007)
DOI: 11685739_3 © Springer-Verlag Berlin Heidelberg 2007

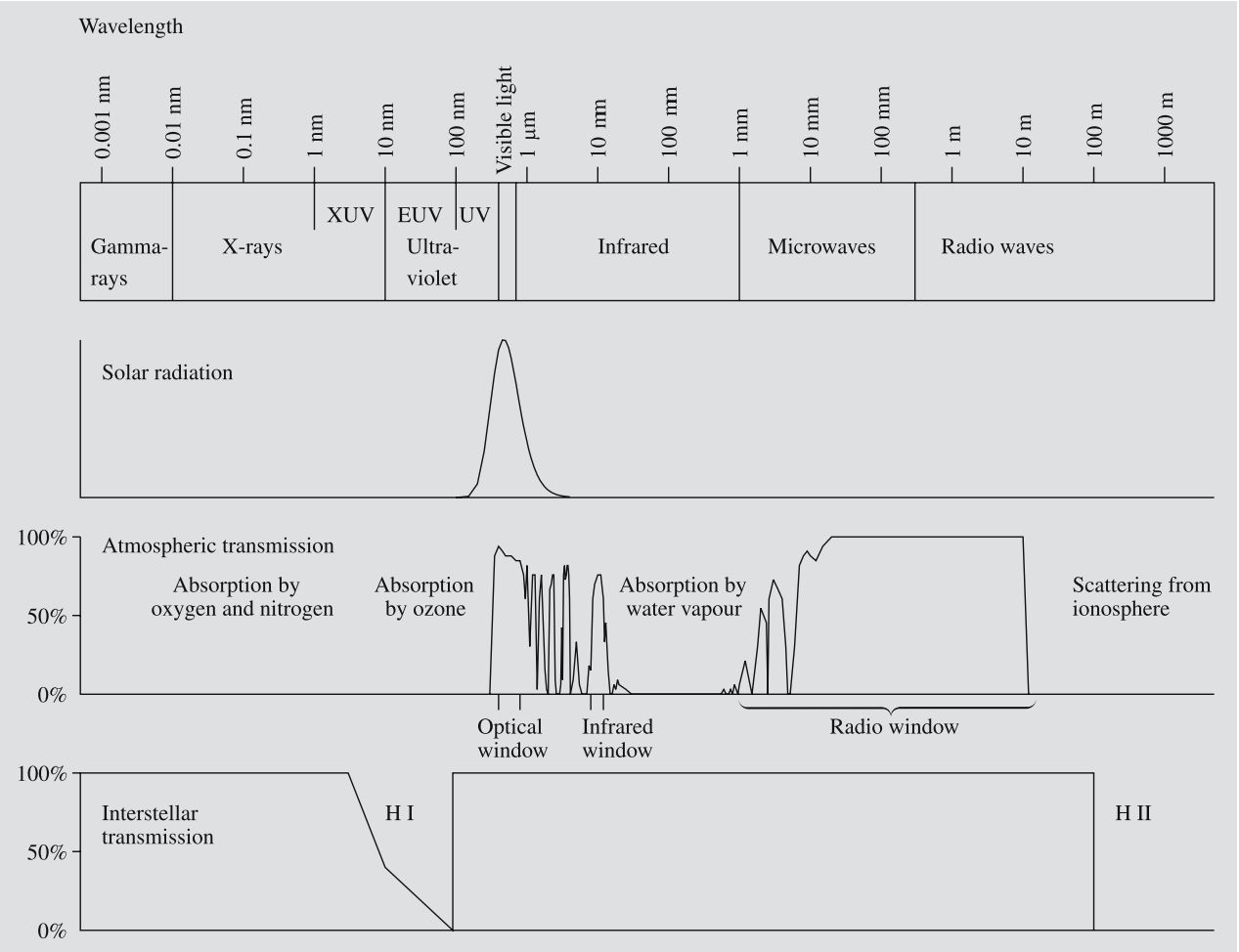


Fig. 3.2. The transparency of the atmosphere at different wavelengths. 100% transmission means that all radiation reaches the surface of the Earth. The radiation is also absorbed by interstellar gas, as shown in the lowermost very schematic figure. The interstellar absorption also varies very much depending on the direction (Chap. 15)

to 1.3 μm. There are some absorption belts caused by water and molecular oxygen, but the atmosphere gets more opaque only at wavelengths of longer than 1.3 μm. At these wavelengths, radiation reaches the lower parts of the atmosphere only in a few narrow windows. All wavelengths between 20 μm and 1 mm are totally absorbed. At wavelengths longer than 1 mm, there is the *radio window* extending up to about 20 m. At still longer wavelengths, the ionosphere in the upper parts of the atmosphere reflects all radiation (Fig. 3.2). The exact upper limit of the radio window depends on the strength of the ionosphere, which varies during the day. (The structure of the atmosphere is described in Chap. 7.)

At optical wavelengths (300–800 nm), light is scattered by the molecules and dust in the atmosphere, and the radiation is attenuated. Scattering and absorption together are called *extinction*. Extinction must be taken into account when one measures the brightness of celestial bodies (Chap. 4).

In the 19th century Lord Rayleigh succeeded in explaining why the sky is blue. Scattering caused by the molecules in the atmosphere is inversely proportional to the fourth power of the wavelength. Thus, blue light is scattered more than red light. The blue light we see all over the sky is scattered sunlight. The same phenomenon colours the setting sun red, because owing to

the long, oblique path through the atmosphere, all the blue light has been scattered away.

In astronomy one often has to observe very faint objects. Thus, it is important that the background sky be as dark as possible, and the atmosphere as transparent as possible. That is why the large observatories have been built on mountain tops far from the cities. The air above an observatory site must be very dry, the number of cloudy nights few, and the seeing good.

Astronomers have looked all over the Earth for optimal conditions and have found some exceptional sites. In the 1970's, several new major observatories were founded at these sites. Among the best sites in the world are: the extinguished volcano Mauna Kea on Hawaii, rising more than 4000 m above the sea; the dry mountains in northern Chile; the Sonoran desert in the U.S., near the border of Mexico; and the mountains on La Palma, in the Canary Islands. Many older observatories are severely plagued by the lights of nearby cities (Fig. 3.3).

In radio astronomy atmospheric conditions are not very critical except when observing at the shortest wavelengths. Constructors of radio telescopes have much greater freedom in choosing their sites than optical astronomers. Still, radio telescopes are also often constructed in uninhabited places to isolate them from disturbing radio and television broadcasts.

Fig. 3.3. Night views from the top of Mount Wilson. The upper photo was taken in 1908, the lower one in 1988. The lights of Los Angeles, Pasadena, Hollywood and more than 40 other towns are reflected in the sky, causing considerable disturbance to astronomical observations. (Photos by Ferdinand Ellerman and International Dark-Sky Association)

3.2 Optical Telescopes

The telescope fulfills three major tasks in astronomical observations:

1. It collects light from a large area, making it possible to study very faint sources.
2. It increases the apparent angular diameter of the object and thus improves resolution.
3. It is used to measure the positions of objects.

The light-collecting surface in a telescope is either a lens or a mirror. Thus, optical telescopes are divided into two types, lens telescopes or *refractors* and mirror telescopes or *reflectors* (Fig. 3.4).

Geometrical Optics. Refractors have two lenses, the *objective* which collects the incoming light and forms an image in the focal plane, and the *eyepiece* which is a small magnifying glass for looking at the image (Fig. 3.5). The lenses are at the opposite ends of a tube which can be directed towards any desired point. The distance between the eyepiece and the focal plane can be adjusted to get the image into focus. The image formed

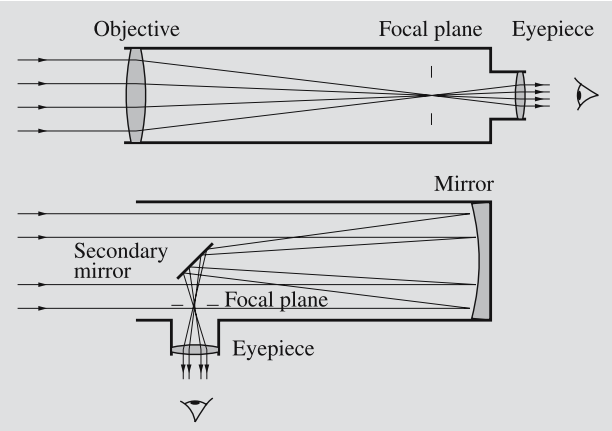


Fig. 3.4. A lens telescope or refractor and a mirror telescope or reflector

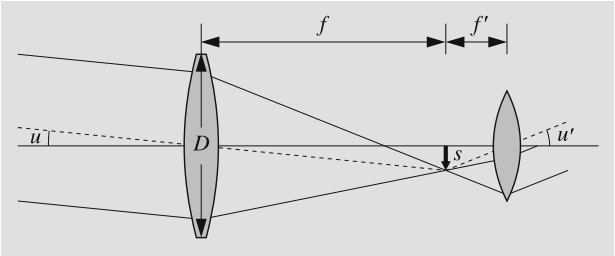


Fig. 3.5. The scale and magnification of a refractor. The object subtends an angle u. The objective forms an image of the object in the focal plane. When the image is viewed through the eyepiece, it is seen at an angle u'

by the objective lens can also be registered, e. g. on a photographic film, as in an ordinary camera.

The diameter of the objective, D, is called the *aperture* of the telescope. The ratio of the aperture D to the focal length f, $F = D/f$, is called the *aperture ratio*. This quantity is used to characterize the light-gathering power of the telescope. If the aperture ratio is large, near unity, one has a powerful, "fast" telescope; this means that one can take photographs using short exposures, since the image is bright. A small aperture ratio (the focal length much greater than the aperture) means a "slow" telescope.

In astronomy, as in photography, the aperture ratio is often denoted by f/n (e. g. $f/8$), where n is the focal length divided by the aperture. For fast telescopes this ratio can be $f/1 \ldots f/3$, but usually it is smaller, $f/8 \ldots f/15$.

The *scale* of the image formed in the focal plane of a refractor can be geometrically determined from Fig. 3.5. When the object is seen at the angle u, it forms an image of height s,

$$s = f \tan u \approx fu \,, \tag{3.1}$$

since u is a very small angle. If the telescope has a focal length of, for instance, 343 cm, one arc minute corresponds to

$$\begin{aligned} s &= 343\,\mathrm{cm} \times 1' \\ &= 343\,\mathrm{cm} \times (1/60) \times (\pi/180) \\ &= 1\,\mathrm{mm}\,. \end{aligned}$$

The *magnification* ω is (from Fig. 3.5)

$$\omega = u'/u \approx f/f' \,, \tag{3.2}$$

where we have used the equation $s = fu$. Here, f is the focal length of the objective and f' that of the eyepiece. For example, if $f = 100$ cm and we use an eyepiece with $f' = 2$ cm, the magnification is 50-fold. The magnification is not an essential feature of a telescope, since it can be changed simply by changing the eyepiece.

A more important characteristic, which depends on the aperture of the telescope, is the *resolving power*, which determines, for example, the minimum angular separation of the components of a binary star that can be seen as two separate stars. The theoretical limit for the resolution is set by the diffraction of light: The telescope does not form a point image of a star, but rather a small disc, since light "bends around the corner" like all radiation (Fig. 3.6).

The theoretical resolution of a telescope is often given in the form introduced by Rayleigh (see *Diffraction by a Circular Aperture, p. 81)

$$\sin\theta \approx \theta = 1.22\,\lambda/D \,, \quad [\theta] = \mathrm{rad}\,. \tag{3.3}$$

As a practical rule, we can say that two objects are seen as separate if the angular distance between them is

$$\theta \gtrsim \lambda/D \,, \quad [\theta] = \mathrm{rad}\,. \tag{3.4}$$

This formula can be applied to optical as well as radio telescopes. For example, if one makes observations at a typical yellow wavelength ($\lambda = 550$ nm), the resolving power of a reflector with an aperture of 1 m is about $0.2''$. However, seeing spreads out the image to a diameter of typically one arc second. Thus, the theoretical diffraction limit cannot usually be reached on the surface of the Earth.

In photography the image is further spread in the photographic plate, decreasing the resolution as compared with visual observations. The grain size of photographic emulsions is about 0.01–0.03 mm, which is also the minimum size of the image. For a focal length of 1 m, the scale is 1 mm = $206''$, and thus 0.01 mm corresponds to about 2 arc seconds. This is similar to the theoretical resolution of a telescope with an aperture of 7 cm in visual observations.

In practice, the resolution of visual observations is determined by the ability of the eye to see details.

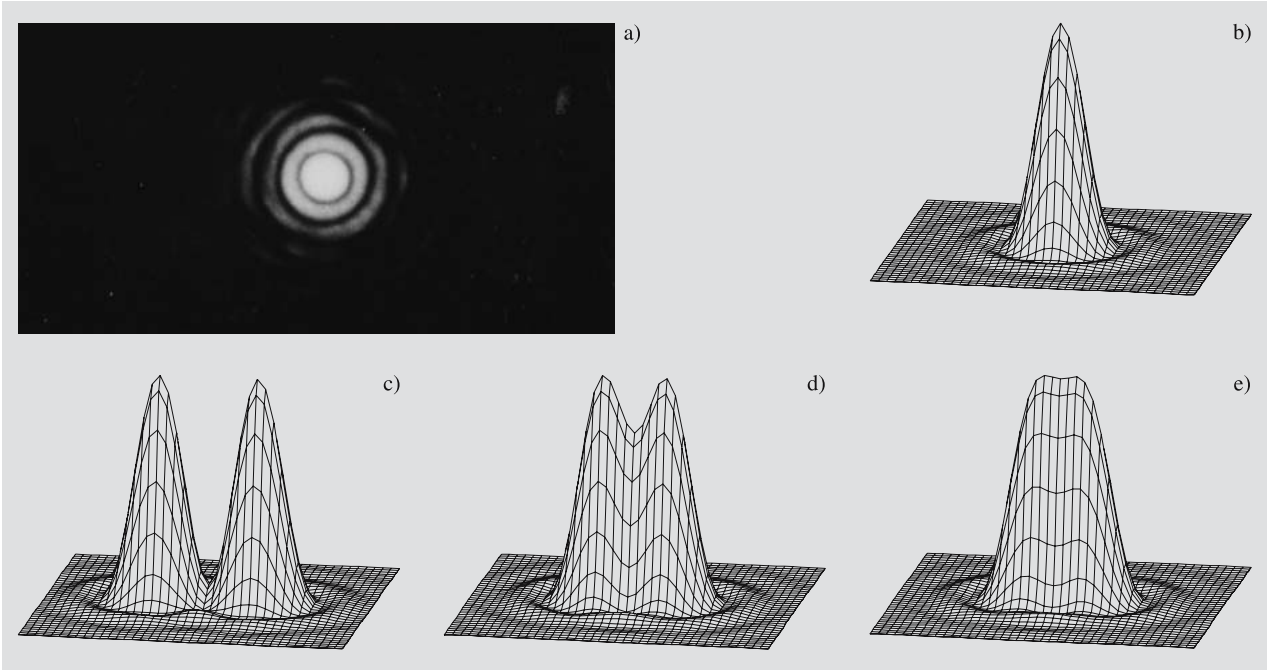


Fig. 3.6a–e. Diffraction and resolving power. The image of a single star (**a**) consists of concentric diffraction rings, which can be displayed as a mountain diagram (**b**). Wide pairs of stars can be easily resolved (**c**). For resolving close binaries, different criteria can be used. One is the Rayleigh limit $1.22\,\lambda/D$ (**d**). In practice, the resolution can be written λ/D, which is near the Dawes limit (**e**). (Photo (a) Sky and Telescope)

In night vision (when the eye is perfectly adapted to darkness) the resolving capability of the human eye is about $2'$.

The *maximum magnification* $\omega_{\max}$ is the largest magnification that is worth using in telescopic observations. Its value is obtained from the ratio of the resolving capability of the eye, $e \approx 2' = 5.8 \times 10^{-4}$ rad, to the resolving power of the telescope, θ,

$$\omega_{\max} = e/\theta \approx eD/\lambda = \frac{5.8 \times 10^{-4}\, D}{5.5 \times 10^{-7}\ \mathrm{m}} \approx D/1\ \mathrm{mm}\ . \tag{3.5}$$

If we use, for example, an objective with a diameter of 100 mm, the maximum magnification is about 100. The eye has no use for larger magnifications.

The *minimum magnification* $\omega_{\min}$ is the smallest magnification that is useful in visual observations. Its value is obtained from the condition that the diameter of the *exit pupil* L of the telescope must be smaller than or equal to the pupil of the eye.

The exit pupil is the image of the objective lens, formed by the eyepiece, through which the light from the objective goes behind the eyepiece. From Fig. 3.7 we obtain

$$L = \frac{f'}{f} D = \frac{D}{\omega}\ . \tag{3.6}$$

Thus the condition $L \le d$ means that

$$\omega \ge D/d\ . \tag{3.7}$$

In the night, the diameter of the pupil of the human eye is about 6 mm, and thus the minimum magnification of a 100 mm telescope is about 17.

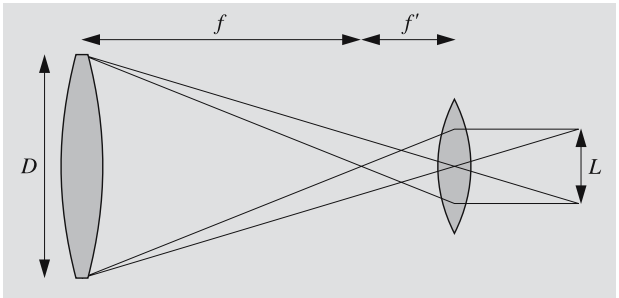


Fig. 3.7. The exit pupil L is the image of the objective lens formed by the eyepiece

Refractors. In the first refractors, which had a simple objective lens, the observations were hampered by the *chromatic aberration*. Since glass refracts different colours by different amounts, all colours do not meet at the same focal point (Fig. 3.8), but the focal length increases with increasing wavelength. To remove this aberration, *achromatic lenses* consisting of two parts were developed in the 18th century. The colour dependence of the focal length is much smaller than in single lenses, and at some wavelength, λ_0, the focal length has an extremum (usually a minimum). Near this point the change of focal length with wavelength is very small (Fig. 3.9). If the telescope is intended for visual observations, we choose $\lambda_0 = 550$ nm, corresponding to the maximum sensitivity of the eye. Objectives for photographic refractors are usually constructed with $\lambda_0 \approx 425$ nm, since normal photographic plates are most sensitive to the blue part of the spectrum.

By combining three or even more lenses of different glasses in the objective, the chromatic aberration can be corrected still better (as in apochromatic objectives). Also, special glasses have been developed where the wavelength dependences of the refractive index cancel out so well that two lenses already give a very good correction of the chromatic aberration. They have, however, hardly been used in astronomy so far.

The largest refractors in the world have an aperture of about one metre (102 cm in the Yerkes Observatory telescope (Fig. 3.10), finished in 1897, and 91 cm in the Lick Observatory telescope (1888)). The aperture ratio is typically $f/10 \ldots f/20$.

The use of refractors is limited by their small field of view and awkwardly long structure. Refractors are used, e. g. for visual observations of binary stars and in various meridian telescopes for measuring the positions of stars. In photography they can be used for accurate position measurements, for example, to find parallaxes.

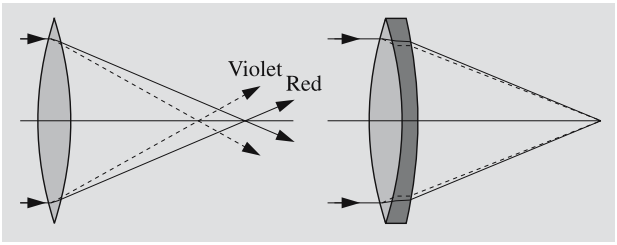


Fig. 3.8. Chromatic aberration. Light rays of different colours are refracted to different focal points (*left*). The aberration can be corrected with an achromatic lens consisting of two parts (*right*)

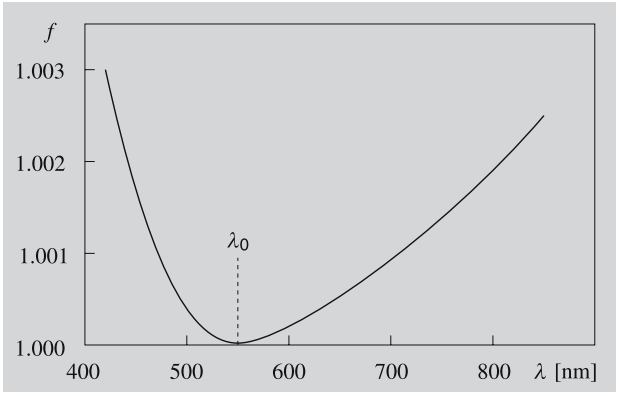


Fig. 3.9. The wavelength dependence of the focal length of a typical achromatic objective for visual observations. The focal length has a minimum near $\lambda = 550$ nm, where the eye is most sensitive. In bluer light ($\lambda = 450$ nm) or in redder light ($\lambda = 800$ nm), the focal length increases by a factor of about 1.002

A wider field of view is obtained by using more complex lens systems, and telescopes of this kind are called *astrographs*. Astrographs have an objective made up of typically 3–5 lenses and an aperture of less than 60 cm. The aperture ratio is $f/5 \ldots f/7$ and the field of view about 5°. Astrographs are used to photograph large areas of the sky, e. g. for proper motion studies and for statistical brightness studies of the stars.

Reflectors. The most common telescope type in astrophysical research is the mirror telescope or reflector. As a light-collecting surface, it employs a mirror coated with a thin layer of aluminium. The form of the mirror is usually parabolic. A parabolic mirror reflects all light rays entering the telescope parallel to the main axis into the same focal point. The image formed at this point can be observed through an eyepiece or registered with a detector. One of the advantages of reflectors is the absence of chromatic aberration, since all wavelengths are reflected to the same point.

In the very largest telescopes, the observer can sit with his instruments in a special cage at the *primary*

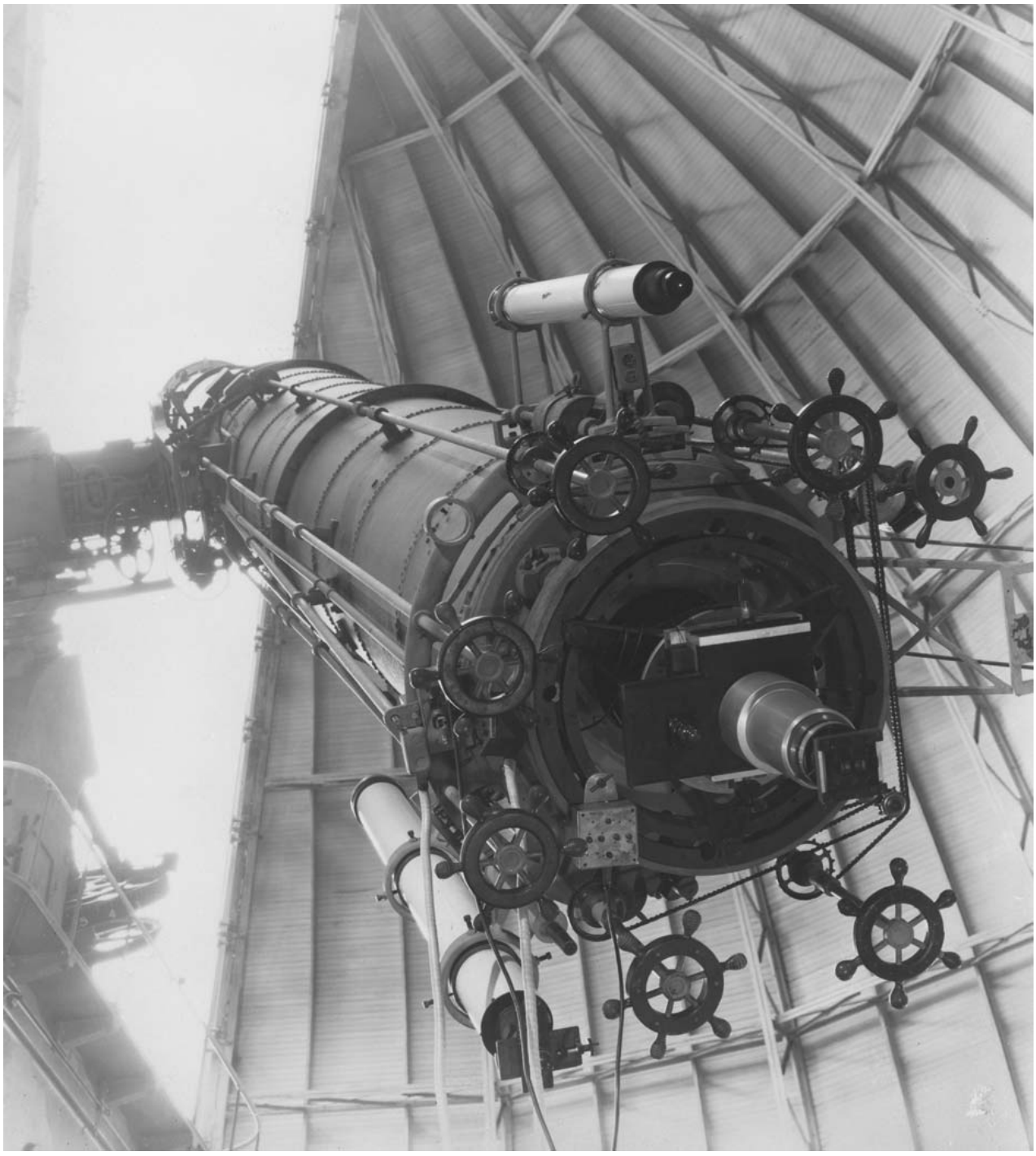


Fig. 3.10. The largest refractor in the world is at the Yerkes Observatory, University of Chicago. It has an objective lens with a diameter of 102 cm. (Photo by Yerkes Observatory)

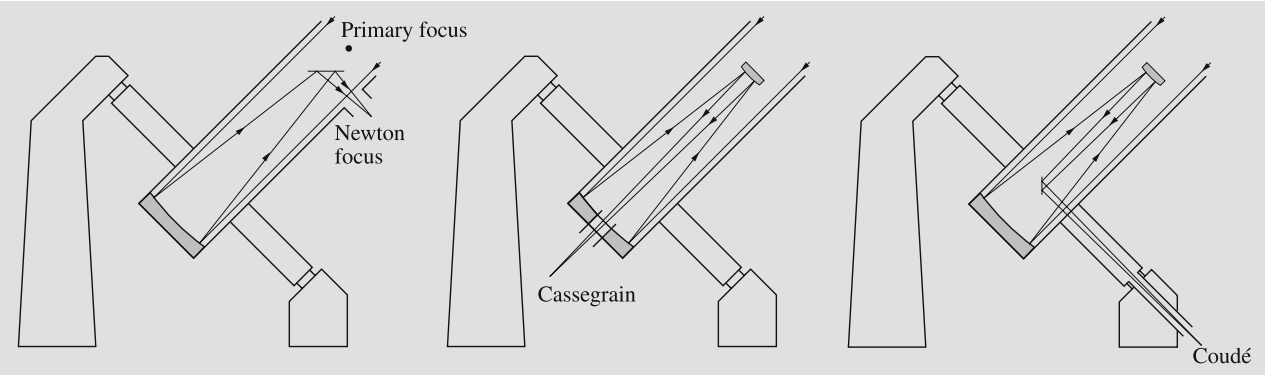


Fig. 3.11. Different locations of the focus in reflectors: primary focus, Newton focus, Cassegrain focus and coudé focus. The coudé system in this figure cannot be used for observations near the celestial pole. More complex coudé systems usually have three flat mirrors after the primary and secondary mirrors

focus (Fig. 3.11) without eclipsing too much of the incoming light. In smaller telescopes, this is not possible, and the image must be inspected from outside the telescope. In modern telescopes instruments are remotely controlled, and the observer must stay away from the telescope to reduce thermal turbulence.

In 1663 James Gregory (1638–1675) described a reflector. The first practical reflector, however, was built by Isaac Newton. He guided the light perpendicularly out from the telescope with a small flat mirror. Therefore the focus of the image in such a system is called the *Newton focus*. A typical aperture ratio of a Newtonian telescope is $f/3 \ldots f/10$. Another possibility is to bore a hole at the centre of the primary mirror and reflect the rays through it with a small hyperbolic secondary mirror in the front end of the telescope. In such a design, the rays meet in the *Cassegrain focus*. Cassegrain systems have aperture ratios of $f/8 \ldots f/15$.

The effective focal length (f_e) of a Cassegrain telescope is determined by the position and convexity of the secondary mirror. Using the notations of Fig. 3.12, we get

$$f_e = \frac{b}{a} f_p \,. \tag{3.8}$$

If we choose $a \ll b$, we have $f_e \gg f_p$. In this way one can construct short telescopes with long focal lengths. Cassegrain systems are especially well suited for spectrographic, photometric and other instruments, which can be mounted in the secondary focus, easily accessible to the observers.

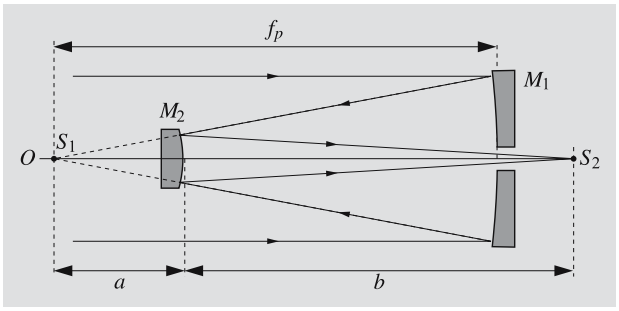


Fig. 3.12. The principle of a Cassegrain reflector. A concave (paraboloid) primary mirror M_1 reflects the light rays parallel to the main axis towards the primary focus S_1. A convex secondary mirror M_2 (hyperboloid) reflects the rays back through a small hole at the centre of the main mirror to the secondary focus S_2 outside the telescope

More complicated arrangements use several mirrors to guide the light through the declination axis of the telescope to a fixed *coudé focus* (from the French word *couder*, to bend), which can even be situated in a separate room near the telescope (Fig. 3.13). The focal length is thus very long and the aperture ratio $f/30 \ldots f/40$. The coudé focus is used mainly for accurate spectroscopy, since the large spectrographs can be stationary and their temperature can be held accurately constant. A drawback is that much light is lost in the reflections in the several mirrors of the coudé system. An aluminized mirror reflects about 80% of the light falling on it, and thus in a coudé system of, e. g. five mirrors (including the primary and secondary mirrors), only $0.8^5 \approx 30\%$ of the light reaches the detector.

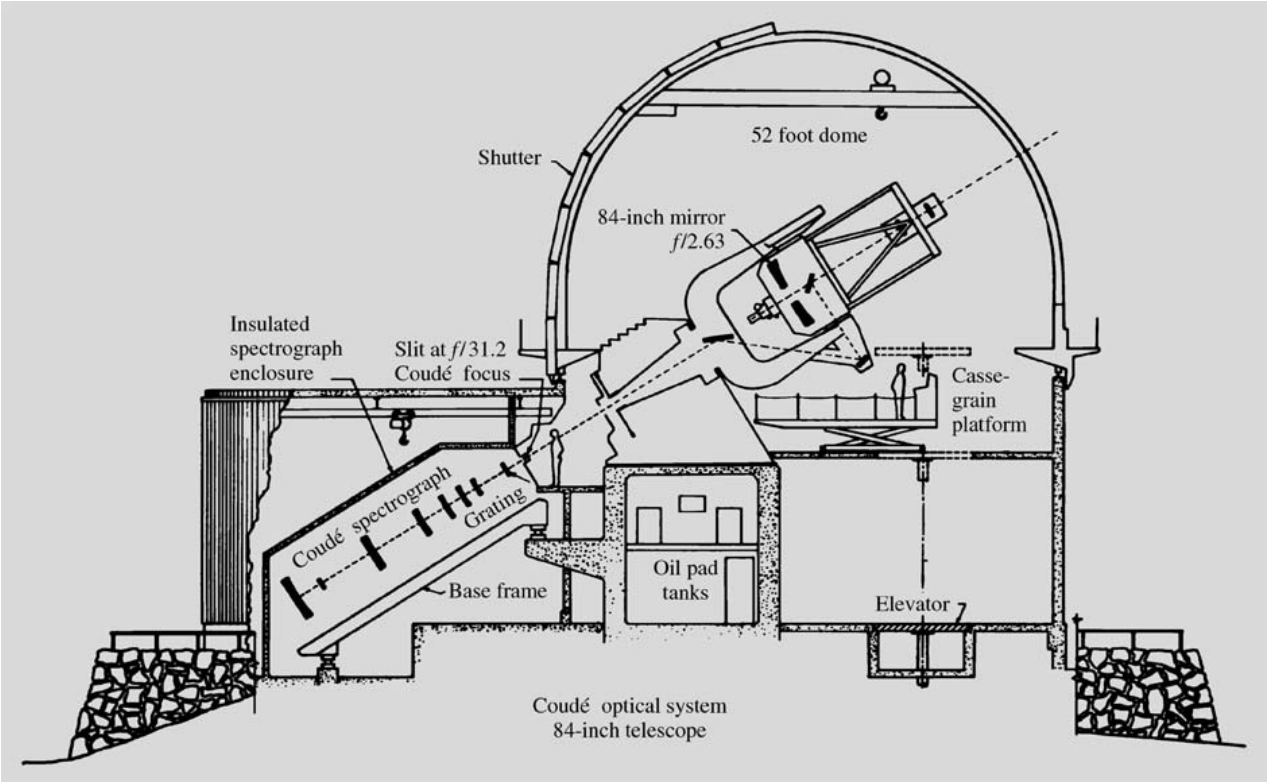


Fig. 3.13. The coudè system of the Kitt Peak 2.1 m reflector. (Drawing National Optical Astronomy Observatories, Kitt Peak National Observatory)

The reflector has its own aberration, *coma*. It affects images displaced from the optical axis. Light rays do not converge at one point, but form a figure like a comet. Due to the coma, the classical reflector with a paraboloid mirror has a very small correct field of view. The coma limits the diameter of the useful field to 2–20 minutes of arc, depending on the aperture ratio of the telescope. The 5 m Palomar telescope, for instance, has a useful field of view of about 4′, corresponding to about one-eighth of the diameter of the Moon. In practice, the small field of view can be enlarged by various correcting lenses.

If the primary mirror were spherical, there would be no coma. However, this kind of mirror has its own error, *spherical aberration*: light rays from the centre and edges converge at different points. To remove the spherical aberration, the Estonian astronomer *Bernhard Schmidt* developed a thin correcting lens that is placed in the way of the incoming light. Schmidt cameras (Figs. 3.14 and 3.15) have a very wide (about 7°), nearly faultless field of view, and the correcting lens is

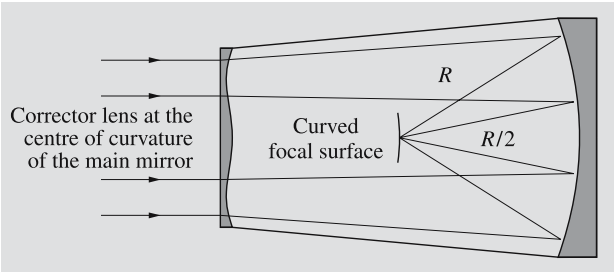


Fig. 3.14. The principle of the Schmidt camera. A correcting glass at the centre of curvature of a concave spherical mirror deviates parallel rays of light and compensates for the spherical aberration of the spherical mirror. (In the figure, the form of the correcting glass and the change of direction of the light rays have been greatly exaggerated.) Since the correcting glass lies at the centre of curvature, the image is practically independent of the incoming angle of the light rays. Thus there is no coma or astigmatism, and the images of stars are points on a spherical surface at a distance of $R/2$, where R is the radius of curvature of the spherical mirror. In photography, the plate must be bent into the form of the focal surface, or the field rectified with a corrector lens

so thin that it absorbs very little light. The images of the stars are very sharp.

In Schmidt telescopes the diaphragm with the correcting lens is positioned at the centre of the radius of curvature of the mirror (this radius equals twice the focal length). To collect all the light from the edges of the field of view, the diameter of the mirror must be larger than that of the correcting glass. The Palomar Schmidt camera, for example, has an aperture of 122 cm (correcting lens)/183 cm (mirror) and a focal length of 300 cm. The largest Schmidt telescope in the world is in Tautenburg, Germany, and its corresponding values are 134/203/400 cm.

A disadvantage of the Schmidt telescope is the curved focal plane, consisting of a part of a sphere. When the telescope is used for photography, the plate must be bent along the curved focal plane. Another possibility of correcting the curvature of the field of view is to use an extra correcting lens near the focal plane. Such a solution was developed by the Finnish astronomer Yrjö Väisälä in the 1930's, independently of Schmidt. Schmidt cameras have proved to be very effective in mapping the sky. They have been used to photograph the Palomar Sky Atlas mentioned in the previous chapter and its continuation, the *ESO/SRC Southern Sky Atlas*.

The Schmidt camera is an example of a *catadioptric telescope*, which has both lenses and mirrors. *Schmidt–Cassegrain telescopes* used by many amateurs are modifications of the Schmidt camera. They have a secondary mirror mounted at the centre of the correcting lens; the mirror reflects the image through a hole in the primary mirror. Thus the effective focal length can be rather long, although the telescope itself is very short. Another common catadioptric telescope is the *Maksutov* telescope. Both surfaces of the correcting lens as well as the primary mirror of a Maksutov telescope are concentric spheres.

Another way of removing the coma of the classical reflectors is to use more complicated mirror surfaces. The *Ritchey–Chrétien* system has hyperboloidal primary and secondary mirrors, providing a fairly wide useful field of view. Ritchey–Chrétien optics are used in many large telescopes.

◄ **Fig. 3.15.** The large Schmidt telescope of the European Southern Observatory. The diameter of the mirror is 1.62 m and of the free aperture 1 m. (Photo ESO)

Mountings of Telescopes. A telescope has to be mounted on a steady support to prevent its shaking, and it must be smoothly rotated during observations. There are two principal types of mounting, *equatorial* and *azimuthal* (Fig. 3.16).

In the equatorial mounting, one of the axes is directed towards the celestial pole. It is called the *polar axis* or *hour axis*. The other one, the *declination axis*, is perpendicular to it. Since the hour axis is parallel to the axis of the Earth, the apparent rotation of the sky can be compensated for by turning the telescope around this axis at a constant rate.

The declination axis is the main technical problem of the equatorial mounting. When the telescope is pointing to the south its weight causes a force perpendicular to the axis. When the telescope is tracking an object and turns westward, the bearings must take an increasing load parallel with the declination axis.

In the azimuthal mounting, one of the axes is vertical, the other one horizontal. This mounting is easier to construct than the equatorial mounting and is more stable for very large telescopes. In order to follow the rotation of the sky, the telescope must be turned around both of the axes with changing velocities. The field of view will also rotate; this rotation must be compensated for when the telescope is used for photography.

If an object goes close to the zenith, its azimuth will change 180° in a very short time. Therefore, around the zenith there is a small region where observations with an azimuthal telescope are not possible.

The largest telescopes in the world were equatorially mounted until the development of computers made possible the more complicated guidance needed for azimuthal mountings. Most of the recently built large telescopes are already azimuthally mounted. Azimuthally mounted telescopes have two additional obvious places for foci, the *Nasmyth foci* at both ends of the horizontal axis.

The *Dobson mounting*, used in many amateur telescopes, is azimuthal. The magnification of the Newtonian telescope is usually small, and the telescope rests on pieces of teflon, which make it very easy to move. Thus the object can easily be tracked manually.

Another type of mounting is the *coelostat*, where rotating mirrors guide the light into a stationary telescope. This system is used especially in solar telescopes.

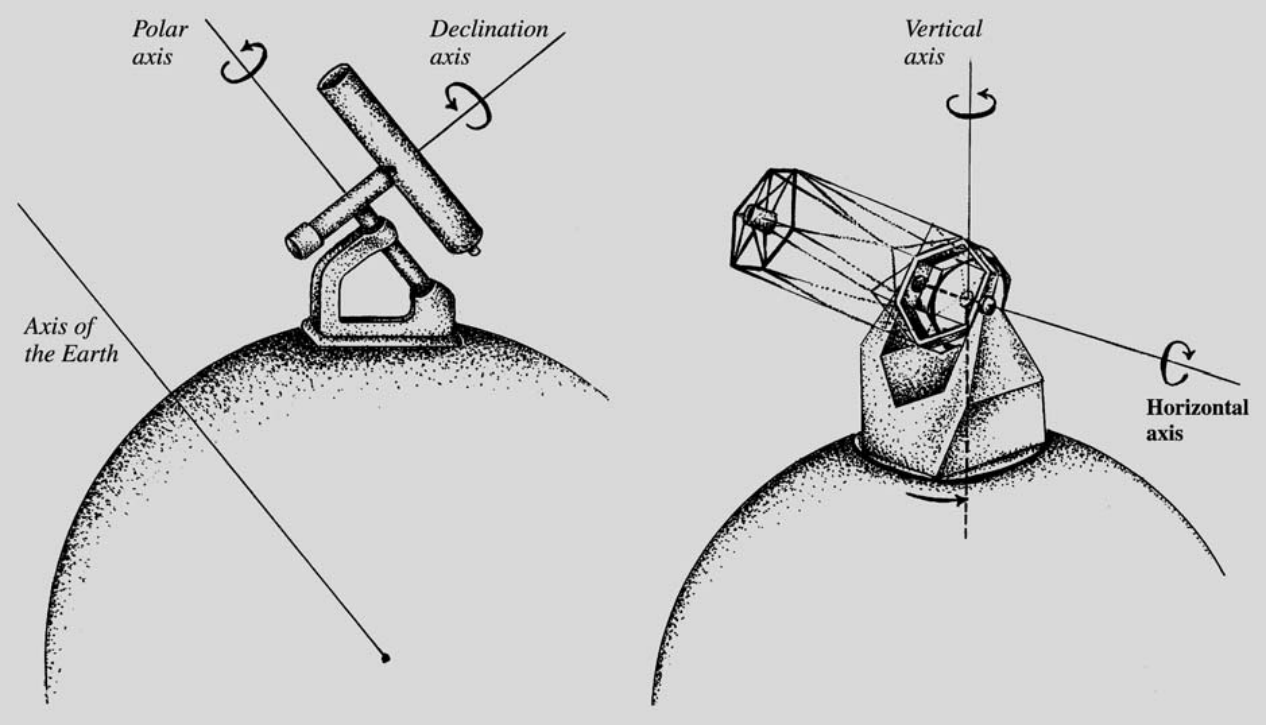


Fig. 3.16. The equatorial mounting (*left*) and the azimuthal mounting (*right*)

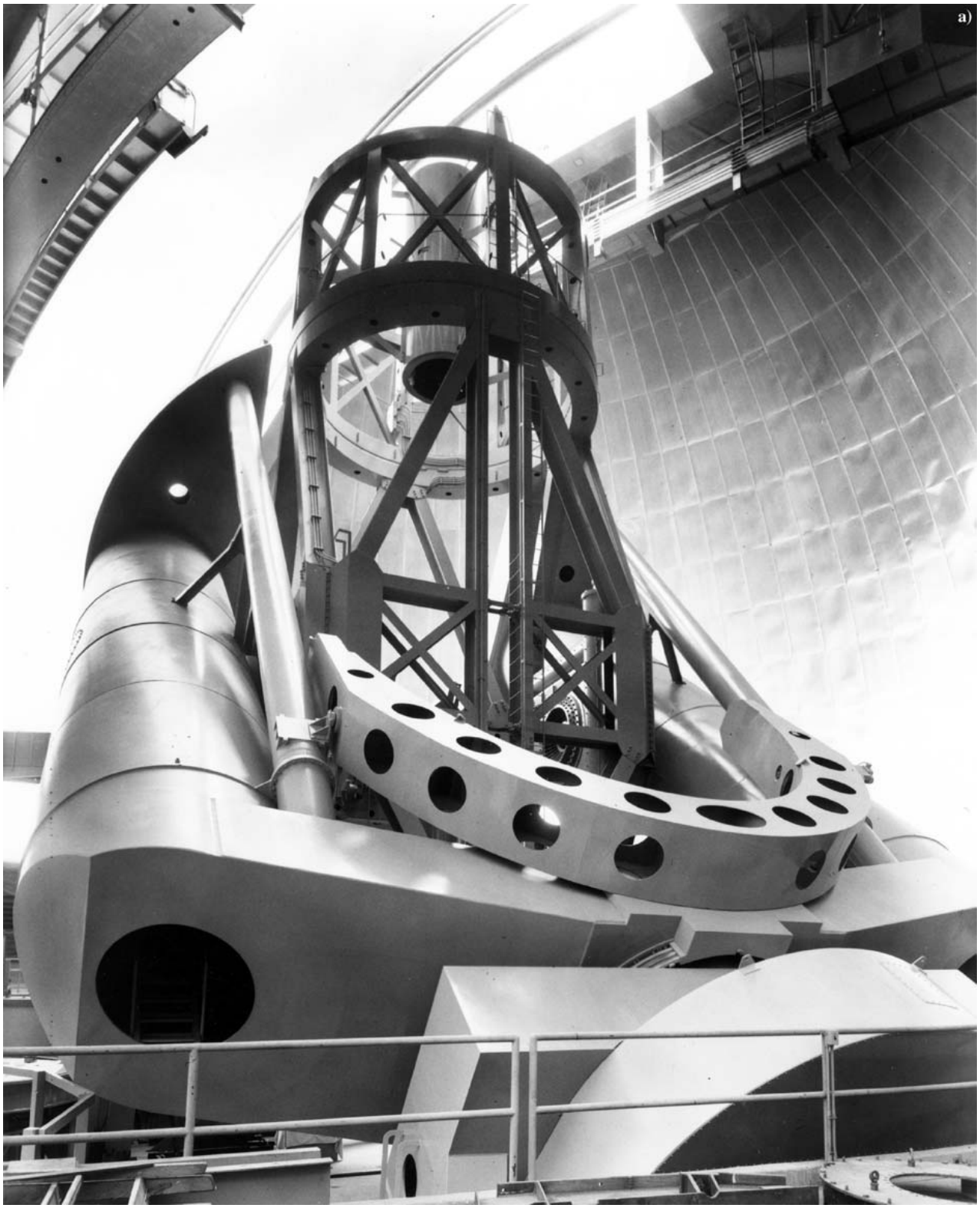
To measure absolute positions of stars and accurate time, telescopes aligned with the north–south direction are used. They can be rotated around one axis only, the east–west horizontal axis. *Meridian circles* or *transit instruments* with this kind of mounting were widely constructed for different observatories during the 19th century. A few are still used for astrometry, but they are now highly automatic like the meridian circle on La Palma funded by the Carlsberg foundation.

New Techniques. Detectors are already approaching the theoretical limit of efficiency, where all incident photons are registered. Ultimately, to detect even fainter objects the only solution is to increase the light gathering area, but also the mirrors are getting close to the practical maximum size. Thus, new technical solutions are needed.

One new feature is *active optics*, used e. g. in the ESO 3.5 metre NTT telescope (New Technology Telescope) at La Silla, Chile. The mirror is very thin, but its shape is kept exactly correct by a computer controlled support mechanism. The weight and production cost of such a mirror are much smaller compared with a conventional thick mirror. Because of the smaller weight also the supporting structure can be made lighter.

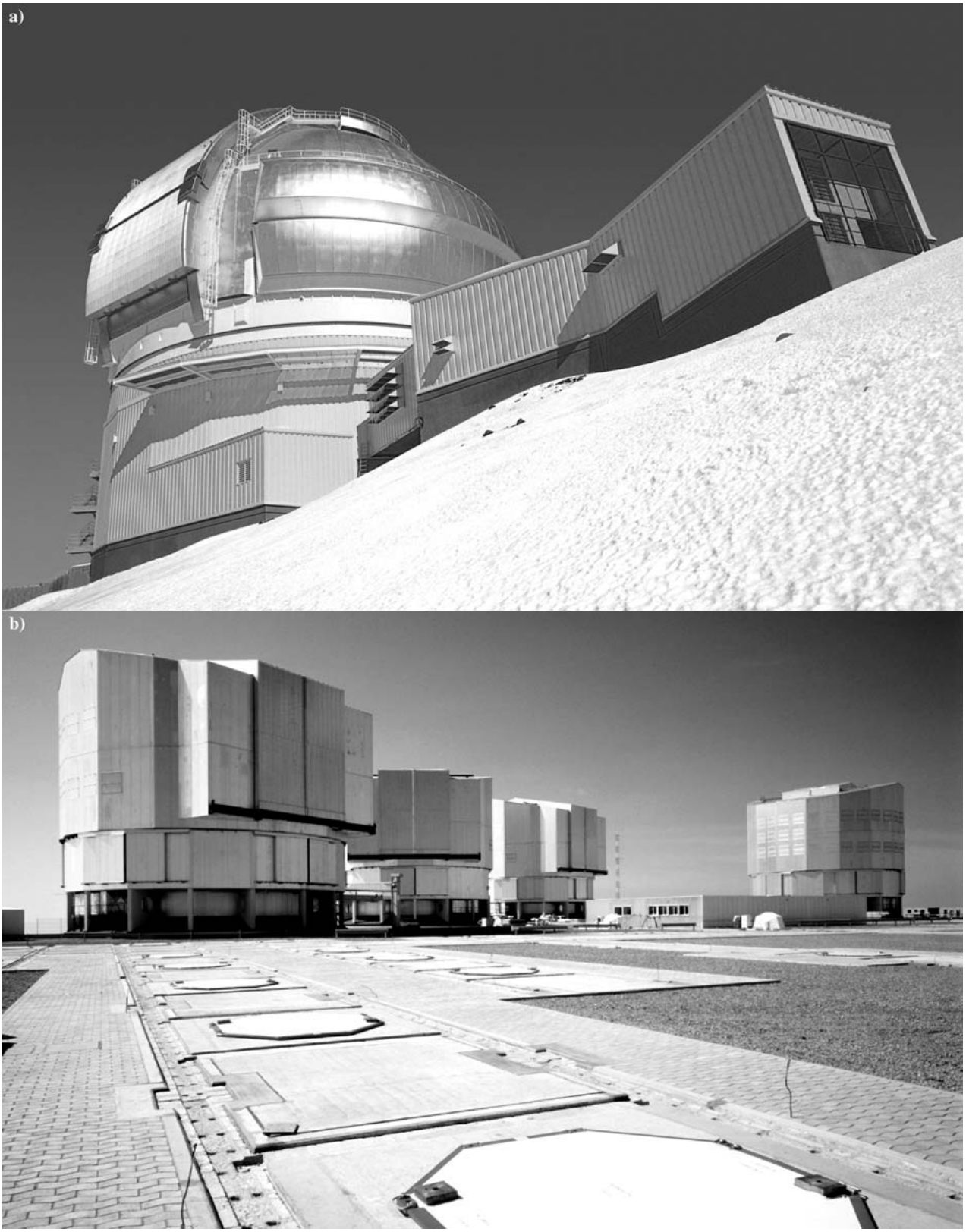
Developing the support mechanism further leads to *adaptive optics*. A reference star (or an artificial beam) is monitored constantly in order to obtain the shape of the seeing disk. The shape of the main mirror or a smaller auxiliary mirror is adjusted up to hundreds of times a second to keep the image as concentrated as possible. Adaptive optics has been taken into use in the largest telescopes of the world from about the year 2000 on.

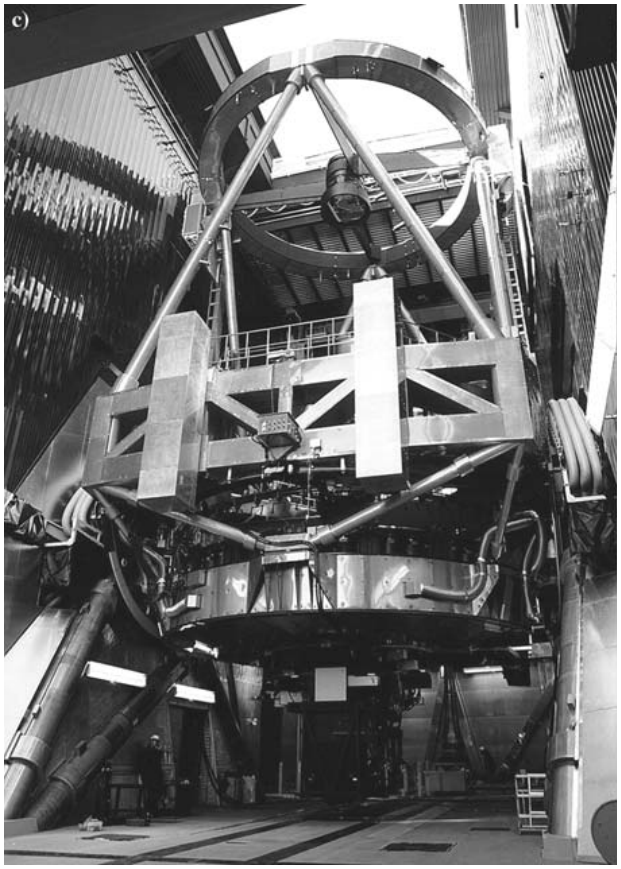
Fig. 3.17a–c. The largest telescopes in the world in 1947–2000. (**a**) For nearly 30 years, the 5.1 m Hale telescope on Mount Palomar, California, USA, was the largest telescope in the world. (**b**) The BTA, Big Azimuthal Telescope, is situated in the Caucasus in the southern Soviet Union. Its mirror has a diameter of 6 m. It was set in operation at the end of 1975. (**c**) The William M. Keck Telescope on the summit of Mauna Kea, Hawaii, was completed in 1992. The 10 m mirror consists of 36 hexagonal segments. (Photos Palomar Observatory, Spetsialnaya Astrofizitsheskaya Observatorya, and Roger Ressmeyer – Starlight for the California Association for Research in Astronomy) ▶


a)


b)
СССР
ЛОМО


c)


a)
b)



◀ **Fig. 3.18a–c.** Some new large telescopes. (**a**) The 8.1 m Gemini North telescope on Mauna Kea, Hawaii, was set in operation in 1999. Its twin, Gemini South, was dedicated in 2000. (**b**) The European Southern Observatory (ESO) was founded by Belgium, France, the Netherlands, Sweden and West Germany in 1962. Other European countries have joined them later. The VLT (Very Large Telescope) on Cerro Paranal in Northern Chile, was inaugurated in 1998–2000. (**c**) The first big Japanese telescope, the 8.3 m Subaru on Mauna Kea, Hawaii, started observations in 1999. (Photos National Optical Astronomy Observatories, European Southern Observatory and Subaru Observatory)

The mirrors of large telescopes need not be monolithic, but can be made of smaller pieces that are, e. g. hexagonal. These *mosaic mirrors* are very light and can be used to build up mirrors with diameters of several tens of metres (Fig. 3.19). Using active optics, the hexagons can be accurately focussed. The California Association for Research in Astronomy has constructed the William M. Keck telescope with a 10 m mosaic mirror. It is located on Mauna Kea, and the last segment was installed in 1992. A second, similar telescope Keck II was completed in 1996, the pair forming a huge binocular telescope.

The reflecting surface does not have to be continuous, but can consist of several separate mirrors. Such

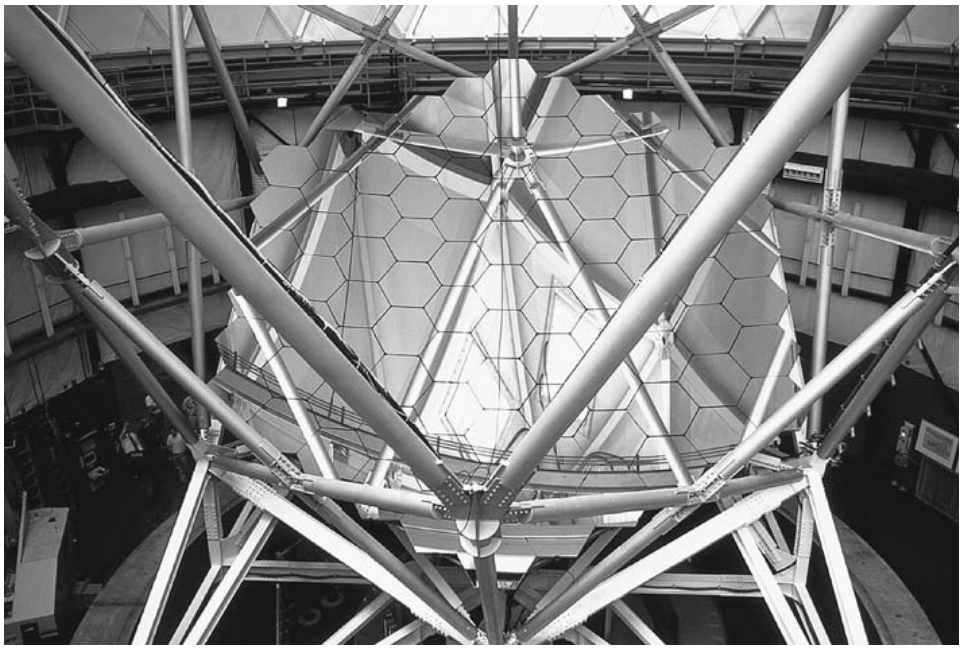


Fig. 3.19. The mirror of a telescope can be made up of several smaller segments, which are much easier to manufacture, as in the Hobby–Eberle Telescope on Mount Fowlkes, Texas. The effective diameter of the mirror is 9.1 m. A similar telescope is being built in South Africa. (Photo MacDonald Observatory)

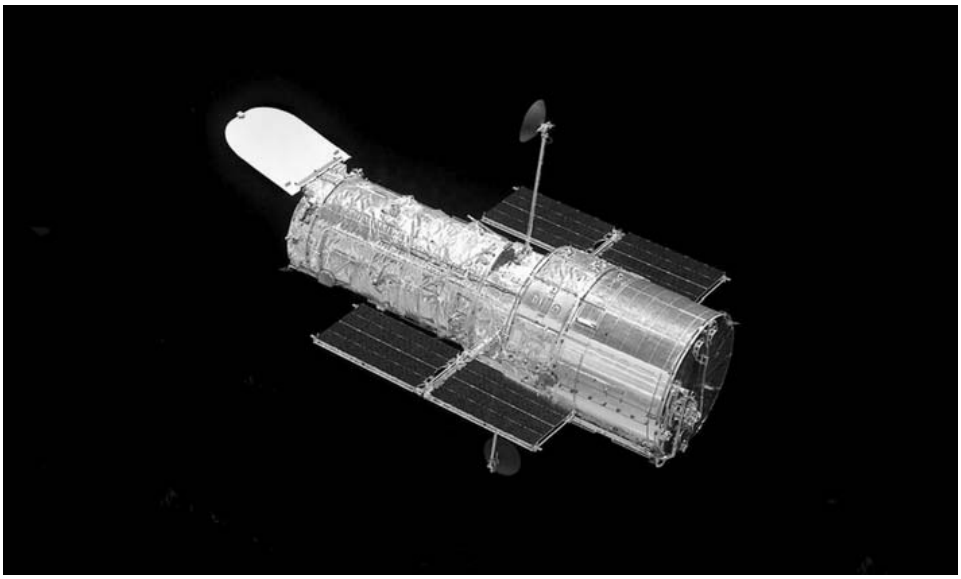


Fig. 3.20. The Hubble Space Telescope after the latest service flight in 2002. The telescope got new solar panels and several other upgrades. (Photo NASA)

a telescope was operating on Mount Hopkins, Arizona, in 1979–1999. It was the *Multiple-Mirror Telescope* (MMT) with six 1.8 m mirrors together corresponding to a single mirror having a diameter of 4.5 m. In 2000 the six mirrors were replaced by one 6.5 m mirror.

The European Southern Observatory has constructed its own multi-mirror telescope. ESO's Very Large Telescope (VLT) has four closely located mirrors (Fig. 3.18). The diameter of each mirror is eight metres, and the total area corresponds to one telescope with a 16 m mirror. The resolution is even better, since the "aperture", i. e. the maximum distance between the mirrors, is several tens of meters.

An important astronomical instruments of the 20th century is the *Hubble Space Telescope*, launched in 1990 (Fig. 3.20). It has a mirror with a diameter of 2.4 m. The resolution of the telescope (after the faulty optics was corrected) is near the theoretical diffraction limit, since there is no disturbing atmosphere. A second generation Space Telescope, now called the *James Webb Space Telescope*, with a mirror of about 6.5 m is planned to be launched in about 2011.

The Hubble Space Telescope was the first large optical telescope in Earth orbit. In the future, satellites will continue to be mainly used for those wavelength regions where the radiation is absorbed by the atmosphere. Due to budgetary reasons, the majority of astronomical observations will still be carried out on the Earth, and great attention will be given to improving ground-based observatories and detectors.

3.3 Detectors and Instruments

Only a limited amount of information can be obtained by looking through a telescope with the unaided eye. Until the end of the 19th century this was the only way to make observations. The invention of photography in the middle of the 19th century brought a revolution in astronomy. The next important step forward in optical astronomy was the development of photoelectric photometry in the 1940's and 1950's. A new revolution, comparable to that caused by the invention of photography, took place in the middle of the 1970's with the introduction of different semiconductor detectors. The sensitivity of detectors has grown so much that today, a 60 cm telescope can be used for observations similar to those made with the Palomar 5 m telescope when it was set in operation in the 1940's.

The Photographic Plate. *Photography* has long been one of the most common methods of observation in astronomy. In astronomical photography glass plates were used, rather than film, since they keep their shape better, but nowadays they are no more manufactured,

and CCD-cameras have largely replaced photography. The sensitive layer on the surface of the film or plate is made up of a silver halide, usually silver bromide, AgBr. A photon absorbed by the halide excites an electron that can move from one atom to another. A silver ion, Ag^+, can catch the electron, becoming a neutral atom. When the necessary amount of silver atoms have been accumulated at one place, they form a latent image. The latent image can be made into a permanent negative by treating the plate after exposure with various chemicals, which transform the silver bromide crystals enclosing the latent image into silver ("development"), and remove the unexposed crystals ("fixing").

The photographic plate has many advantages over the human eye. The plate can register up to millions of stars (picture elements) at one time, while the eye can observe at most one or two objects at a time. The image on a plate is practically permanent – the picture can be studied at any time. In addition, the photographic plate is cheap and easy to use, as compared to many other detectors. The most important feature of a plate is its capability to collect light over an extended time: the longer exposures are used, the more silver atoms are formed on the plate (the plate darkens). By increasing the exposure times, fainter objects can be photographed. The eye has no such capacity: if a faint object does not show through a telescope, it cannot been seen, no matter how long one stares.

One disadvantage of the photographic plate is its low sensitivity. Only one photon in a thousand causes a reaction leading to the formation of a silver grain. Thus the *quantum efficiency* of the plate is only 0.1%. Several chemical treatments can be used to sensitize the plate before exposure. This brings the quantum efficiency up to a few percent. Another disadvantage is the fact that a silver bromide crystal that has been exposed once does not register anything more, i. e. a saturation point is reached. On the other hand, a certain number of photons are needed to produce an image. Doubling the number of photons does not necessarily double the density (the 'blackness' of the image): the density of the plate depends nonlinearly on the amount of incoming light. The sensitivity of the plate is also strongly dependent on the wavelength of the light. For the reasons mentioned above the accuracy with which brightness can be measured on a photographic plate is usually worse than about 5%. Thus the photographic plate makes a poor photometer, but it can be excellently used, e. g. for measuring the positions of stars (positional astronomy) and for mapping the sky.

Photocathodes, Photomultipliers. A *photocathode* is a more effective detector than the photographic plate. It is based on the photoelectric effect. A light quantum, or photon, hits the photocathode and loosens an electron. The electron moves to the positive electrode, or anode, and gives rise to an electric current that can be measured. The quantum efficiency of a photocathode is about 10–20 times better than that of a photographic plate; optimally, an efficiency of 30% can be reached. A photocathode is also a linear detector: if the number of electrons is doubled, the outcoming current is also doubled.

The *photomultiplier* is one of the most important applications of the photocathode. In this device, the electrons leaving the photocathode hit a dynode. For each electron hitting the dynode, several others are released. When there are several dynodes in a row, the original weak current can be intensified a millionfold. The photomultiplier measures all the light entering it, but does not form an image. Photomultipliers are mostly used in photometry, and an accuracy of 0.1–1% can be attained.

Photometers, Polarimeters. A detector measuring brightness, *a photometer*, is usually located behind the telescope in the Cassegrain focus. In the focal plane there is a small hole, the *diaphragm*, which lets through light from the object under observation. In this way, light from other stars in the field of view can be prevented from entering the photometer. A *field lens* behind the diaphragm refracts the light rays onto a photocathode. The outcoming current is intensified further in a preamplifier. The photomultiplier needs a voltage of 1000–1500 volts.

Observations are often made in a certain wavelength interval, instead of measuring all the radiation entering the detector. In this case a *filter* is used to prevent other wavelengths from reaching the photomultiplier. A photometer can also consist of several photomultipliers (Fig. 3.21), which measure simultaneously different wavelength bands. In such an instrument beam splitters or semitransparent mirrors split the light beam through fixed filters to the photomultipliers.

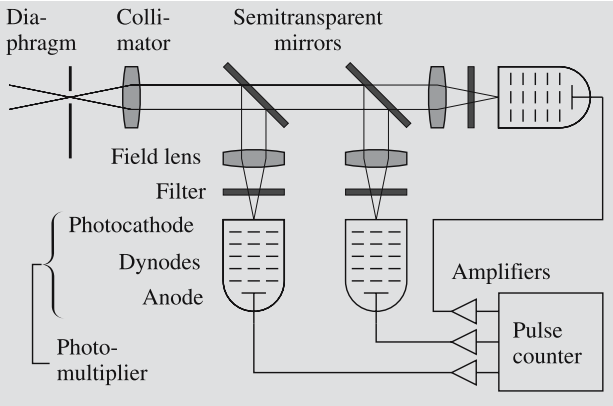


Fig. 3.21. The principle of a photoelectric multicolour photometer. Light collected by the telescope arrives from the left. The light enters the photometer through a small hole in the focal plane, the diaphragm. A lens collimates the light into a parallel beam. Semitransparent mirrors divide the beam to several photomultipliers. A field lens guides the light through a filter onto the photocathode of the photomultiplier. The quanta of light, photons, release electrons from the cathodes. The electrons are accelerated towards the dynodes with a voltage of about 1500 V. The electrons hitting the dynodes release still more electrons, and the current is greatly enhanced. Every electron emitted from the cathode gives rise to a pulse of up to 10^8 electrons at the anode; the pulse is amplified and registered by a pulse counter. In this way, the photons from the star are counted

In a device called the *photopolarimeter*, a *polarizing filter* is used, either alone or in combination with other filters. The degree and direction of polarization can be found by measuring the intensity of the radiation with different orientations of the polarizers.

In practice, the diaphragm of a photometer will always also let through part of the background sky around the observed object. The measured brightness is in reality the combined brightness of the object and the sky. In order to find the brightness of the object, the background brightness must be measured separately and subtracted from the combined brightness. The accuracy of the measurements is decreased if long observation times are used and the background brightness undergoes fast changes. The problem can be solved by observing the brightness of the background sky and the object simultaneously.

Photometric observations are often relative. If one is observing, e. g. a variable star, a *reference star* close to the actual target is observed at regular intervals. Using the observations of this reference star it is possible to derive a model for the slow changes in the atmospheric extinction (see Chap. 4) and remove their effect. The instrument can be calibrated by observing some *standard stars*, whose brightness is known very accurately.

Image Intensifiers. Different *image intensifiers* based on the photocathode have been used since the 1960's. In the intensifier the information about the starting point of the electron on the photocathode is preserved and the intensified image is formed on a fluorescent screen. The image can then be registered, e. g. with a CCD camera. One of the advantages of the image intensifier is that even faint objects can be imaged using relatively short exposures, and observations can be made at wavelengths where the detector is insensitive.

Another common type of detector is based on the TV camera (*Vidicon camera*). The electrons released from the photocathode are accelerated with a voltage of a few kilovolts before they hit the electrode where they form an image in the form of an electric charge distribution. After exposure, the charge at different points of the electrode is read by scanning its surface with an electron beam row by row. This produces a video signal, which can be transformed into a visible image on a TV tube. The information can also be saved in digital form. In the most advanced systems, the scintillations caused by single electrons on the fluorescent screen of the image intensifier can be registered and stored in the memory of a computer. For each point in the image there is a memory location, called a picture element or *pixel*.

Since the middle of the 1970's, detectors using semiconductor techniques began to be used in increasing numbers. With semiconductor detectors a quantum efficiency of about 70–80% can be attained; thus, sensitivity cannot be improved much more. The wavelength regions suitable for these new detectors are much wider than in the case of the photographic plate. The detectors are also linear. Computers are used for collecting, saving and analyzing the output data available in digital form.

CCD Camera. The most important new detector is the *CCD camera* (Charge Coupled Device). The detector consists of a surface made up of light sensitive silicon

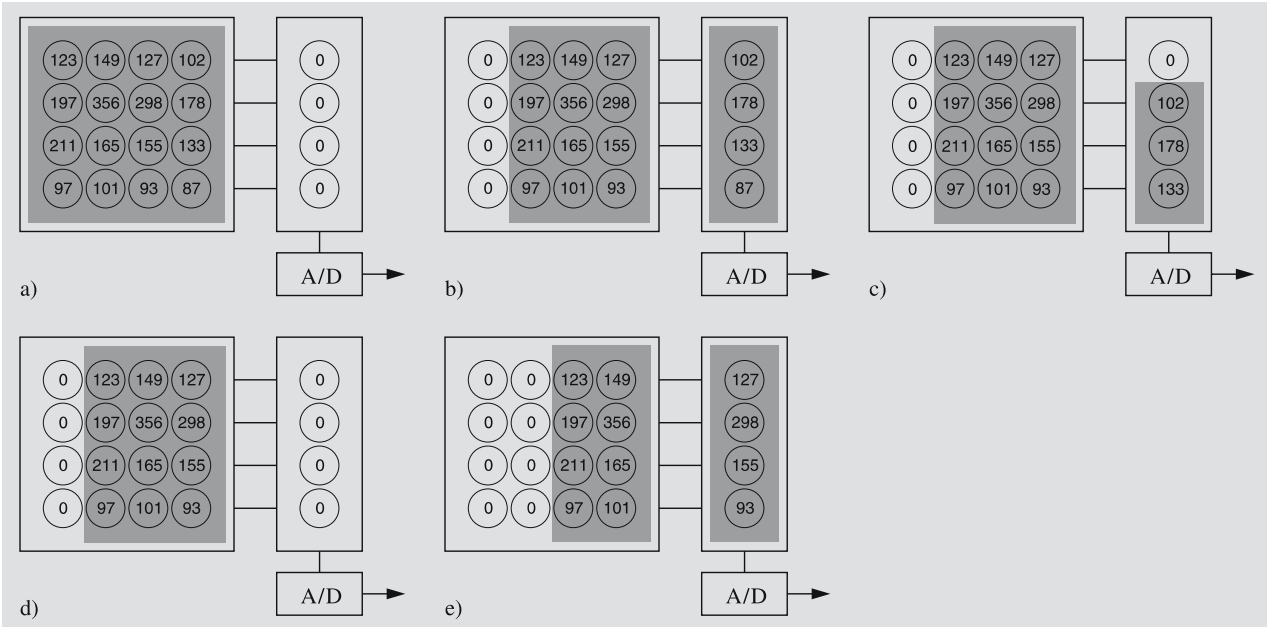


Fig. 3.22a–e. The principle of reading a CCD camera. (**a**) During an exposure electrons are trapped in potential wells corresponding to pixels of the camera. The number at each pixel shows the number of electrons. (**b**) After the exposure each horizontal line is moved one pixel to the right; the rightmost row moves to the readout buffer. (**c**) The contents of the buffer is moved down by one pixel. The lowermost charge moves to the A/D converter, which sends the number of electrons to the computer. (**d**) After moving the buffer down several times one vertical row has been read. (**e**) The image is again shifted right by one pixel. This procedure is repeated till the whole image is read

diodes, arranged in a rectangular array of image elements or pixels. The largest cameras can have as many as 4096×4096 pixels, although most are considerably smaller.

A photon hitting the detector can release an electron, which will remain trapped inside a pixel. After the exposure varying potential differences are used to move the accumulated charges row by row to a readout buffer. In the buffer the charges are moved pixel by pixel to an analogy/digital converter, which transmits the digital value to a computer. Reading an image also clears the detector (Fig. 3.22). If the exposures are very short the readout times may take a substantial part of the observing time.

The CCD camera is nearly linear: the number of electrons is directly proportional to the number of photons. Calibration of the data is much easier than with photographic plates.

The quantum efficiency, i. e. the number of electrons per incident photon, is high, and the CCD camera is much more sensitive than a photographic plate. The sensitivity is highest in the red wavelength range, about 600–800 nm, where the quantum efficiency can be 80–90% or even higher.

The range of the camera extends far to the infrared. In the ultraviolet the sensitivity drops due to the absorption of the silicon very rapidly below about 500 nm. Two methods have been used to avoid this problem. One is to use a coating that absorbs the ultraviolet photons and emits light of longer wavelength. Another possibility is to turn the chip upside down and make it very thin to reduce the absorption.

The thermal noise of the camera generates *dark current* even if the camera is in total darkness. To reduce the noise the camera must be cooled. Astronomical CCD cameras are usually cooled with liquid nitrogen, which efficiently removes most of the dark current. However, the sensitivity is also reduced when the camera is cooled; so too cold is not good either. The temperature must be kept constant in order to obtain consistent

data. For amateurs there are already moderately priced CCD cameras, which are electrically cooled. Many of them are good enough also for scientific work, if very high sensitivity is not required.

The dark current can easily be measured by taking exposures with the shutter closed. Subtracting this from the observed image gives the real number of electrons due to incident light.

The sensitivity of individual pixels may be slightly different. This can be corrected for by taking an image of an evenly illuminated field, like a twilight sky. This image is called a *flat-field*. When observations are divided by the flat-field, the error caused by different pixels is removed.

The CCD camera is very stable. Therefore it is not necessary to repeat the dark current and flat-field observations very frequently. Typically these calibration exposures are taken during evening and morning twilights, just before and after actual observations.

Cosmic rays are charged particles that can produce extraneous bright dots in CCD images. They are usually limited to one or two pixels, and are easily identified. Typically a short exposure of a few minutes contains a few traces of cosmic rays. Instead of a single long exposure it is usually better to take several short ones, clean the images from cosmic rays, and finally add the images on a computer.

A more serious problem is the *readout noise* of the electronics. In the first cameras it could be hundreds of electrons per pixel. In modern cameras it is a few electrons. This gives a limit to the faintest detectable signal: if the signal is weaker than the readout noise, it is indistinguishable from the noise.

Although the CCD camera is a very sensitive detector, even bright light cannot damage it. A photomultiplier, on the other hand, can be easily destroyed by letting in too much light. However, one pixel can only store a certain number of electrons, after which it becomes *saturated*. Excessive saturation can make the charge to overflow also to the neighboring pixels. If the camera becomes badly saturated it may have to be read several times to completely remove the charges.

The largest CCD cameras are quite expensive, and even they are still rather small compared with photographic plates and films. Therefore photography still has some use in recording extended objects.

Spectrographs. The simplest spectrograph is a prism that is placed in front of a telescope. This kind of device is called the *objective prism spectrograph*. The prism spreads out the different wavelengths of light into a spectrum which can be registered. During the exposure, the telescope is usually slightly moved perpendicularly to the spectrum, in order to increase the width of the spectrum. With an objective prism spectrograph, large numbers of spectra can be photographed, e. g. for spectral classification.

For more accurate information the *slit spectrograph* must be used (Fig. 3.23). It has a narrow slit in the focal plane of the telescope. The light is guided through the slit to a collimator that reflects or refracts all the light rays into a parallel beam. After this, the light is dispersed into a spectrum by a prism and focused with a camera onto a detector, which nowadays is usually a CCD camera. A comparison spectrum is exposed next to the stellar spectrum to determine the precise wavelengths. In modern spectrographs using CCD cameras, the comparison spectrum is usually exposed as a separate image. A big slit spectrograph is often placed at the coudé or Nasmyth focus of the telescope.

Instead of the prism a *diffraction grating* can be used to form the spectrum. A grating has narrow grooves, side by side, typically several hundred per millimetre. When light is reflected by the walls of the grooves, the adjoining rays interfere with each other and give rise to spectra of different orders. There are two kinds of gratings: *reflection* and *transmission gratings*. In a reflection grating no light is absorbed by the glass as in the prism or transmission grating. A grating usually has higher dispersion, or ability to spread the spectrum,

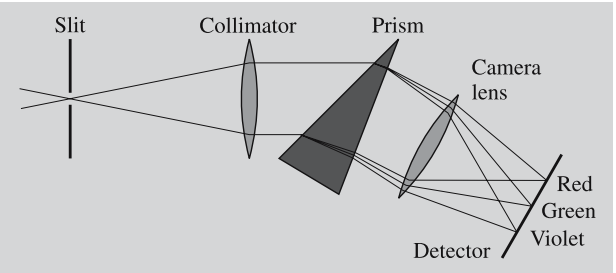


Fig. 3.23. The principle of the slit spectrograph. Light rays entering through a slit are collimated (made parallel to each other), dispersed into a spectrum by a prism and projected onto a photographic plate or a CCD

than a prism. The dispersion can be increased by increasing the density of the grooves of the grating. In slit spectrographs the reflection grating is most commonly used.

Interferometers. The resolution of a big telescope is in practice limited by seeing, and thus increasing the aperture does not necessarily improve the resolution. To get nearer to the theoretical resolution limit set by diffraction (Fig. 3.6), different *interferometers* can be used.

There are two types of optical interferometers. One kind uses an existing large telescope; the other a system of two or more separate telescopes. In both cases the light rays are allowed to interfere. By analyzing the outcoming interference pattern, the structures of close binaries can be studied, apparent angular diameters of the stars can be measured, etc.

One of the earliest interferometers was the *Michelson interferometer* that was built shortly before 1920 for the largest telescope of that time. In front of the telescope, at the ends of a six metre long beam, there were flat mirrors reflecting the light into the telescope. The form of the interference pattern changed when the separation of the mirrors was varied. In practice, the interference pattern was disturbed by seeing, and only a few positive results were obtained with this instrument.

The diameters of over 30 of the brightest stars have been measured using *intensity interferometers*. Such a device consists of two separate telescopes that can be moved in relation to each other. This method is suitable for the brightest objects only.

In 1970 the Frenchman *Antoine Labeyrie* introduced the principle of *speckle interferometry*. In traditional imaging the pictures from long exposures consist of a large number of instantaneous images, "speckles", that together form the seeing disc. In speckle interferometry very short exposures and large magnifications are used and hundreds of pictures are taken. When these pictures are combined and analyzed (usually in digital form), the actual resolution of the telescope can nearly be reached.

The accuracy of interferometric techniques was improved at the beginning of 00's. The first experiments to use the two 10 m Keck telescopes as one interferometer, were made in 2001. Similarly, the ESO VLT will be used as an interferometer.

3.4 Radio Telescopes

Radio astronomy represents a relatively new branch of astronomy. It covers a frequency range from a few megahertz (100 m) up to frequencies of about 300 GHz (1 mm), thereby extending the observable electromagnetic spectrum by many orders of magnitude. The low-frequency limit of the radio band is determined by the opacity of the ionosphere, while the high-frequency limit is due to the strong absorption from oxygen and water bands in the lower atmosphere. Neither of these limits is very strict, and under favourable conditions radio astronomers can work into the submillimetre region or through ionospheric holes during sunspot minima.

At the beginning of the 20th century attempts were made to observe radio emission from the Sun. These experiments, however, failed because of the low sensitivity of the antenna–receiver systems, and because of the opaqueness of the ionosphere at the low frequencies at which most of the experiments were carried out. The first observations of cosmic radio emission were later made by the American engineer Karl G. Jansky in 1932, while studying thunderstorm radio disturbances at a frequency of 20.5 MHz (14.6 m). He discovered radio emission of unknown origin, which varied within a 24 hour period. Somewhat later he identified the source of this radiation to be in the direction of the centre of our Galaxy.

The real birth of radio astronomy may perhaps be dated to the late 1930's, when Grote Reber started systematic observations with his homemade 9.5 m paraboloid antenna. Thereafter radio astronomy developed quite rapidly and has greatly improved our knowledge of the Universe.

Observations are made both in the continuum (broad band) and in spectral lines (radio spectroscopy). Much of our knowledge about the structure of our Milky Way comes from radio observations of the 21 cm line of neutral hydrogen and, more recently, from the 2.6 mm line of the carbon monoxide molecule. Radio astronomy has resulted in many important discoveries; e. g. both pulsars and quasars were first found by radio astronomical observations. The importance of the field can also be seen from the fact that the Nobel prize in physics has recently been awarded twice to radio astronomers.

A radio telescope collects radiation in an aperture or antenna, from which it is transformed to an electric

signal by a receiver, called a radiometer. This signal is then amplified, detected and integrated, and the output is registered on some recording device, nowadays usually by a computer. Because the received signal is very weak, one has to use sensitive receivers. These are often cooled to minimize the noise, which could otherwise mask the signal from the source. Because radio waves are electromagnetic radiation, they are reflected and refracted like ordinary light waves. In radio astronomy, however, mostly reflecting telescopes are used.

At low frequencies the antennas are usually dipoles (similar to those used for radio or TV), but in order to increase the collecting area and improve the resolution, one uses dipole arrays, where all dipole elements are connected to each other.

The most common antenna type, however, is a parabolic reflector, which works exactly as an optical mirror telescope. At long wavelengths the reflecting surface does not need to be solid, because the long wavelength photons cannot see the holes in the reflector, and the antenna is therefore usually made in the form of a metal mesh. At high frequencies the surface has to be smooth, and in the millimetre-submillimetre range, radio astronomers even use large optical telescopes, which they equip with their own radiometers. To ensure a coherent amplification of the signal, the surface irregularities should be less than one-tenth of the wavelength used.

The main difference between a radio telescope and an optical telescope is in the recording of the signal. Radio telescopes are not imaging telescopes (except for synthesis telescopes, which will be described later); instead, a feed horn, which is located at the antenna focus, transfers the signal to a receiver. The wavelength and phase information is, however, preserved.

The resolving power of a radio telescope, θ, can be deduced from the same formula (3.4) as for optical telescopes, i.e. λ/D, where λ is the wavelength used and D is the diameter of the aperture. Since the wavelength ratio between radio and visible light is of the order of 10,000, radio antennas with diameters of several kilometres are needed in order to achieve the

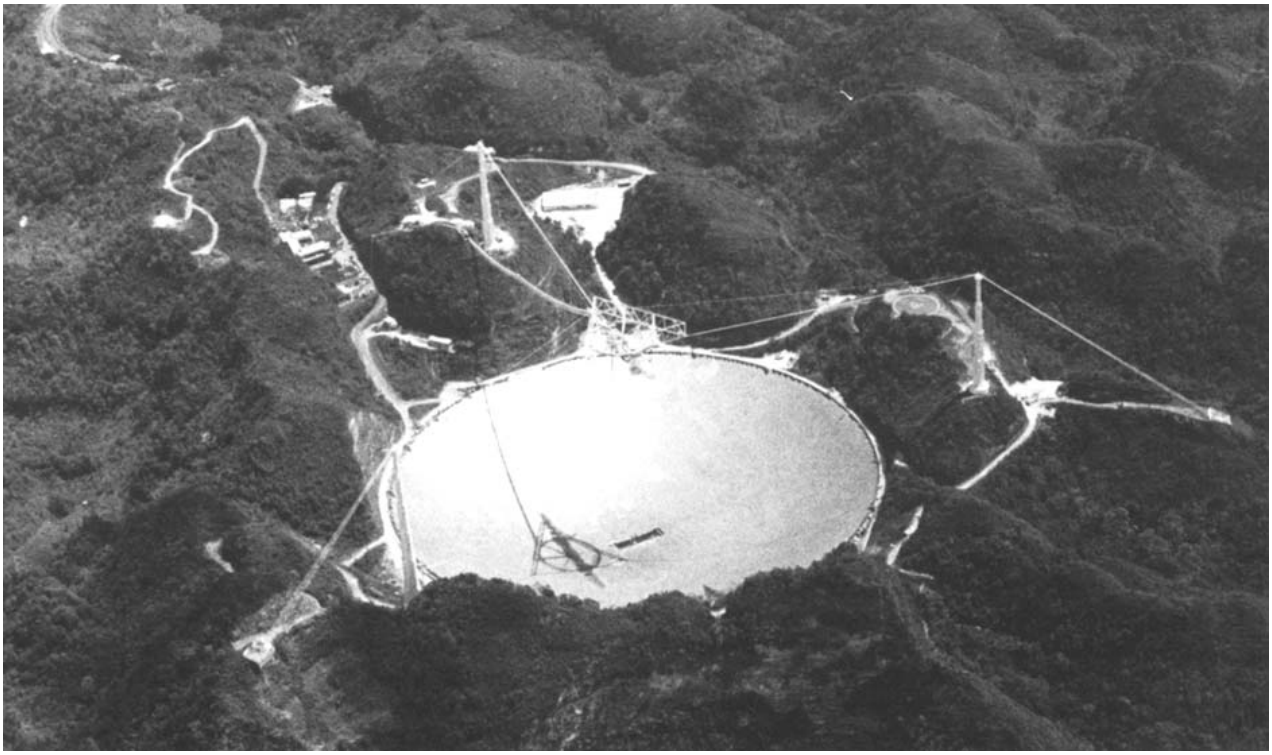


Fig. 3.24. The largest radio telescope in the world is the Arecibo dish in Puerto Rico. It has been constructed over a natural bowl and is 300 m in diameter. (Photo Arecibo Observatory)

Fig. 3.25. The largest fully steerable radio telescope is in Green Bank, Virginia. Its diameter is 100×110 m. (Photo NRAO)

same resolution as for optical telescopes. In the early days of radio astronomy poor resolution was the biggest drawback for the development and recognition of radio astronomy. For example, the antenna used by Jansky had a fan beam with a resolution of about 30° in the narrower direction. Therefore radio observations could not be compared with optical observations. Neither was it possible to identify the radio sources with optical counterparts.

The world's biggest radio telescope is the Arecibo antenna in Puerto Rico, whose main reflector is fixed and built into a 305 m diameter, natural round valley covered by a metal mesh (Fig. 3.24). In the late 1970's the antenna surface and receivers were upgraded, enabling the antenna to be used down to wavelengths of 5 cm. The mirror of the Arecibo telescope is not parabolic but spherical, and the antenna is equipped with a movable feed system, which makes observations possible within a 20° radius around the zenith.

The biggest completely steerable radio telescope is the Green Bank telescope in Virginia, U.S.A., dedicated at the end of 2000. It is slightly asymmetric with a diameter of 100×110 m (Fig. 3.25). Before the Green Bank telescope, for over two decades the largest telescope was the Effelsberg telescope in Germany. This antenna has a parabolic main reflector with a diameter of 100 m. The inner 80 m of the dish is made of solid aluminium panels, while the outmost portion of the disk is a metal mesh structure. By using only the inner portion of the telescope, it has been possible to observe down to wavelengths of 4 mm. The oldest and perhaps best-known big radio telescope is the 76 m antenna at Jodrell Bank in Britain, which was completed in the end of the 1950's.

The biggest telescopes are usually incapable of operating below wavelengths of 1 cm, because the surface cannot be made accurate enough. However, the millimetre range has become more and more important. In this wavelength range there are many transitions of interstellar molecules, and one can achieve quite high angular resolution even with a single dish telescope. At present, the typical size of a mirror of a millimetre telescope is about 15 m. The development of this field is rapid, and at present several big millimetre telescopes are in opera-

Fig. 3.26. The 15 metre Maxwell submillimetre telescope on Mauna Kea, Hawaii, is located in a dry climate at an altitude of 4100 m. Observations can be made down to wavelengths of 0.5 mm. (Photo Royal Observatory, Edinburgh)

tion (Table C.24). Among them are the 40 m Nobeyama telescope in Japan, which can be used down to 3 mm, the 30 m IRAM telescope at Pico Veleta in Spain, which is usable down to 1 mm, and the 15 m UK James Clerk Maxwell Telescope on Mauna Kea, Hawaii, operating down to 0.5 mm (Fig. 3.26). The largest project in the first decade of the 21st century is ALMA (Atacama Large Millimetre Array), which comprises of 50 telescopes with a diameter of 12 m (Fig. 3.27). It will be built as an international project by the United States, Europe and Japan.

As already mentioned, the resolving power of a radio telescope is far poorer than that of an optical telescope. The biggest radio telescopes can at present reach a resolution of 5 arc seconds, and that only at the very highest frequencies. To improve the resolution by increasing the size is difficult, because the present telescopes are already close to the practical upper limit. However, by combining radio telescopes and interferometers, it is possible to achieve even better resolution than with optical telescopes.

As early as 1891 Michelson used an interferometer for astronomical purposes. While the use of interferometers has proved to be quite difficult in the optical wavelength regime, interferometers are extremely useful in the radio region. To form an interferometer, one needs at least two antennas coupled together. The spacing between the antennas, D, is called the baseline. Let us first assume that the baseline is perpendicular to the line of sight (Fig. 3.28). Then the radiation arrives at

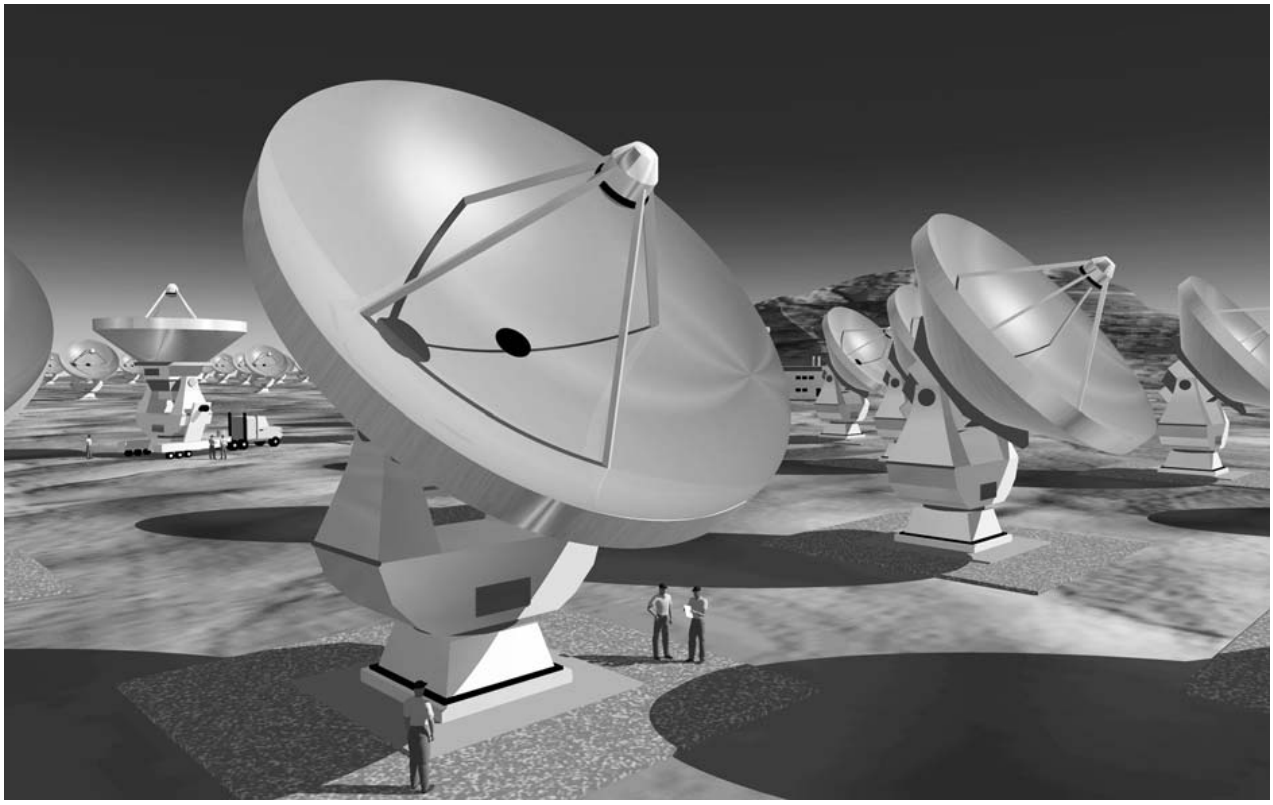


Fig. 3.27. The Atacama Large Millimetre Array (ALMA) will be built in cooperation by Europe, U.S.A. and Japan. The original plan was to have 64 antennas, but for financial reasons the number has been reduced to 50. (Drawing ESO/NOAJ)

both antennas with the same phase, and the summed signal shows a maximum. However, due to the rotation of the Earth, the direction of the baseline changes, producing a phase difference between the two signals. The result is a sinusoidal interference pattern, in which minima occur when the phase difference is 180 degrees. The distance between the peaks is given by

$$\theta D = \lambda \,,$$

where θ is the angle the baseline has turned and λ is the wavelength of the received signal. The resolution of the interferometer is thus equal to that of an antenna with a linear size equal to D.

If the source is not a point source, the radiation emitted from different parts of the source will have phase differences when it enters the antennas. In this case the minima of the interference pattern will not be zero, but will have some positive value $P_{\min}$. If we denote the maximum value of the interference pattern by $P_{\max}$, the ratio

$$\frac{P_{\max} - P_{\min}}{P_{\max} + P_{\min}}$$

gives a measure of the source size (fringe visibility).

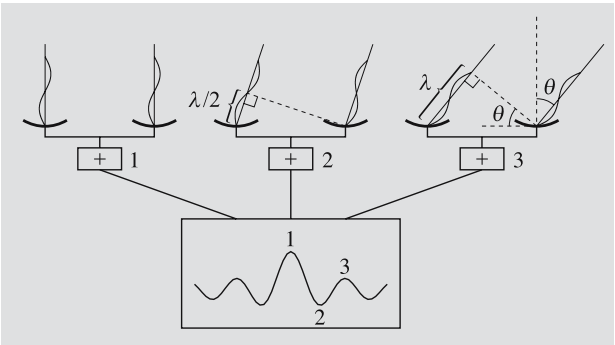


Fig. 3.28. The principle of an interferometer. If the radiation reaches the radio telescopes in the same phase, the waves amplify each other and a maximum is obtained in the combined radiation (cases *1* and *3*). If the incoming waves are in opposite phase, they cancel each other (case *2*)

More accurate information about the source structure can be obtained by changing the spacing between the antennas, i. e. by moving the antennas with respect to each other. If this is done, interferometry is transformed into a technique called *aperture synthesis*.

The theory and techniques of aperture synthesis were developed by the British astronomer Sir Martin Ryle. In Fig. 3.29 the principle of aperture synthesis is illustrated. If the telescopes are located on an east–west track, the spacing between them, projected onto the sky, will describe a circle or an ellipse, depending on the position of the source as the the Earth rotates around its axis. If one varies the distance between the telescopes, one will get a series of circles or ellipses on the sky during a 12 hour interval. As we can see from Fig. 3.29, one does not have to cover all the spacings between the telescopes, because any antenna combination which has the same relative distance will describe the same path on the sky. In this way one can synthesize an antenna, a filled aperture, with a size equal to the maximum spacing between the telescopes. Interferometers working according to this principle are called *aperture synthesis telescopes*. If one covers all the spacings up to the maximum baseline, the result will be an accurate map of the source over the primary beam of an individual antenna element. Aperture synthesis telescopes therefore produce an image of the sky, i. e. a "*radio photograph*".

A typical aperture synthesis telescope consists of one fixed telescope and a number of movable telescopes, usually located on an east–west track, although T or Y configurations are also quite common. The number of telescopes used determines how fast one can synthesize a larger disk, because the number of possible antenna combinations increases as $n(n-1)$, where n is the number of telescopes. It is also possible to synthesize a large telescope with only one fixed and one movable telescope by changing the spacing between the telescopes every 12 hours, but then a full aperture synthesis can require several months of observing time. In order for this technique to work, the source must be constant, i. e. the signal cannot be time variable during the observing session.

The most efficient aperture synthesis telescope at present is the VLA (Very Large Array) in New Mexico, USA (Fig. 3.30). It consists of 27 paraboloid antennas, each with a diameter of 25 m, which are located on a Y-shaped track. The Y-formation was chosen because it provides a full aperture synthesis in 8 hours. Each antenna can be moved by a specially built carrier,

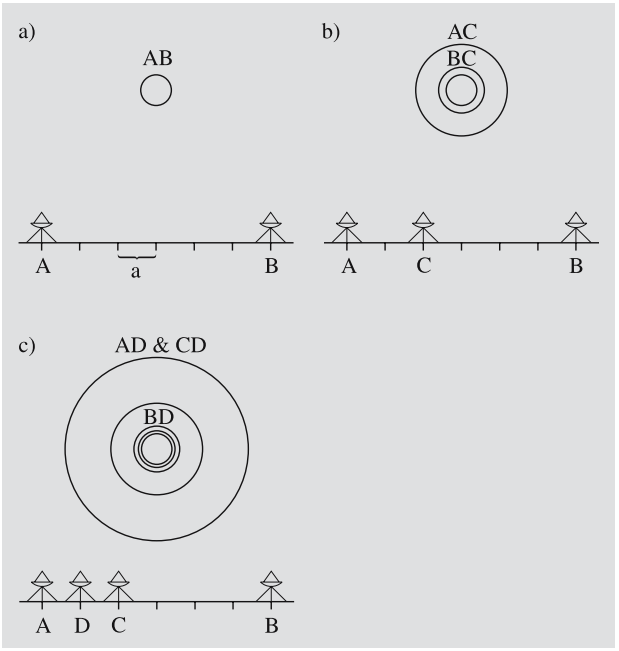


Fig. 3.29a–c. To illustrate the principle of aperture synthesis, let us consider an east–west oriented interferometer pointed towards the celestial north. Each antenna is identical, has a diameter D and operates at a wavelength λ. The minimum spacing between each antenna element is a, and the maximum spacing is $6a$. In (**a**) there are only two antennas, A and B, displaced by the maximum spacing $6a$. When the earth rotates, antennas A and B will, in the course of 12 hours, track a circle on the plane of the sky with a diameter $\lambda/(6a)$, the maximum resolution that can be achieved with this interferometer. In (**b**) the antenna C is added to the interferometer, thus providing two more baselines, which track the circles AC and BC with radii of $\lambda/(2a)$ and $\lambda/(4a)$, respectively. In (**c**) there is still another antenna D added to the interferometer. In this case two of the baselines are equal, AD and CD, and therefore only two new circles are covered on the plane of the sky. By adding more interferometer elements, one can fill in the missing parts within the primary beam, i. e. the beam of one single dish, and thus obtain a full coverage of the beam. It is also evident from (**c**), that not all of the antenna positions are needed to provide all the different spacings; some antenna spacings will in such a case be equal and therefore provide no additional information. Obtaining a full aperture synthesis with an east–west interferometer always takes 12 hours, if all spacings are available. Usually, however, several antenna elements are movable, in which case a full aperture synthesis can take a long time before all spacings are filled in

Fig. 3.30. The VLA at Socorro, New Mexico, is a synthesis telescope consisting of 27 movable antennas

and the locations of the telescopes are chosen to give optimal spacings for each configuration. In the largest configuration each arm is about 21 km long, thereby resulting in an antenna with an effective diameter of 35 km. If the VLA is used in its largest configuration and at its highest frequency, 23 GHz (1.3 cm), the resolution achieved is 0.1 arc second, clearly superior to any optical telescope. Similar resolution can also be obtained with the British MERLIN telescope, where already existing telescopes have been coupled together by radio links. Other well-known synthesis telescopes are the Cambridge 5 km array in Britain and the Westerbork array in the Netherlands, both located on east–west tracks.

Even higher resolution can be obtained with an extension of the aperture synthesis technique, called *VLBI* (Very Long Baseline Interferometry). With the VLBI technique the spacing between the antennas is restricted only by the size of the Earth. VLBI uses existing antennas (often on different continents), which are all pointed towards the same source. In this case the signal is recorded together with accurate timing signals from atomic clocks. The data files are correlated against each other, resulting in maps similar to those obtained with a normal aperture synthesis telescope. With VLBI techniques it is possible to achieve resolutions of $0.0001''$. Because interferometry is very sensitive to the distance between the telescopes, the VLBI technique also provides one of the most accurate methods to measure distances. Currently one can measure distances with an accuracy of a few centimetres on intercontinental baselines. This is utilized in geodetic VLBI experiments, which study continental drift and polar motion as a function of time.

In radio astronomy the maximum size of single antennas has also been reached. The trend is to build synthesis antennas, similar to the VLA in New Mexico. In the 1990's The United States built a chain of antennas extending across the whole continent, and the Australians have constructed a similar, but north–south antenna chain across their country.

More and more observations are being made in the submillimetre region. The disturbing effect of atmospheric water vapour becomes more serious at shorter wavelengths; thus, submillimetre telescopes must be lo-

cated on mountain tops, like optical telescopes. All parts of the mirror are actively controlled in order to accurately maintain the proper form like in the new optical telescopes. Several new submillimetre telescopes are under construction.

3.5 Other Wavelength Regions

All wavelengths of the electromagnetic spectrum enter the Earth from the sky. However, as mentioned in Sect. 3.1, not all radiation reaches the ground. The wavelength regions absorbed by the atmosphere have been studied more extensively since the 1970's, using Earth-orbiting satellites. Besides the optical and radio regions, there are only some narrow wavelength ranges in the infrared that can be observed from high mountain tops.

The first observations in each new wavelength region were usually carried out from balloons, but not until rockets came into use could observations be made from outside the atmosphere. The first actual observations of an X-ray source, for instance, were made on a rocket flight in June 1962, when the detector rose above the atmosphere for about 6 minutes. Satellites have made it possible to map the whole sky in the wavelength regions invisible from the ground.

Gamma Radiation. Gamma ray astronomy studies radiation quanta with energies of 10^5–10^{14} eV. The boundary between gamma and X-ray astronomy, 10^5 eV, corresponds to a wavelength of 10^{-11} m. The boundary is not fixed; the regions of hard (= high-energy) X-rays and soft gamma rays partly overlap.

While ultraviolet, visible and infrared radiation are all produced by changes in the energy states of the electron envelopes of atoms, gamma and hard X-rays are produced by transitions in atomic nuclei or in mutual interactions of elementary particles. Thus observations of the shortest wavelengths give information on processes different from those giving rise to longer wavelengths.

The first observations of gamma sources were obtained at the end of the 1960's, when a device in the OSO 3 satellite (Orbiting Solar Observatory) detected gamma rays from the Milky Way. Later on, some satellites were especially designed for gamma astronomy, notably SAS 2, COS B, HEAO 1 and 3, and the Compton Gamma Ray Observatory. The most effective satellite at present is the European Integral, launched in 2002.

The quanta of gamma radiation have energies a million times greater than those of visible light, but they cannot be observed with the same detectors. These observations are made with various *scintillation detectors*, usually composed of several layers of detector plates, where gamma radiation is transformed by the photoelectric effect into visible light, detectable by photomultipliers.

The energy of a gamma quantum can be determined from the depth to which it penetrates the detector. Analyzing the trails left by the quanta gives information on their approximate direction. The field of view is limited by the grating. The directional accuracy is low, and in gamma astronomy the resolution is far below that in other wavelength regions.

X-rays. The observational domain of X-ray astronomy includes the energies between 10^2 and 10^5 eV, or the wavelengths 10–0.01 nm. The regions 10–0.1 nm and 0.1–0.01 nm are called *soft* and *hard X-rays*, respectively. X-rays were discovered in the late 19th century. Systematic studies of the sky at X-ray wavelengths only became possible in the 1970's with the advent of satellite technology.

The first all-sky mapping was made in the early 1970's by SAS 1 (Small Astronomical Satellite), also called Uhuru. At the end of the 1970's, two High-Energy Astronomy Observatories, HEAO 1 and 2 (the latter called Einstein), mapped the sky with much higher sensitivity than Uhuru.

The Einstein Observatory was able to detect sources about a thousand times fainter than earlier X-ray telescopes. In optical astronomy, this would correspond to a jump from a 15 cm reflector to a 5 m telescope. Thus X-ray astronomy has developed in 20 years as much as optical astronomy in 300 years.

The latest X-ray satellites have been the American Chandra and the European XMM-Newton, both launched in 1999.

Besides satellites mapping the whole sky, there have been several satellites observing the X-ray radiation of the Sun. The first effective telescopes were installed in

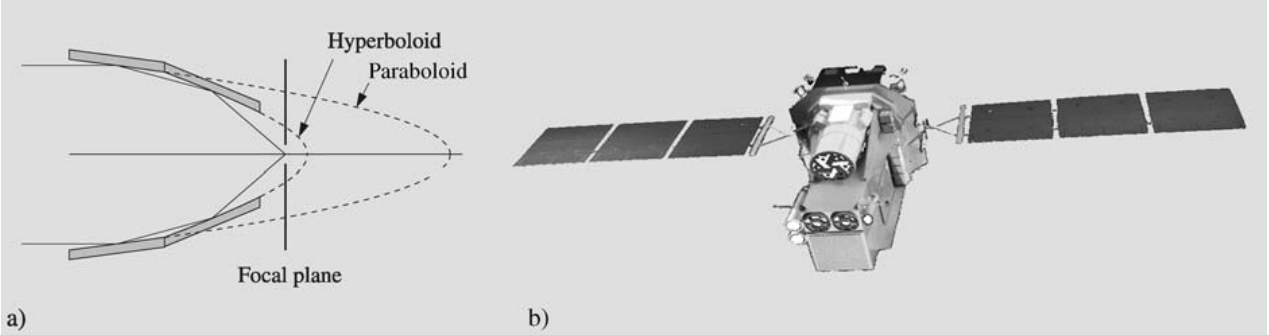


Fig. 3.31. (**a**) X-rays are not reflected by an ordinary mirror, and the principle of grazing reflection must be used for collecting them. Radiation meets the paraboloid mirror at a very small angle, is reflected onto a hyperboloid mirror and further to a focal point. In practice, several mirrors are placed one inside another, collecting radiation in a common focus (**b**) The European Integral gamma ray observatory was launched in 2002. (Picture ESA)

the Skylab space station, and they were used to study the Sun in 1973–74. In the 1990's, the European Soho started making regular X-ray observations of the Sun.

The first X-ray telescopes used detectors similar to those in gamma astronomy. Their directional accuracy was never better than a few arc minutes. The more precise X-ray telescopes utilize the principle of *grazing reflection* (Fig. 3.31). An X-ray hitting a surface perpendicularly is not reflected, but absorbed. If, however, X-rays meet the mirror nearly parallel to its surface, just grazing it, a high quality surface can reflect the ray.

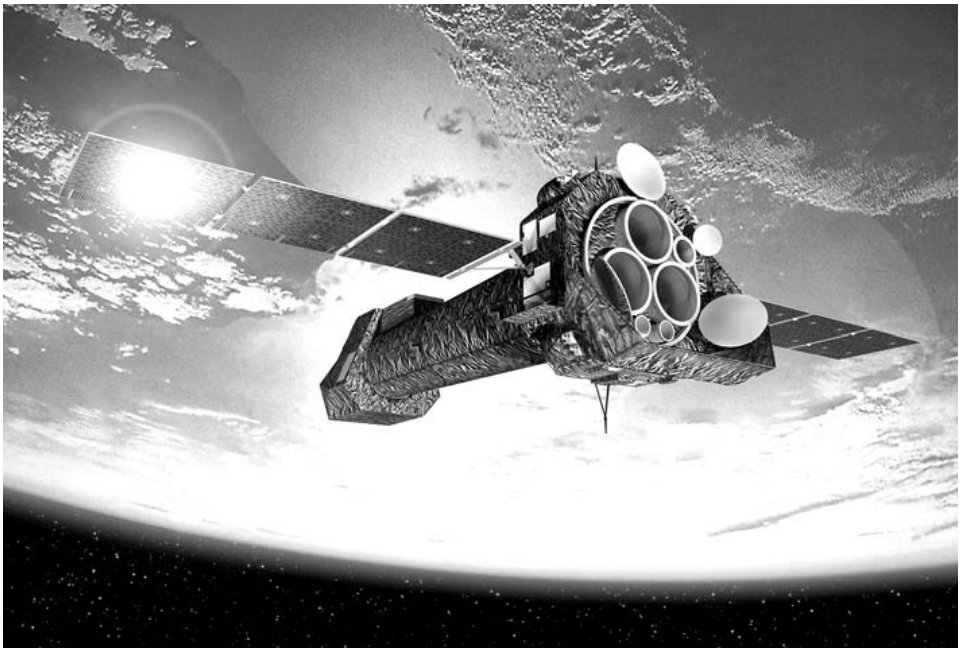
The mirror of an X-ray reflector is on the inner surface of a slowly narrowing cone. The outer part of the surface is a paraboloid and the inner part a hyperboloid. The rays are reflected by both surfaces and meet at a focal plane. In practice, several tubes are installed one within another. For instance, the four cones of the Einstein Observatory had as much polished optical surface as a normal telescope with a diameter of 2.5 m. The resolution in X-ray telescopes is of the order of a few arc seconds and the field of view about 1 deg.

The detectors in X-ray astronomy are usually *Geiger–Müller counters*, *proportional counters* or scintillation detectors. Geiger–Müller and proportional counters are boxes filled with gas. The walls form a cathode, and an anode wire runs through the middle of the box; in more accurate counters, there are several anode wires. An X-ray quantum entering the box ionizes the gas, and the potential difference between the anode and cathode gives rise to a current of electrons and positive ions.

Ultraviolet Radiation. Between X-rays and the optical region lies the domain of ultraviolet radiation, with wavelengths between 10 and 400 nm. Most ultraviolet observations have been carried out in the *soft UV* region, at wavelengths near those of optical light, since most of the UV radiation is absorbed by the atmosphere. The wavelengths below 300 nm are completely blocked out. The short wavelength region from 10 to 91.2 nm is called the *extreme ultraviolet* (EUV, XUV).

Extreme ultraviolet was one of the last regions of the electromagnetic radiation to be observed systematically. The reason for this is that the absorption of interstellar hydrogen makes the sky practically opaque at these wavelengths. The visibility in most directions is limited to some hundred light years in the vicinity of the Sun. In some directions, however, the density of the interstellar gas is so low that even extragalactic objects can be seen. The first dedicated EUV satellite was the Extreme Ultraviolet Explorer (EUVE), operating in 1992–2000. It observed about a thousand EUV sources. In EUV grazing reflection telescopes similar to those used in X-ray astronomy are employed.

In nearly all branches of astronomy important information is obtained by observations of ultraviolet radiation. Many emission lines from stellar chromospheres or coronas, the Lyman lines of atomic hydrogen, and most of the radiation from hot stars are found in the UV domain. In the *near-ultraviolet*, telescopes can be made similar to optical telescopes and, equipped with a photometer or spectrometer, installed in a satellite orbiting the Earth.



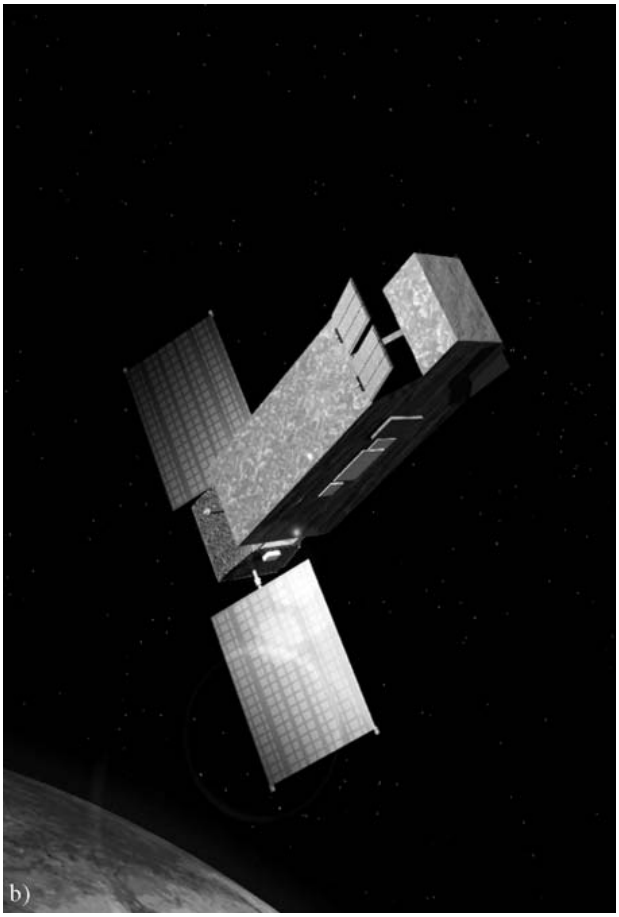


Fig. 3.32. (**a**) The European X-ray satellite XMM-Newton was launched in 1999. (Drawing D. Ducros, XMM Team, ESA) (**b**) FUSE satellite has photographed far ultraviolet objects from Earth orbit since 1999. (Graphics NASA/JHU Applied Physics Laboratory)

The most effective satellites in the UV have been the European TD-1, the American Orbiting Astronomical Observatories OAO 2 and 3 (Copernicus), the International Ultraviolet Explorer IUE and the Soviet Astron. The instruments of the TD-1 satellite included both a photometer and a spectrometer. The satellite measured the magnitudes of over 30,000 stars in four different spectral regions between 135 and 274 nm, and registered UV spectra from over 1000 stars. The OAO satellites were also used to measure magnitudes and spectra, and OAO 3 worked for over eight years.

The IUE satellite, launched in 1978, was one of the most successful astronomical satellites. IUE had a 45 cm Ritchey-Chrétien telescope with an aperture ratio of $f/15$ and a field of view of 16 arc minutes. The satellite had two spectrographs to measure spectra of higher or lower resolution in wavelength intervals of 115–200 nm or 190–320 nm. For registration of the spectra, a Vidicon camera was used. IUE worked on the orbit for 20 years.

Infrared Radiation. Radiation with longer wavelengths than visible light is called infrared radiation. This region extends from about 1 micrometre to 1 millimetre, where the radio region begins. Sometimes the *near-infrared*, at wavelengths below 5 m,

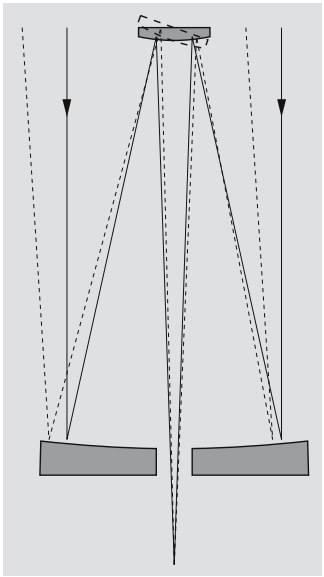


Fig. 3.33. Refractors are not suitable for infrared telescopes, because infrared radiation cannot penetrate glass. The Cassegrain reflectors intended especially for infrared observations have secondary mirrors nodding rapidly back and forth between the object and the background near the object. By subtracting the brightness of the background from the brightness of the object, the background can be eliminated

and the submillimetre domain, at wavelengths between 0.1 and 1 mm, are considered separate wavelength regions.

In infrared observations radiation is collected by a telescope, as in the optical region. The incoming radiation consists of radiation from the object, from the background and from the telescope itself. Both the source and the background must be continually measured, the difference giving the radiation from the object. The background measurements are usually made with a Cassegrain secondary mirror oscillating between the source and the background at a rate of, say, 100 oscillations per second, and thus the changing background can be eliminated. To register the measurements, semiconductor detectors are used. The detector must always be cooled to minimize its own thermal radiation. Sometimes the whole telescope is cooled.

Infrared observatories have been built on high mountain tops, where most of the atmospheric water vapour remains below. Some favourable sites are, e. g. Mauna Kea on Hawaii, Mount Lemon in Arizona and Pico del Teide on Tenerife. For observations in the far-infrared these mountains are not high enough; these observations are carried out, e. g. on aeroplanes. One of the best-equipped planes is the Kuiper Airborne Observatory, named after the well-known planetary scientist Gerard Kuiper.

Fig. 3.34. The most effective infrared satellite at present is the American Spitzer, launched in 2003. (Drawing NASA)

Balloons and satellites are also used for infrared observations. The most successful infrared observatories so far have been the InfraRed Astronomy Satellite IRAS, the European Infrared Space Observatory ISO, and the present-day Spitzer (originally SIRTF, Space InfraRed Telescope Facility). A very succesful satellite was the 1989 launched COBE (Cosmic Background Explorer), which mapped the background radiation in submillimetre and infrared wavelengths. The Microwave Anisotropy Probe (MAP) has continued the work of COBE, starting in 2001.

3.6 Other Forms of Energy

Besides electromagnetic radiation, energy arrives from space in other forms: particles (*cosmic rays*, *neutrinos*) and *gravitational radiation*.

Fig. 3.35. The LIGO Livingston Observatory seen from the air. (Photo LIGO/Caltech)

Cosmic Rays. Cosmic rays, consisting of electrons and totally ionized nuclei of atoms, are received in equal amounts from all directions. Their incoming directions do not reveal their origin, since cosmic rays are electrically charged; thus their paths are continually changed when they move through the magnetic fields of the Milky Way. The high energies of cosmic rays mean that they have to be produced by high-energy phenomena like supernova explosions. The majority of cosmic rays are protons (nearly 90%) and helium nuclei (10%), but some are heavier nuclei; their energies lie between 10^8 and 10^{20} eV.

The most energetic cosmic rays give rise to *secondary radiation* when they hit molecules of the atmosphere. This secondary radiation can be observed from the ground, but primary cosmic rays can only be directly observed outside the atmosphere. The detectors used to observe cosmic rays are similar to those used in particle physics. Since Earth-based accelerators reach energies of only about 10^{12} eV, cosmic rays offer an excellent "natural" laboratory for particle physics. Many satellites and spacecraft have detectors for cosmic rays.

Neutrinos. Neutrinos are elementary particles with no electric charge and a mass equal to zero or, at any rate, less than 1/10,000 of the mass of the electron. Most neutrinos are produced in nuclear reactions within stars; since they react very weakly with other matter, they escape directly from the stellar interior.

Neutrinos are very difficult to observe; the first method of detection was the radiochemical method. As a reactive agent, e. g. tetrachloroethene (C_2Cl_4) can be used. When a neutrino hits a chlorine atom, the chlorine is transformed into argon, and an electron is freed:

$$^{37}\mathrm{Cl} + \nu \rightarrow {}^{37}\mathrm{Ar} + \mathrm{e}^- \, .$$

The argon atom is radioactive and can be observed. Instead of chlorine, lithium and gallium might be used to detect neutrinos. The first gallium detectors have been running in Italy and Russia from the end of the 1980's.

Another observation method is based on the Čerenkov radiation produced by neutrinos in extremely pure water. The flashes of light are registered with photomultipliers, and thus it is possible to find out the direction of the radiation. This method is used e. g. in the Japanese Kamiokande detector.

Neutrino detectors must be located deep under the ground to protect them from the secondary radiation caused by cosmic rays.

The detectors have observed neutrinos from the Sun, and the Supernova 1987A in the Large Magellanic Cloud was also observed in 1987.

Gravitational Radiation. Gravitational astronomy is as young as neutrino astronomy. The first attempts to measure gravitational waves were made in the 1960's. Gravitational radiation is emitted by accelerating masses, just as electromagnetic radiation is emitted by electric charges in accelerated motion. Detection of gravitational waves is very difficult, and they have yet to be directly observed.

The first type of gravitational wave antenna was the *Weber cylinder*. It is an aluminium cylinder which starts vibrating at its proper frequency when hit by a gravitational pulse. The distance between the ends of the cylinder changes by about 10^{-17} m, and the changes in the length are studied by strain sensors welded to the side of the cylinder.

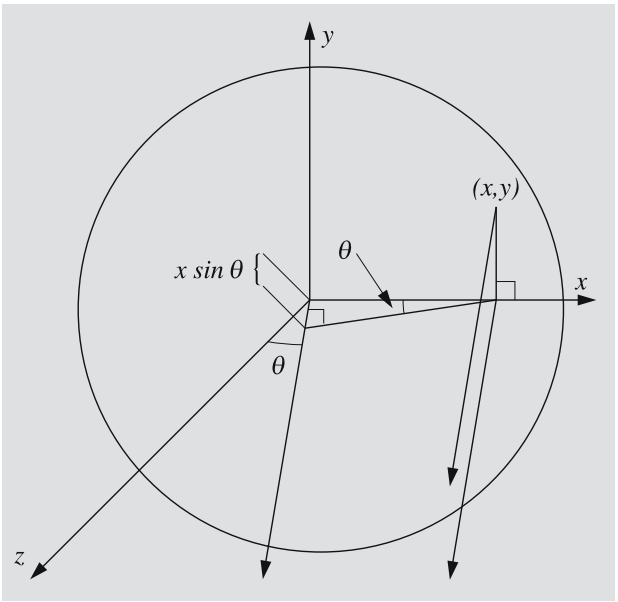
Another type of modern gravity radiation detectors measures "spatial strain" induced by gravity waves and consists of two sets of mirrors in directions perpendicular to each other (Michelson interferometer), or one set of parallel mirrors (Fabry–Perot interferometer). The relative distances between the mirrors are moni-

tored by laser interferometers. If a gravity pulse passes the detector, the distances change and the changes can be measured. The longest baseline between the mirrors is in the American LIGO (Laser Interferometer Gravitational-wave Observatory) system, about 25 km (Fig. 3.35). LIGO made the first scientific observations in 2002.

* Diffraction by a Circular Aperture

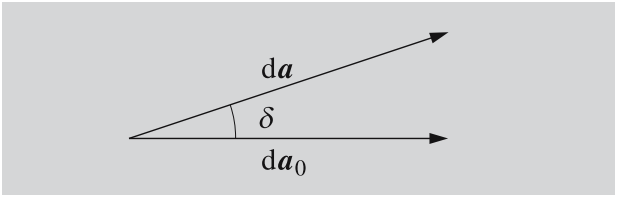
Consider a circular hole of radius R in the xy plane. Coherent light enters the hole from the direction of the negative z axis (see figure). We consider light rays leaving the hole parallel to the xz plane forming an angle θ with the z axis. The light waves interfere on a screen far away. The phase difference between a wave through a point (x, y) and a wave going through the centre of the hole can be calculated from the different path lengths $s = x \sin\theta$:

$$\delta = \frac{s}{\lambda} 2\pi = \frac{2\pi \sin\theta}{\lambda} x \equiv kx \; .$$



Thus, the phase difference δ depends on the x coordinate only. The sum of the amplitudes of the waves from a small surface element is proportional to the area of the element $\mathrm{d}x\,\mathrm{d}y$. Let the amplitude coming through the centre of the hole be $\mathrm{d}\boldsymbol{a}_0 = \mathrm{d}x\,\mathrm{d}y\hat{\boldsymbol{i}}$. The amplitude coming from the point (x, y) is then

$$\mathrm{d}\boldsymbol{a} = \mathrm{d}x\,\mathrm{d}y\,(\cos\delta\,\hat{\boldsymbol{i}} + \sin\delta\,\hat{\boldsymbol{j}}) \; .$$



We sum up the amplitudes coming from different points of the hole:

$$\boldsymbol{a} = \int\limits_{\text{Aperture}} \mathrm{d}\boldsymbol{a}$$

$$= \int\limits_{x=-R}^{R} \int\limits_{y=-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} (\cos kx\,\hat{\boldsymbol{i}} + \sin kx\,\hat{\boldsymbol{j}})\,\mathrm{d}y\,\mathrm{d}x$$

$$= 2\int\limits_{-R}^{R} \sqrt{R^2-x^2}(\cos kx\,\hat{\boldsymbol{i}} + \sin kx\,\hat{\boldsymbol{j}})\,\mathrm{d}x \; .$$

Since sine is an odd function ($\sin(-kx) = -\sin(kx)$), we get zero when we integrate the second term. Cosine is an even function, and so

$$a \propto \int\limits_{0}^{R} \sqrt{R^2-x^2}\cos kx\,\mathrm{d}x \; .$$

We substitute $x = Rt$ and define $p = kR = (2\pi r \sin\theta)/\lambda$, thus getting

$$a \propto \int\limits_{0}^{1} \sqrt{1-t^2}\cos pt\,\mathrm{d}t \; .$$

The zero points of the intensity observed on the screen are obtained from the zero points of the amplitude,

$$J(p) = \int\limits_{0}^{1} \sqrt{1-t^2}\cos pt\,\mathrm{d}t = 0 \; .$$

Inspecting the function $J(p)$, we see that the first zero is at $p = 3.8317$, or

$$\frac{2\pi R \sin\theta}{\lambda} = 3.8317 \, .$$

The radius of the diffraction disc in angular units can be estimated from the condition

$$\sin\theta = \frac{3.8317\lambda}{2\pi R} \approx 1.22 \, \frac{\lambda}{D} \, ,$$

where $D = 2R$ is the diameter of the hole.

In mirror telescopes diffraction is caused also by the support structure of the secondary mirror. If the aperture is more complex and only elementary mathematics is used calculations may become rather cumbersome. However, it can be shown that the diffraction pattern can be obtained as the Fourier transform of the aperture.

3.7 Examples

Example 3.1 The distance between the components of the binary star ζ Herculis is $1.38''$. What should the diameter of a telescope be to resolve the binary? If the focal length of the objective is 80 cm, what should the focal length of the eyepiece be to resolve the components, when the resolution of the eye is $2'$?

In the optical region, we can use the wavelength value of $\lambda \approx 550$ nm. The diameter of the objective is obtained from the equation for the resolution (3.4),

$$D \approx \frac{\lambda}{\theta} = \frac{550 \times 10^{-9}}{(1.38/3600) \times (\pi/180)} \, \mathrm{m}$$

$$= 0.08 \, \mathrm{m} = 8 \, \mathrm{cm} \, .$$

The required magnification is

$$\omega = \frac{2'}{1.38''} = 87 \, .$$

The magnification is given by

$$\omega = \frac{f}{f'} \, ,$$

and, thus, the focal length of the eyepiece should be

$$f' = \frac{f}{\omega} = \frac{80 \, \mathrm{cm}}{87} = 0.9 \, \mathrm{cm} \, .$$

Example 3.2 A telescope has an objective with a diameter of 90 mm and focal length of 1200 mm.

a) What is the focal length of an eyepiece, the exit pupil of which is 6 mm (about the size of the pupil of the eye)?
b) What is the magnification of such an eyepiece?
c) What is the angular diameter of the Moon seen through this telescope and eyepiece?

a) From Fig. 3.7 we get

$$L = \frac{f'}{f} D \, ,$$

whence

$$f' = f \frac{L}{D} = 1200 \, \mathrm{mm} \, \frac{6 \, \mathrm{mm}}{90 \, \mathrm{mm}}$$

$$= 80 \, \mathrm{mm} \, .$$

b) The magnification is $\omega = f/f' = 1200 \, \mathrm{mm}/80 \, \mathrm{mm} = 15$.
c) Assuming the angular diameter of the Moon is $\alpha = 31' = 0.52°$, its diameter through the telescope is $\omega\alpha = 7.8°$.

3.8 Exercises

Exercise 3.1 The Moon was photographed with a telescope, the objective of which had a diameter of 20 cm and focal length of 150 cm. The exposure time was 0.1 s.

a) What should the exposure time be, if the diameter of the objective were 15 cm and focal length 200 cm?
b) What is the size of the image of the Moon in both cases?
c) Both telescopes are used to look at the Moon with an eyepiece the focal length of which is 25 mm. What are the magnifications?

Exercise 3.2 The radio telescopes at Amherst, Massachusetts, and Onsala, Sweden, are used as an interferometer, the baseline being 2900 km.

a) What is the resolution at 22 GHz in the direction of the baseline?
b) What should be the size of an optical telescope with the same resolution?

4. Photometric Concepts and Magnitudes

Most astronomical observations utilize electromagnetic radiation in one way or another. We can obtain information on the physical nature of a radiation source by studying the energy distribution of its radiation. We shall now introduce some basic concepts that characterize electromagnetic radiation.

4.1 Intensity, Flux Density and Luminosity

Let us assume we have some radiation passing through a surface element $\mathrm{d}A$ (Fig. 4.1). Some of the radiation will leave $\mathrm{d}A$ within a solid angle $\mathrm{d}\omega$; the angle between $\mathrm{d}\omega$ and the normal to the surface is denoted by θ. The amount of energy with frequency in the range $[\nu, \nu+\mathrm{d}\nu]$ entering this solid angle in time $\mathrm{d}t$ is

$$\mathrm{d}E_\nu = I_\nu \cos\theta \, \mathrm{d}A \, \mathrm{d}\nu \, \mathrm{d}\omega \, \mathrm{d}t \, . \tag{4.1}$$

Here, the coefficient I_ν is the *specific intensity* of the radiation at the frequency ν in the direction of the solid angle $\mathrm{d}\omega$. Its dimension is $\mathrm{W\,m^{-2}\,Hz^{-1}\,sterad^{-1}}$.

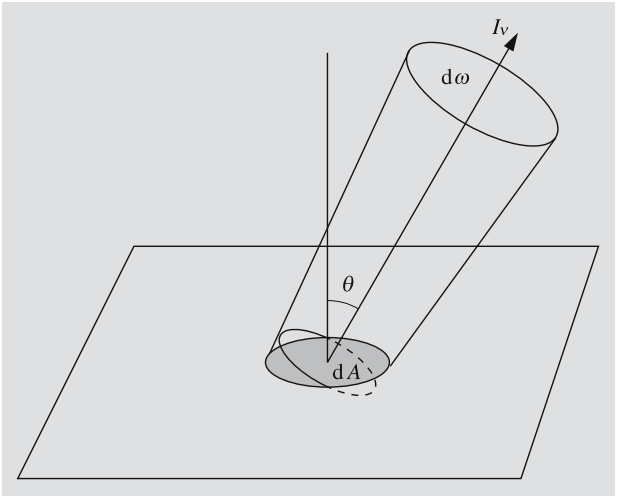


Fig. 4.1. The intensity I_ν of radiation is related to the energy passing through a surface element $\mathrm{d}A$ into a solid angle $\mathrm{d}\omega$, in a direction θ

The projection of the surface element $\mathrm{d}A$ as seen from the direction θ is $\mathrm{d}A_n = \mathrm{d}A\cos\theta$, which explains the factor $\cos\theta$. If the intensity does not depend on direction, the energy $\mathrm{d}E_\nu$ is directly proportional to the surface element perpendicular to the direction of the radiation.

The intensity including all possible frequencies is called the *total intensity* I, and is obtained by integrating I_ν over all frequencies:

$$I = \int_0^\infty I_\nu \, \mathrm{d}\nu \, .$$

More important quantities from the observational point of view are the *energy flux* (L_ν, L) or, briefly, the *flux* and the *flux density* (F_ν, F). The flux density gives the power of radiation per unit area; hence its dimension is $\mathrm{W\,m^{-2}\,Hz^{-1}}$ or $\mathrm{W\,m^{-2}}$, depending on whether we are talking about the flux density at a certain frequency or about the total flux density.

Observed flux densities are usually rather small, and $\mathrm{W\,m^{-2}}$ would be an inconveniently large unit. Therefore, especially in radio astronomy, flux densities are often expressed in *Janskys*; one Jansky (Jy) equals $10^{-26}\,\mathrm{W\,m^{-2}\,Hz^{-1}}$.

When we are observing a radiation source, we in fact measure the energy collected by the detector during some period of time, which equals the flux density integrated over the radiation-collecting area of the instrument and the time interval.

The flux density F_ν at a frequency ν can be expressed in terms of the intensity as

$$\begin{aligned} F_\nu &= \frac{1}{\mathrm{d}A\,\mathrm{d}\nu\,\mathrm{d}t} \int_S \mathrm{d}E_\nu \\ &= \int_S I_\nu \cos\theta \, \mathrm{d}\omega \, , \end{aligned} \tag{4.2}$$

where the integration is extended over all possible directions. Analogously, the total flux density is

$$F = \int_S I \cos\theta \, \mathrm{d}\omega \, .$$

Hannu Karttunen et al. (Eds.), *Photometric Concepts and Magnitudes*.
In: Hannu Karttunen et al. (Eds.), Fundamental Astronomy, 5th Edition. pp. 83–93 (2007)
DOI: 11685739_4 © Springer-Verlag Berlin Heidelberg 2007

For example, if the radiation is *isotropic*, i.e. if I is independent of the direction, we get

$$F = \int_S I \cos\theta \, \mathrm{d}\omega = I \int_S \cos\theta \, \mathrm{d}\omega \, . \tag{4.3}$$

The solid angle element $\mathrm{d}\omega$ is equal to a surface element on a unit sphere. In spherical coordinates it is (Fig. 4.2; also c.f. Appendix A.5):

$$\mathrm{d}\omega = \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \, .$$

Substitution into (4.3) gives

$$F = I \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} \cos\theta \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi = 0 \, ,$$

so there is no net flux of radiation. This means that there are equal amounts of radiation entering and leaving the surface. If we want to know the amount of radiation passing through the surface, we can find, for example, the radiation leaving the surface. For isotropic radiation this is

$$F_\uparrow = I \int_{\theta=0}^{\pi/2} \int_{\phi=0}^{2\pi} \cos\theta \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi = \pi I \, . \tag{4.4}$$

In the astronomical literature, terms such as intensity and brightness are used rather vaguely. Flux density is hardly ever called flux density but intensity or (with luck) flux. Therefore the reader should always carefully check the meaning of these terms.

Flux means the power going through some surface, expressed in watts. The flux emitted by a star into a solid angle ω is $L = \omega r^2 F$, where F is the flux density observed at a distance r. *Total flux* is the flux passing through a closed surface encompassing the source. Astronomers usually call the total flux of a star the *luminosity* L. We can also talk about the luminosity L_ν at a frequency ν ($[L_\nu] = \mathrm{W\,Hz^{-1}}$). (This must not be confused with the luminous flux used in physics; the latter takes into account the sensitivity of the eye.)

If the source (like a typical star) radiates isotropically, its radiation at a distance r is distributed evenly on a spherical surface whose area is $4\pi r^2$ (Fig. 4.3). If the flux density of the radiation passing through this surface is F, the total flux is

$$L = 4\pi r^2 F \, . \tag{4.5}$$

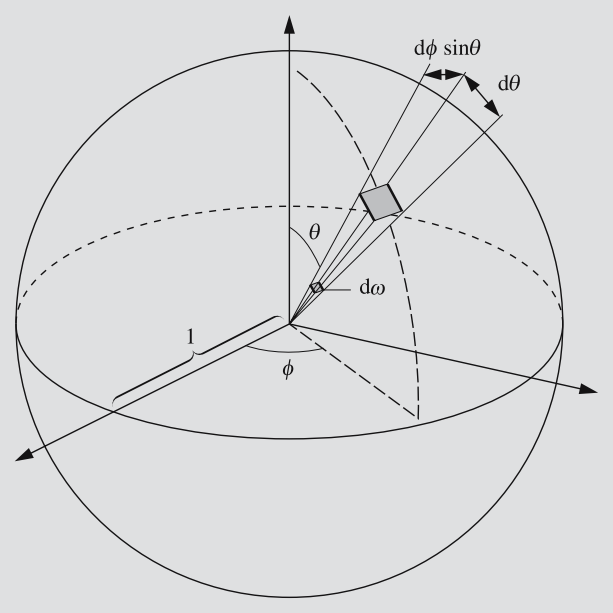


Fig. 4.2. An infinitesimal solid angle $\mathrm{d}\omega$ is equal to the corresponding surface element on a unit sphere: $\mathrm{d}\omega = \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi$

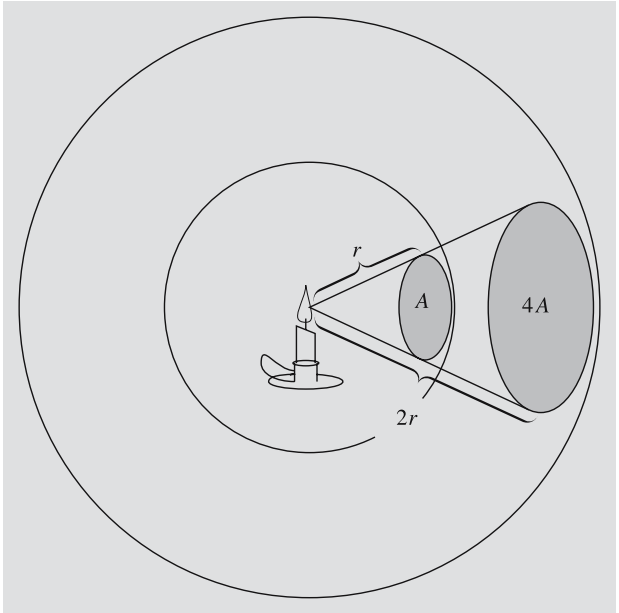


Fig. 4.3. An energy flux which at a distance r from a point source is distributed over an area A is spread over an area $4A$ at a distance $2r$. Thus the flux density decreases inversely proportional to the distance squared

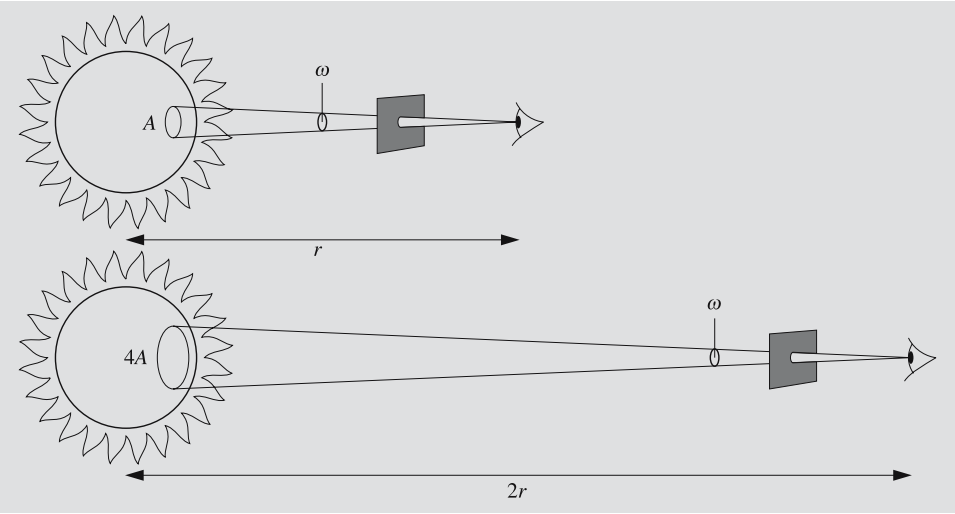


Fig. 4.4. An observer sees radiation coming from a constant solid angle ω. The area giving off radiation into this solid angle increases when the source moves further away ($A \propto r^2$). Therefore the surface brightness or the observed flux density per unit solid angle remains constant

If we are outside the source, where radiation is not created or destroyed, the luminosity does not depend on distance. The flux density, on the other hand, falls off proportional to $1/r^2$.

For extended objects (as opposed to objects such as stars visible only as points) we can define the *surface brightness* as the flux density per unit solid angle (Fig. 4.4). Now the observer is at the apex of the solid angle. The surface brightness is independent of distance, which can be understood in the following way. The flux density arriving from an area A is inversely proportional to the distance squared. But also the solid angle subtended by the area A is proportional to $1/r^2$ ($\omega = A/r^2$). Thus the surface brightness $B = F/\omega$ remains constant.

The *energy density* u of radiation is the amount of energy per unit volume ($\mathrm{J\,m^{-3}}$):

$$u = \frac{1}{c} \int_S I \, \mathrm{d}\omega \, . \tag{4.6}$$

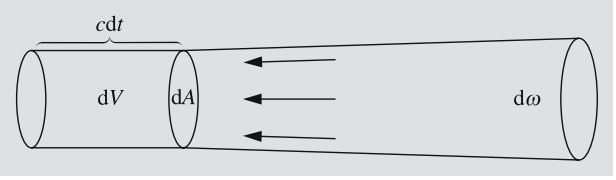


Fig. 4.5. In time $\mathrm{d}t$, the radiation fills a volume $\mathrm{d}V = c\,\mathrm{d}t\,\mathrm{d}A$, where $\mathrm{d}A$ is the surface element perpendicular to the propagation direction of the radiation

This can be seen as follows. Suppose we have radiation with intensity I arriving from a solid angle $\mathrm{d}\omega$ perpendicular to the surface $\mathrm{d}A$ (Fig. 4.5). In the time $\mathrm{d}t$, the radiation travels a distance $c\,\mathrm{d}t$ and fills a volume $\mathrm{d}V = c\,\mathrm{d}t\,\mathrm{d}A$. Thus the energy in the volume $\mathrm{d}V$ is (now $\cos\theta = 1$)

$$\mathrm{d}E = I\,\mathrm{d}A\,\mathrm{d}\omega\,\mathrm{d}t = \frac{1}{c} I\,\mathrm{d}\omega\,\mathrm{d}V \, .$$

Hence the energy density $\mathrm{d}u$ of the radiation arriving from the solid angle $\mathrm{d}\omega$ is

$$\mathrm{d}u = \frac{\mathrm{d}E}{\mathrm{d}V} = \frac{1}{c} I\,\mathrm{d}\omega \, ,$$

and the total energy density is obtained by integrating this over all directions. For isotropic radiation we get

$$u = \frac{4\pi}{c} I \, . \tag{4.7}$$

4.2 Apparent Magnitudes

As early as the second century B. C., Hipparchos divided the visible stars into six classes according to their apparent brightness. The first class contained the brightest stars and the sixth the faintest ones still visible to the naked eye.

The response of the human eye to the brightness of light is not linear. If the flux densities of three stars are in the proportion 1:10:100, the brightness difference of the

first and second star seems to be equal to the difference of the second and third star. Equal brightness ratios correspond to equal apparent brightness differences: the human perception of brightness is logarithmic.

The rather vague classification of Hipparchos was replaced in 1856 by Norman R. Pogson. The new, more accurate classification followed the old one as closely as possible, resulting in another of those illogical definitions typical of astronomy. Since a star of the first class is about one hundred times brighter than a star of the sixth class, Pogson defined the ratio of the brightnesses of classes n and $n+1$ as $\sqrt[5]{100} = 2.512$.

The brightness class or *magnitude* can be defined accurately in terms of the observed flux density F ($[F] = \mathrm{W\,m^{-2}}$). We decide that the magnitude 0 corresponds to some preselected flux density F_0. All other magnitudes are then defined by the equation

$$m = -2.5 \lg \frac{F}{F_0} \,. \tag{4.8}$$

Note that the coefficient is exactly 2.5, not 2.512! Magnitudes are dimensionless quantities, but to remind us that a certain value is a magnitude, we can write it, for example, as 5 mag or 5^{m}.

It is easy to see that (4.8) is equivalent to Pogson's definition. If the magnitudes of two stars are m and $m+1$ and their flux densities F_m and F_{m+1}, respectively, we have

$$\begin{aligned} m - (m+1) &= -2.5 \lg \frac{F_m}{F_0} + 2.5 \lg \frac{F_{m+1}}{F_0} \\ &= -2.5 \lg \frac{F_m}{F_{m+1}} \,, \end{aligned}$$

whence

$$\frac{F_m}{F_{m+1}} = \sqrt[5]{100} \,.$$

In the same way we can show that the magnitudes m_1 and m_2 of two stars and the corresponding flux densities F_1 and F_2 are related by

$$m_1 - m_2 = -2.5 \lg \frac{F_1}{F_2} \,. \tag{4.9}$$

Magnitudes extend both ways from the original six values. The magnitude of the brightest star, Sirius, is in fact negative -1.5. The magnitude of the Sun is -26.8 and that of a full moon -12.5. The magnitude of the faintest objects observed depends on the size of the telescope, the sensitivity of the detector and the exposure time. The limit keeps being pushed towards fainter objects; currently the magnitudes of the faintest observed objects are over 30.

4.3 Magnitude Systems

The *apparent magnitude* m, which we have just defined, depends on the instrument we use to measure it. The sensitivity of the detector is different at different wavelengths. Also, different instruments detect different wavelength ranges. Thus the flux measured by the instrument equals not the total flux, but only a fraction of it. Depending on the method of observation, we can define various magnitude systems. Different magnitudes have different zero points, i. e. they have different flux densities F_0 corresponding to the magnitude 0. The zero points are usually defined by a few selected standard stars.

In daylight the human eye is most sensitive to radiation with a wavelength of about 550 nm, the sensitivity decreasing towards red (longer wavelengths) and violet (shorter wavelengths). The magnitude corresponding to the sensitivity of the eye is called the *visual magnitude* m_{v}.

Photographic plates are usually most sensitive at blue and violet wavelengths, but they are also able to register radiation not visible to the human eye. Thus the *photographic magnitude* m_{pg} usually differs from the visual magnitude. The sensitivity of the eye can be simulated by using a yellow filter and plates sensitised to yellow and green light. Magnitudes thus observed are called *photovisual magnitudes* m_{pv}.

If, in ideal case, we were able to measure the radiation at all wavelengths, we would get the *bolometric magnitude* m_{bol}. In practice this is very difficult, since part of the radiation is absorbed by the atmosphere; also, different wavelengths require different detectors. (In fact there is a gadget called the bolometer, which, however, is not a real bolometer but an infrared detector.) The bolometric magnitude can be derived from the visual magnitude if we know the *bolometric correction* BC:

$$m_{\mathrm{bol}} = m_{\mathrm{v}} - \mathrm{BC} \,. \tag{4.10}$$

By definition, the bolometric correction is zero for radiation of solar type stars (or, more precisely, stars of the spectral class F5). Although the visual and bolometric

magnitudes can be equal, the flux density corresponding to the bolometric magnitude must always be higher. The reason of this apparent contradiction is in the different values of F_0.

The more the radiation distribution differs from that of the Sun, the higher the bolometric correction is. The correction is positive for stars both cooler or hotter than the Sun. Sometimes the correction is defined as $m_{\rm bol} = m_{\rm v} + {\rm BC}$ in which case ${\rm BC} \leq 0$ always. The chance for errors is, however, very small, since we must have $m_{\rm bol} \leq m_{\rm v}$.

The most accurate magnitude measurements are made using photoelectric photometers. Usually filters are used to allow only a certain wavelength band to enter the detector. One of the multicolour magnitude systems used widely in photoeletric photometry is the UBV system developed in the early 1950's by *Harold L. Johnson* and *William W. Morgan*. Magnitudes are measured through three filters, U = ultraviolet, B = blue and V = visual. Figure 4.6 and Table 4.1 give the wavelength bands of these filters. The magnitudes observed through these filters are called U, B and V magnitudes, respectively.

The UBV system was later augmented by adding more bands. One commonly used system is the five colour UBVRI system, which includes R = red and I = infrared filters.

There are also other broad band systems, but they are not as well standardised as the UBV, which has been defined moderately well using a great number of standard stars all over the sky. The magnitude of an object is obtained by comparing it to the magnitudes of standard stars.

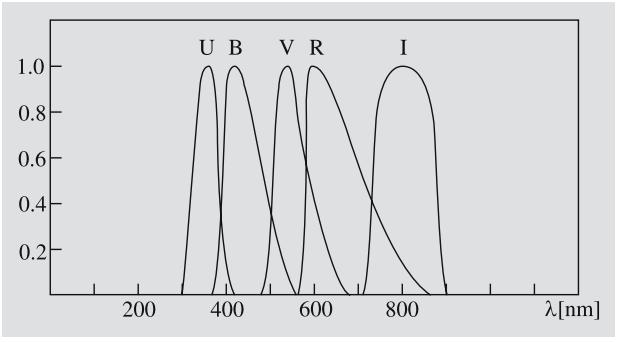


Fig. 4.6. Relative transmission profiles of filters used in the UBVRI magnitude system. The maxima of the bands are normalized to unity. The R and I bands are based on the system of Johnson, Cousins and Glass, which includes also infrared bands J, H, K, L and M. Previously used R and I bands differ considerably from these

Table 4.1. Wavelength bands of the UBVRI and uvby filters and their effective (≈ average) wavelengths

Magnitude		Band width [nm]	Effective wavelength [nm]
U	ultraviolet	66	367
B	blue	94	436
V	visual	88	545
R	red	138	638
I	infrared	149	797
u	ultraviolet	30	349
v	violet	19	411
b	blue	18	467
y	yellow	23	547

In *Strömgren's* four-colour or *uvby system*, the bands passed by the filters are much narrower than in the UBV system. The uvby system is also well standardized, but it is not quite as common as the UBV. Other narrow band systems exist as well. By adding more filters, more information on the radiation distribution can be obtained.

In any multicolour system, we can define *colour indices*; a colour index is the difference of two magnitudes. By subtracting the B magnitude from U we get the colour index $U - B$, and so on. If the UBV system is used, it is common to give only the V magnitude and the colour indices $U - B$ and $B - V$.

The constants F_0 in (4.8) for U, B and V magnitudes have been selected in such a way that the colour indices $B - V$ and $U - B$ are zero for stars of spectral type A0 (for spectral types, see Chap. 8). The surface temperature of such a star is about 10,000 K. For example, Vega (α Lyr, spectral class A0V) has $V = 0.03$, $B - V = U - B = 0.00$. The Sun has $V = -26.8$, $B - V = 0.62$ and $U - B = 0.10$.

Before the UBV system was developed, a colour index C.I., defined as

$$\text{C.I.} = m_{\rm pg} - m_{\rm v} \, ,$$

was used. Since $m_{\rm pg}$ gives the magnitude in blue and $m_{\rm v}$ in visual, this index is related to $B - V$. In fact,

$$\text{C.I.} = (B - V) - 0.11 \, .$$

4.4 Absolute Magnitudes

Thus far we have discussed only apparent magnitudes. They do not tell us anything about the true brightness of stars, since the distances differ. A quantity measuring the intrinsic brightness of a star is the *absolute magnitude*. It is defined as the apparent magnitude at a distance of 10 parsecs from the star (Fig. 4.7).

We shall now derive an equation which relates the apparent magnitude m, the absolute magnitude M and the distance r. Because the flux emanating from a star into a solid angle ω has, at a distance r, spread over an area ωr^2, the flux density is inversely proportional to the distance squared. Therefore the ratio of the flux density at a distance r, $F(r)$, to the flux density at a distance of 10 parsecs, $F(10)$, is

$$\frac{F(r)}{F(10)} = \left(\frac{10\,\mathrm{pc}}{r}\right)^2 .$$

Thus the difference of magnitudes at r and 10 pc, or the *distance modulus* $m - M$, is

$$m - M = -2.5 \lg \frac{F(r)}{F(10)} = -2.5 \lg \left(\frac{10\,\mathrm{pc}}{r}\right)^2$$

or

$$m - M = 5 \lg \frac{r}{10\,\mathrm{pc}} . \qquad (4.11)$$

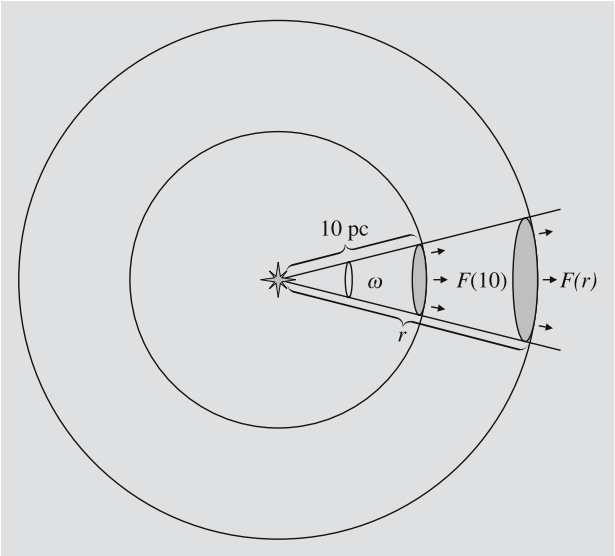


Fig. 4.7. The flux density at a distance of 10 parsecs from the star defines its absolute magnitude

For historical reasons, this equation is almost always written as

$$m - M = 5 \lg r - 5 , \qquad (4.12)$$

which is valid *only* if the distance is expressed in parsecs. (The logarithm of a dimensional quantity is, in fact, physically absurd.) Sometimes the distance is given in kiloparsecs or megaparsecs, which require different constant terms in (4.12). To avoid confusion, we highly recommend the form (4.11).

Absolute magnitudes are usually denoted by capital letters. Note, however, that the U, B and V magnitudes are apparent magnitudes. The corresponding absolute magnitudes are M_U, M_B and M_V.

The absolute bolometric magnitude can be expressed in terms of the luminosity. Let the total flux density at a distance $r = 10$ pc be F and let $F_\odot$ be the equivalent quantity for the Sun. Since the luminosity is $L = 4\pi r^2 F$, we get

$$M_\mathrm{bol} - M_{\mathrm{bol},\odot} = -2.5 \lg \frac{F}{F_\odot} = -2.5 \lg \frac{L/4\pi r^2}{L_\odot/4\pi r^2} ,$$

or

$$M_\mathrm{bol} - M_{\mathrm{bol},\odot} = -2.5 \lg \frac{L}{L_\odot} . \qquad (4.13)$$

The absolute bolometric magnitude $M_\mathrm{bol} = 0$ corresponds to a luminosity $L_0 = 3.0 \times 10^{28}$ W.

4.5 Extinction and Optical Thickness

Equation (4.11) shows how the apparent magnitude increases (and brightness decreases!) with increasing distance. If the space between the radiation source and the observer is not completely empty, but contains some interstellar medium, (4.11) no longer holds, because part of the radiation is absorbed by the medium (and usually re-emitted at a different wavelength, which may be outside the band defining the magnitude), or scattered away from the line of sight. All these radiation losses are called the *extinction*.

Now we want to find out how the extinction depends on the distance. Assume we have a star radiating a flux L_0 into a solid angle ω in some wavelength range. Since the medium absorbs and scatters radiation, the

flux L will now decrease with increasing distance r (Fig. 4.8). In a short distance interval $[r, r+\mathrm{d}r]$, the extinction $\mathrm{d}L$ is proportional to the flux L and the distance travelled in the medium:

$$\mathrm{d}L = -\alpha L\,\mathrm{d}r\ . \tag{4.14}$$

The factor α tells how effectively the medium can obscure radiation. It is called the *opacity*. From (4.14) we see that its dimension is $[\alpha] = \mathrm{m}^{-1}$. The opacity is zero for a perfect vacuum and approaches infinity when the substance becomes really murky. We can now define a dimensionless quantity, the *optical thickness* τ by

$$\mathrm{d}\tau = \alpha\,\mathrm{d}r\ . \tag{4.15}$$

Substituting this into (4.14) we get

$$\mathrm{d}L = -L\,\mathrm{d}\tau\ .$$

Next we integrate this from the source (where $L = L_0$ and $r = 0$) to the observer:

$$\int_{L_0}^{L} \frac{\mathrm{d}L}{L} = -\int_0^{\tau} \mathrm{d}\tau\ ,$$

which gives

$$L = L_0 \mathrm{e}^{-\tau}\ . \tag{4.16}$$

Here, τ is the optical thickness of the material between the source and the observer and L, the observed flux. Now, the flux L falls off exponentially with increasing optical thickness. Empty space is perfectly transparent, i. e. its opacity is $\alpha = 0$; thus the optical thickness does not increase in empty space, and the flux remains constant.

Let F_0 be the flux density on the surface of a star and $F(r)$, the flux density at a distance r. We can express the fluxes as

$$L = \omega r^2 F(r)\ , \quad L_0 = \omega R^2 F_0\ ,$$

where R is the radius of the star. Substitution into (4.16) gives

$$F(r) = F_0 \frac{R^2}{r^2} \mathrm{e}^{-\tau}\ .$$

For the absolute magnitude we need the flux density at a distance of 10 parsecs, $F(10)$, which is still evaluated without extinction:

$$F(10) = F_0 \frac{R^2}{(10\,\mathrm{pc})^2}\ .$$

The distance modulus $m - M$ is now

$$\begin{aligned} m - M &= -2.5 \lg \frac{F(r)}{F(10)} \\ &= 5 \lg \frac{r}{10\,\mathrm{pc}} - 2.5 \lg \mathrm{e}^{-\tau} \\ &= 5 \lg \frac{r}{10\,\mathrm{pc}} + (2.5 \lg \mathrm{e})\tau \end{aligned}$$

or

$$m - M = 5 \lg \frac{r}{10\,\mathrm{pc}} + A\ , \tag{4.17}$$

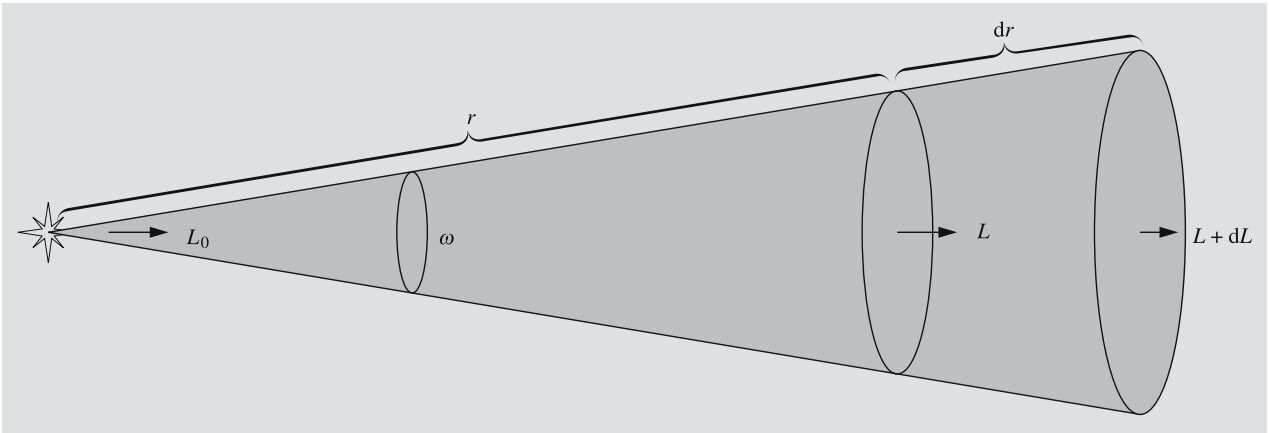


Fig. 4.8. The interstellar medium absorbs and scatters radiation; this usually reduces the energy flux L in the solid angle ω ($\mathrm{d}L \leq 0$)

where $A \geq 0$ is the extinction in magnitudes due to the entire medium between the star and the observer. If the opacity is constant along the line of sight, we have

$$\tau = \alpha \int_0^r \mathrm{d}r = \alpha r \,,$$

and (4.17) becomes

$$m - M = 5 \lg \frac{r}{10\,\mathrm{pc}} + ar \,, \tag{4.18}$$

where the constant $a = 2.5\alpha \lg \mathrm{e}$ gives the extinction in magnitudes per unit distance.

Colour Excess. Another effect caused by the interstellar medium is the *reddening of light*: blue light is scattered and absorbed more than red. Therefore the colour index $B - V$ increases. The visual magnitude of a star is, from (4.17),

$$V = M_\mathrm{V} + 5 \lg \frac{r}{10\,\mathrm{pc}} + A_\mathrm{V} \,, \tag{4.19}$$

where M_V is the absolute visual magnitude and A_V is the extinction in the V passband. Similarly, we get for the blue magnitudes

$$B = M_\mathrm{B} + 5 \lg \frac{r}{10\,\mathrm{pc}} + A_\mathrm{B} \,.$$

The observed colour index is now

$$B - V = M_\mathrm{B} - M_\mathrm{V} + A_\mathrm{B} - A_\mathrm{V} \,,$$

or

$$B - V = (B - V)_0 + E_{\mathrm{B-V}} \,, \tag{4.20}$$

where $(B - V)_0 = M_\mathrm{B} - M_\mathrm{V}$ is the *intrinsic colour* of the star and $E_{\mathrm{B-V}} = (B - V) - (B - V)_0$ is the *colour excess*. Studies of the interstellar medium show that the ratio of the visual extinction A_V to the colour excess $E_{\mathrm{B-V}}$ is almost constant for all stars:

$$R = \frac{A_\mathrm{V}}{E_{\mathrm{B-V}}} \approx 3.0 \,.$$

This makes it possible to find the visual extinction if the colour excess is known:

$$A_\mathrm{V} \approx 3.0\, E_{\mathrm{B-V}} \,. \tag{4.21}$$

When A_V is obtained, the distance can be solved directly from (4.19), when V and M_V are known.

We shall study interstellar extinction in more detail in Sect. 15.1 (“Interstellar Dust”).

Atmospheric Extinction. As we mentioned in Sect. 3.1, the Earth’s atmosphere also causes extinction. The observed magnitude m depends on the location of the observer and the zenith distance of the object, since these factors determine the distance the light has to travel in the atmosphere. To compare different observations, we must first *reduce* them, i. e. remove the atmospheric effects somehow. The magnitude m_0 thus obtained can then be compared with other observations.

If the zenith distance z is not too large, we can approximate the atmosphere by a plane layer of constant thickness (Fig. 4.9). If the thickness of the atmosphere is used as a unit, the light must travel a distance

$$X = 1/\cos z = \sec z \tag{4.22}$$

in the atmosphere. The quantity X is the *air mass*. According to (4.18), the magnitude increases linearly with the distance X:

$$m = m_0 + kX \,, \tag{4.23}$$

where k is the *extinction coefficient*.

The extinction coefficient can be determined by observing the same source several times during a night with as wide a zenith distance range as possible. The observed magnitudes are plotted in a diagram as a function of the air mass X. The points lie on a straight line the slope of which gives the extinction coefficient k. When

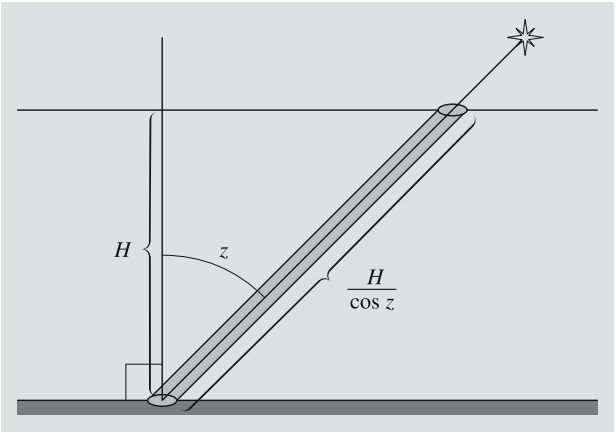


Fig. 4.9. If the zenith distance of a star is z, the light of the star travels a distance $H/\cos z$ in the atmosphere; H is the height of the atmosphere

this line is extrapolated to $X = 0$, we get the magnitude m_0, which is the apparent magnitude outside the atmosphere.

In practice, observations with zenith distances higher than 70° (or altitudes less than 20°) are not used to determine k and m_0, since at low altitudes the curvature of the atmosphere begins to complicate matters. The value of the extinction coefficient k depends on the observation site and time and also on the wavelength, since extinction increases strongly towards short wavelengths.

4.6 Examples

Example 4.1 Show that intensity is independent of distance.

Suppose we have some radiation leaving the surface element $\mathrm{d}A$ in the direction θ. The energy entering the solid angle $\mathrm{d}\omega$ in time $\mathrm{d}t$ is

$$\mathrm{d}E = I \cos\theta \, \mathrm{d}A \, \mathrm{d}\omega \, \mathrm{d}t \ ,$$

where I is the intensity. If we have another surface $\mathrm{d}A'$ at a distance r receiving this radiation from direction θ', we have

$$\mathrm{d}\omega = \mathrm{d}A' \cos\theta' / r^2 \ .$$

The definition of the intensity gives

$$\mathrm{d}E = I' \cos\theta' \, \mathrm{d}A' \, d\omega' \, \mathrm{d}t \ ,$$

where I' is the intensity at $\mathrm{d}A'$ and

$$\mathrm{d}\omega' = \mathrm{d}A \cos\theta / r^2 \ .$$

Substitution of $\mathrm{d}\omega$ and $\mathrm{d}\omega'$ into the expressions of $\mathrm{d}E$ gives

$$I \cos\theta \, \mathrm{d}\theta \, \mathrm{d}A \frac{\mathrm{d}A' \cos\theta'}{r^2} \, \mathrm{d}t$$
$$= I' \cos\theta' \, \mathrm{d}A' \frac{\mathrm{d}A \cos\theta}{r^2} \, \mathrm{d}t \quad \Rightarrow \quad I' = I \ .$$

Thus the intensity remains constant in empty space.

Example 4.2 *Surface Brightness of the Sun*

Assume that the Sun radiates isotropically. Let R be the radius of the Sun, $F_\odot$ the flux density on the surface of the Sun and F the flux density at a distance r. Since the luminosity is constant,

$$L = 4\pi R^2 F_\odot = 4\pi r^2 F \ ,$$

the flux density equals

$$F = F_\odot \frac{R^2}{r^2} \ .$$

At a distance $r \gg R$, the Sun subtends a solid angle

$$\omega = \frac{A}{r^2} = \frac{\pi R^2}{r^2} \ ,$$

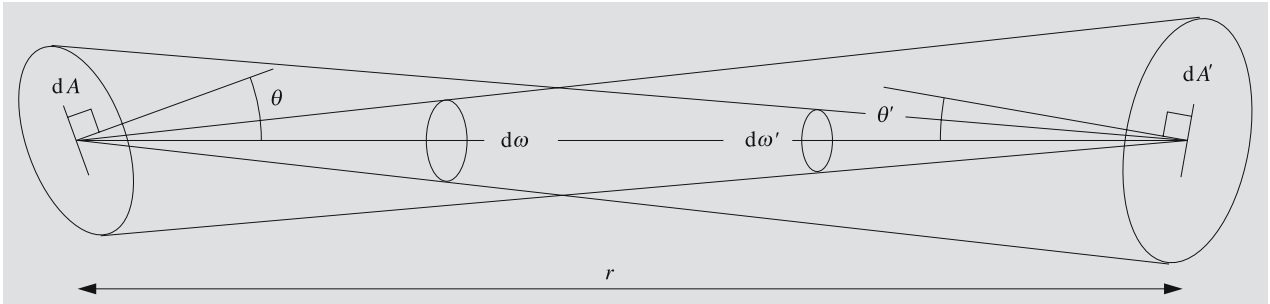
where $A = \pi R^2$ is the cross section of the Sun. The surface brightness B is

$$B = \frac{F}{\omega} = \frac{F_\odot}{\pi} \ .$$

Applying (4.4) we get

$$B = I_\odot \ .$$

Thus the surface brightness is independent of distance and equals the intensity. We have found a simple



interpretation for the somewhat abstract concept of intensity.

The flux density of the Sun on the Earth, the *solar constant*, is $S_\odot \approx 1370\ \mathrm{W\,m^{-2}}$. The angular diameter of the Sun is $\alpha = 32'$, whence

$$\frac{R}{r} = \frac{\alpha}{2} = \frac{1}{2} \times \frac{32}{60} \times \frac{\pi}{180} = 0.00465\ \mathrm{rad}\ .$$

The solid angle subtended by the Sun is

$$\omega = \pi \left(\frac{R}{r}\right)^2 = \pi \times 0.00465^2$$

$$= 6.81 \times 10^{-5}\ \mathrm{sterad}\ .$$

and the surface brightness

$$B = \frac{S_\odot}{\omega} = 2.01 \times 10^7\ \mathrm{W\,m^{-2}\,sterad^{-1}}\ .$$

Example 4.3 *Magnitude of a Binary Star*

Since magnitudes are logarithmic quantities, they can be a little awkward for some purposes. For example, we cannot add magnitudes like flux densities. If the magnitudes of the components of a binary star are 1 and 2, the total magnitude is certainly not 3. To find the total magnitude, we must first solve the flux densities from

$$1 = -2.5 \lg \frac{F_1}{F_0}\ , \quad 2 = -2.5 \lg \frac{F_2}{F_0}\ ,$$

which give

$$F_1 = F_0 \times 10^{-0.4}\ , \quad F_2 = F_0 \times 10^{-0.8}\ .$$

Thus the total flux density is

$$F = F_1 + F_2 = F_0(10^{-0.4} + 10^{-0.8})$$

and the total magnitude,

$$m = -2.5 \lg \frac{F_0(10^{-0.4} + 10^{-0.8})}{F_0}$$

$$= -2.5 \lg 0.5566 = 0.64\ .$$

Example 4.4 The distance of a star is $r = 100\,\mathrm{pc}$ and its apparent magnitude $m = 6$. What is its absolute magnitude?

Substitution into (4.11)

$$m - M = 5 \lg \frac{r}{10\,\mathrm{pc}}$$

gives

$$M = 6 - 5 \lg \frac{100}{10} = 1\ .$$

Example 4.5 The absolute magnitude of a star is $M = -2$ and the apparent magnitude $m = 8$. What is the distance of the star?

We can solve the distance r from (4.11):

$$r = 10\,\mathrm{pc} \times 10^{(m-M)/5} = 10 \times 10^{10/5}\mathrm{pc}$$

$$= 1000\,\mathrm{pc} = 1\,\mathrm{kpc}\ .$$

Example 4.6 Although the amount of interstellar extinction varies considerably from place to place, we can use an average value of 2 mag/kpc near the galactic plane. Find the distance of the star in Example 4.5, assuming such extinction.

Now the distance must be solved from (4.18):

$$8 - (-2) = 5 \lg \frac{r}{10} + 0.002\,r\ ,$$

where r is in parsecs. This equation cannot be solved analytically, but we can always use a numerical method. We try a simple iteration (Appendix A.7), rewriting the equation as

$$r = 10 \times 10^{2 - 0.0004\,r}\ .$$

The value $r = 1000$ pc found previously is a good initial guess:

$$r_0 = 1000$$

$$r_1 = 10 \times 10^{2\,-\,0.0004\,\times\,1000} = 398$$

$$r_2 = 693$$

$$\vdots$$

$$r_{12} = r_{13} = 584\ .$$

The distance is $r \approx 580\,\mathrm{pc}$, which is much less than our earlier value 1000 pc. This should be quite obvious,

since due to extinction, radiation is now reduced much faster than in empty space.

Example 4.7 What is the optical thickness of a layer of fog, if the Sun seen through the fog seems as bright as a full moon in a cloudless sky?

The apparent magnitudes of the Sun and the Moon are -26.8 and -12.5, respectively. Thus the total extinction in the cloud must be $A = 14.3$. Since

$$A = (2.5 \lg \mathrm{e})\,\tau \ ,$$

we get

$$\tau = A/(2.5 \lg \mathrm{e}) = 14.3/1.086 = 13.2 \ .$$

The optical thickness of the fog is 13.2. In reality, a fraction of the light scatters several times, and a few of the multiply scattered photons leave the cloud along the line of sight, reducing the total extinction. Therefore the optical thickness must be slightly higher than our value.

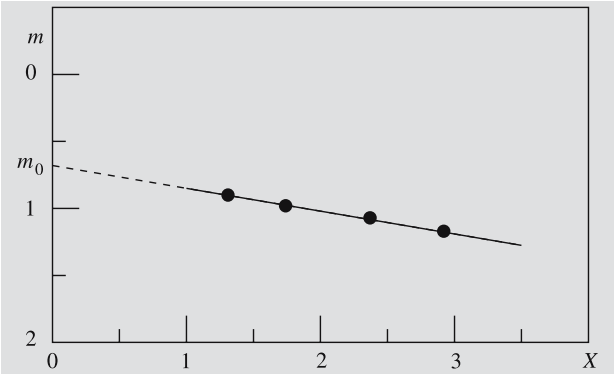
Example 4.8 *Reduction of Observations*

The altitude and magnitude of a star were measured several times during a night. The results are given in the following table.

Altitude	Zenith distance	Air mass	Magnitude
50°	40°	1.31	0.90
35°	55°	1.74	0.98
25°	65°	2.37	1.07
20°	70°	2.92	1.17

By plotting the observations as in the following figure, we can determine the extinction coefficient k and the magnitude m_0 outside the atmosphere. This can be done graphically (as here) or using a least-squares fit.

Extrapolation to the air mass $X = 0$ gives $m_0 = 0.68$. The slope of the line gives $k = 0.17$.



4.7 Exercises

Exercise 4.1 The total magnitude of a triple star is 0.0. Two of its components have magnitudes 1.0 and 2.0. What is the magnitude of the third component?

Exercise 4.2 The absolute magnitude of a star in the Andromeda galaxy (distance 690 kpc) is $M = 5$. It explodes as a supernova, becoming one billion (10^9) times brighter. What is its apparent magnitude?

Exercise 4.3 Assume that all stars have the same absolute magnitude and stars are evenly distributed in space. Let $N(m)$ be the number of stars brighter than m magnitudes. Find the ratio $N(m+1)/N(m)$.

Exercise 4.4 The V magnitude of a star is 15.1, $B - V = 1.6$, and absolute magnitude $M_\mathrm{V} = 1.3$. The extinction in the direction of the star in the visual band is $a_\mathrm{V} = 1\ \mathrm{mag\,kpc^{-1}}$. What is the intrinsic colour of the star?

Exercise 4.5 Stars are observed through a triple window. Each surface reflects away 15% of the incident light.

a) What is the magnitude of Regulus ($M_\mathrm{V} = 1.36$) seen through the window?
b) What is the optical thickness of the window?

5. Radiation Mechanisms

In the previous chapters we have studied the physical properties and detection of electromagnetic radiation. Next we shall briefly discuss concepts related to emission and absorption of radiation. Since we can give here only a summary of some essential results without delving into quantum mechanical explanations, the reader interested in the details is advised to consult any good physics textbook.

5.1 Radiation of Atoms and Molecules

Electromagnetic radiation is emitted or absorbed when an atom or a molecule moves from one energy level to another. If the energy of the atom decreases by an amount ΔE, the atom emits or radiates a quantum of electromagnetic radiation, called a *photon*, whose frequency ν is given by the equation

$$\Delta E = h\nu \,, \tag{5.1}$$

where h is the *Planck constant*, $h = 6.6256 \times 10^{-34}$ J s. Similarly, if the atom receives or absorbs a photon of a frequency ν, its energy increases by $\Delta E = h\nu$.

The classical model describes an atom as a nucleus surrounded by a swarm of electrons. The nucleus consists of Z protons, each having a charge $+e$ and N electrically neutral neutrons; Z is the *charge number* of the atom and $A = Z + N$ is its *mass number*. A neutral atom has as many electrons (charge $-e$) as protons.

An energy level of an atom usually refers to an energy level of its electrons. The energy E of an electron cannot take arbitrary values; only certain energies are allowed: the energy levels are *quantized*. An atom can emit or absorb radiation only at certain frequencies $\nu_{\rm if}$ corresponding to energy differences between some initial and final states i and f: $|E_{\rm i} - E_{\rm f}| = h\nu_{\rm if}$. This gives rise to the *line spectrum*, specific for each element (Fig. 5.1). Hot gas under low pressure produces an *emission spectrum* consisting of such discrete lines. If the same gas is cooled down and observed against a source of white light (which has a continuous spectrum), the same lines are seen as dark *absorption lines*.

At low temperatures most atoms are in their lowest energy state, the *ground state*. Higher energy levels are *excitation states*; a transition from lower to higher state is called *excitation*. Usually the excited atom will return to the lower state very rapidly, radiating a photon (*spontaneous emission*); a typical lifetime of an excited state might be 10^{-8} seconds. The frequency of the emitted photon is given by (5.1). The atom may return to the lower state directly or through some intermediate states, emitting one photon in each transition.

Downward transitions can also be induced by radiation. Suppose our atom has swallowed a photon and become excited. Another photon, whose frequency ν corresponds to some possible downward transition from the excited state, can now irritate the atom, causing it to jump to a lower state, emitting a photon with the same frequency ν. This is called *induced* or *stimulated emission*. Photons emitted spontaneously leave the atom randomly in all directions with random phases: the radiation is isotropic and incoherent. Induced radiation, on the other hand, is *coherent*; it propagates in the same direction as and in phase with the inducing radiation.

The zero level of the energy states is usually chosen so that a bound electron has negative energy and a free electron positive energy (cf. the energy integral of planetary orbits, Chap. 6). If an electron with energy $E < 0$ receives more energy than $|E|$, it will leave the atom, which becomes an ion. In astrophysics ionization is often called a *bound-free* transition (Fig. 5.2). Unlike in excitation all values of energy ($E > 0$) are now possible. The extraneous part of the absorbed energy goes to the kinetic energy of the liberated electron. The inverse process, in which an atom captures a free electron, is the *recombination* or free–bound transition.

When an electron scatters from a nucleus or an ion without being captured, the electromagnetic interaction can change the kinetic energy of the electron producing *free–free* radiation. In a very hot gas ($T > 10^6$ K) hydrogen is fully ionized, and the free–free radiation is the most important source of emission. It is then usually called *thermal bremsstrahlung*. The latter part of the name derives from the fact that decelerating electrons hitting the anode of an X-ray tube emit similar

Hannu Karttunen et al. (Eds.), *Radiation Mechanisms*.
In: Hannu Karttunen et al. (Eds.), Fundamental Astronomy, 5th Edition. pp. 95–112 (2007)
DOI: 11685739_5 © Springer-Verlag Berlin Heidelberg 2007

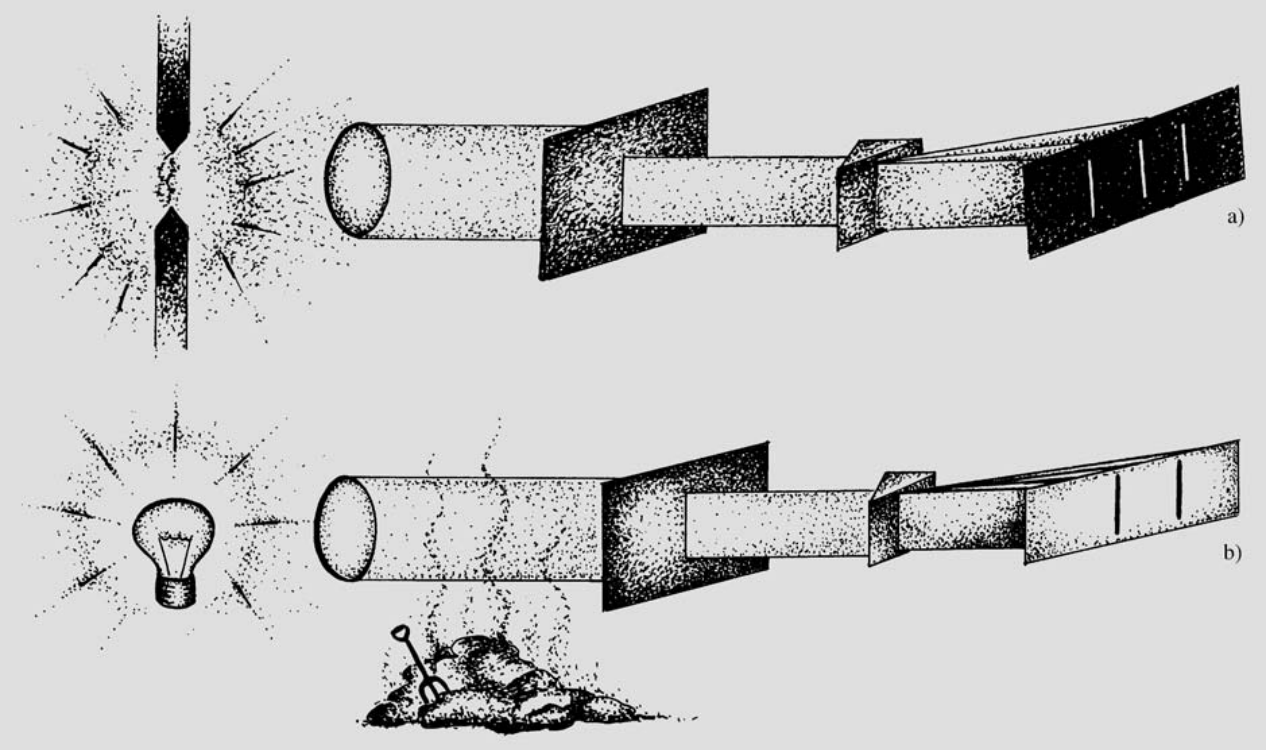


Fig. 5.1a,b. Origin of line spectra. (**a**) Emission spectrum. Atoms of glowing gas returning from excited states to lower states emit photons with frequencies corresponding to the energy difference of the states. Each element emits its own characteristic wavelengths, which can be measured by spreading the light into a spectrum with a prism or diffraction grating. (**b**) Absorption spectrum. When white light containing all wavelengths travels through gas, the wavelengths characteristic of the gas are absorbed

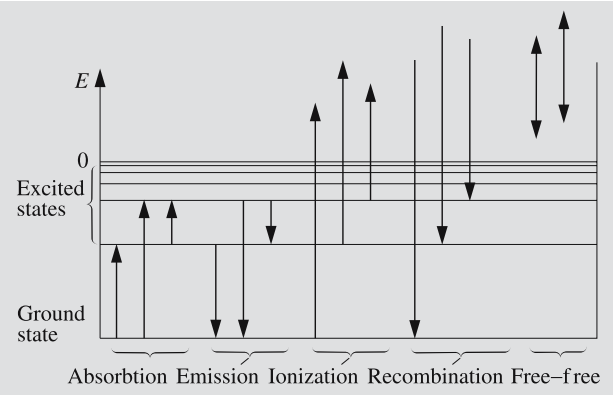


Fig. 5.2. Different kinds of transitions between energy levels. Absorption and emission occur between two bound states, whereas ionization and recombination occur between a bound and a free state. Interaction of an atom with an free electron can result in a free–free transition

radiation. In an analogous way the absorption process can be called a bound–bound transition.

Electromagnetic radiation is transverse wave motion; the electric and magnetic fields oscillate perpendicular to each other and also perpendicular to the direction of propagation. The light of an ordinary incandescent lamp has a random distribution of electric fields vibrating in all directions. If the directions of electric fields in the plane perpendicular to the direction of propagation are not evenly distributed, the radiation is *polarized* (Fig. 5.3). The direction of polarization of *linearly polarized* light means the plane determined by the electric vector and the direction of the light ray. If the electric vector describes a circle, the radiation is *circularly polarized*. If the amplitude of the electric field varies at the same time, the polarization is *elliptic*.

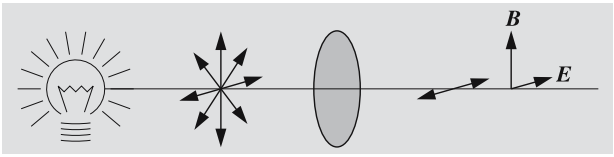


Fig. 5.3. Polarization of light. The light of an incandescent bulb contains all possible directions of vibration and is therefore unpolarized. Some crystals, for example, pass electric fields oscillating only in certain directions, and the transmitted part of the light becomes linearly polarized. ***E*** is the electric field and ***B*** the magnetic field

If polarized radiation travels through a magnetic field, the direction of the polarization will rotate. The amount of such *Faraday rotation* is proportional to the component of the magnetic field parallel to the line of sight, number of electrons along the line of sight, distance travelled, and square of the wavelength of the radiation.

Scattering is an absorption followed by an instantaneous emission at the same wavelength but usually in a new direction. On the macroscopic scale, radiation seems to be reflected by the medium. The light coming from the sky is sunlight scattered from atmospheric molecules. Scattered light is always polarized, the degree of polarization being highest in the direction perpendicular to the direction of the original radiation.

5.2 The Hydrogen Atom

The hydrogen atom is the simplest atom, consisting of a proton and an electron. According to the Bohr model the electron orbits the proton in a circular orbit. (In spite of the fact that this model has very little to do with reality, it can be successfully used to predict some properties of the hydrogen atom.) Bohr's first postulate says that the angular momentum of the electron must be a multiple of $\hbar$:

$$mvr = n\hbar \,, \tag{5.2}$$

where

m = mass of the electron ,
v = speed of the electron ,
r = radius of the orbit ,
n = the principal quantum number ,
$\quad n = 1, 2, 3, \ldots$,
$\hbar = h/2\pi$,
h = the Planck constant .

The quantum mechanical interpretation of Bohr's first postulate is obvious: the electron is represented as a standing wave, and the "length of the orbit" must be a multiple of the de Broglie wavelength, $\lambda = \hbar/p = \hbar/mv$.

A charged particle in a circular orbit (and thus in accelerated motion) should emit electromagnetic radiation, losing energy, were it to obey the rules of classical electrodynamics. Therefore our electron should spiral down towards the nucleus. But obviously, Nature does not behave this way, and we have to accept Bohr's second postulate, which says that an electron moving in an allowed orbit around a nucleus does not radiate. Radiation is emitted only when the electron jumps from a higher energy state to a lower one. The emitted quantum has an energy $h\nu$, equal to the energy difference of these states:

$$h\nu = E_{n_2} - E_{n_1} \,. \tag{5.3}$$

We shall now try to find the energy of an electron in the state E_n. Coulomb's law gives the force pulling the electron towards the proton:

$$F = \frac{1}{4\pi\epsilon_0}\frac{e^2}{r_n^2} \,, \tag{5.4}$$

where

ϵ_0 = the vacuum permittivity
$\quad = 8.85 \times 10^{-12}\,\mathrm{N^{-1}\,m^{-2}\,C^2}$,
e = the charge of the electron $= 1.6 \times 10^{-19}\,\mathrm{C}$,
r_n = the distance between the electron
$\quad$ and the proton .

The acceleration of a particle moving in a circular orbit of radius r_n is

$$a = \frac{v_n^2}{r_n} \,,$$

and applying Newton's second law ($F = ma$), we get

$$\frac{mv_n^2}{r_n} = \frac{1}{4\pi\epsilon_0}\frac{e^2}{r_n^2}. \tag{5.5}$$

From (5.2) and (5.5) it follows that

$$v_n = \frac{e^2}{4\pi\epsilon_0\hbar}\frac{1}{n}\,, \quad r_n = \frac{4\pi\epsilon_0\hbar^2}{me^2}n^2\,.$$

The total energy of an electron in the orbit n is now

$$\begin{aligned} E_n = T + V &= \frac{1}{2}mv_n^2 - \frac{1}{4\pi\epsilon_0}\frac{e^2}{r_n} \\ &= -\frac{me^4}{32\pi^2\epsilon_0^2\hbar^2}\frac{1}{n^2} \equiv -C\frac{1}{n^2}\,, \end{aligned} \tag{5.6}$$

where C is a constant. For the ground state ($n = 1$), we get from (5.6)

$$E_1 = -2.18\times10^{-18}\,\mathrm{J} = -13.6\,\mathrm{eV}\,.$$

From (5.3) and (5.6) we get the energy of the quantum emitted in the transition $E_{n_2} \to E_{n_1}$:

$$h\nu = E_{n_2} - E_{n_1} = C\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)\,. \tag{5.7}$$

In terms of the wavelength λ this can be expressed as

$$\frac{1}{\lambda} = \frac{\nu}{c} = \frac{C}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right) \equiv R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)\,, \tag{5.8}$$

where R is the *Rydberg constant*, $R = 1.097\times10^7\ \mathrm{m}^{-1}$.

Equation (5.8) was derived experimentally for $n_1 = 2$ by Johann Jakob Balmer as early as 1885. That is why we call the set of lines produced by transitions $E_n \to E_2$ the *Balmer series*. These lines are in the visible part of the spectrum. For historical reasons the Balmer lines are often denoted by symbols H_α, H_β, H_γ etc. If the electron returns to its ground state ($E_n \to E_1$), we get the *Lyman*

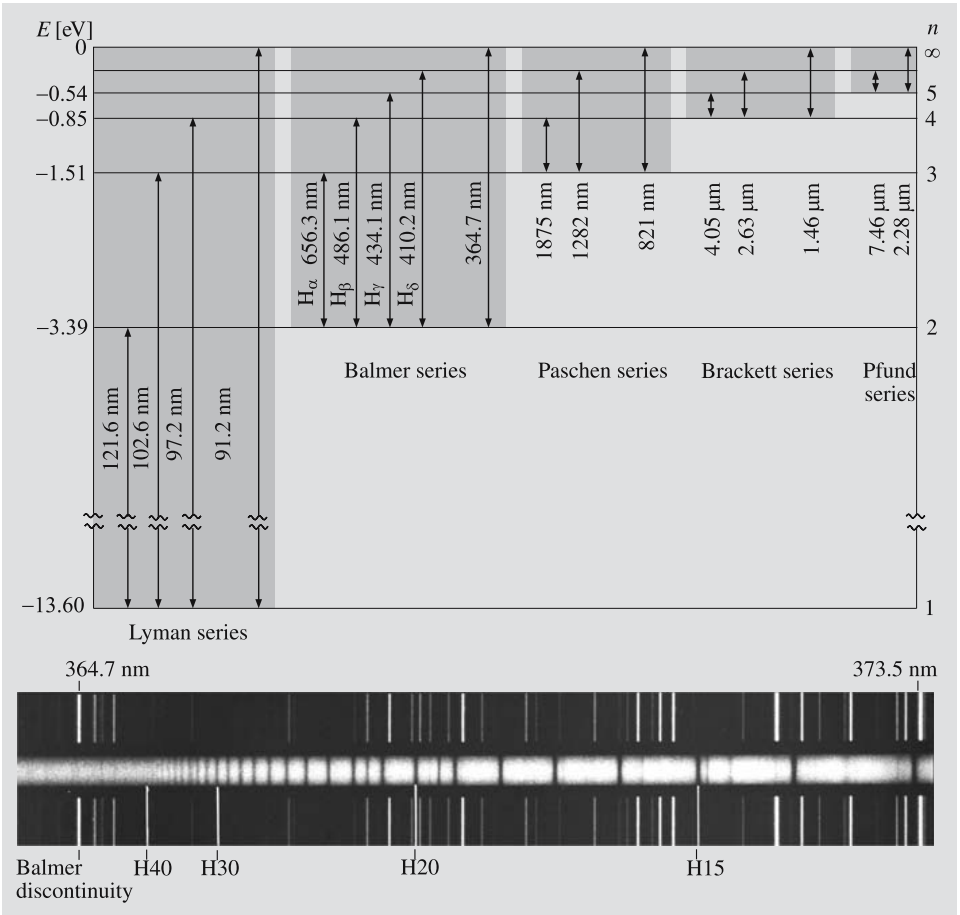


Fig. 5.4. Transitions of a hydrogen atom. The lower picture shows a part of the spectrum of the star HD193182. On both sides of the stellar spectrum we see an emission spectrum of iron. The wavelengths of the reference lines are known, and they can be used to calibrate the wavelengths of the observed stellar spectrum. The hydrogen Balmer lines are seen as dark absorption lines converging towards the Balmer ionization limit (also called the Balmer discontinuity) at $\lambda = 364.7$ nm to the left. The numbers $(15, \ldots, 40)$ refer to the quantum number n of the higher energy level. (Photo by Mt. Wilson Observatory)

series, which is in the ultraviolet. The other series with specific names are the *Paschen series* ($n_1 = 3$), *Bracket series* ($n_1 = 4$) and *Pfund series* ($n_1 = 5$) (see Fig. 5.4).

5.3 Line Profiles

The previous discussion suggests that spectral lines would be infinitely narrow and sharp. In reality, however, they are somewhat broadened. We will now consider briefly the factors affecting the shape of a spectral line, called a *line profile*. An exact treatment would take us too deep into quantum mechanics, so we cannot go into the details here.

According to quantum mechanics everything cannot be measured accurately at the same time. For example, even in principle, there is no way to determine the x coordinate and the momentum p_x in the direction of the x axis with arbitrary precision simultaneously. These quantities have small uncertainties Δx and Δp_x, such that

$$\Delta x \Delta p_x \approx \hbar \, .$$

A similar relation holds for other directions, too. Time and energy are also connected by an uncertainty relation,

$$\Delta E \Delta t \approx \hbar \, .$$

The natural width of spectral lines is a consequence of this *Heisenberg uncertainty principle*.

If the average lifetime of an excitation state is T, the energy corresponding to the transition can only be determined with an accuracy of $\Delta E = \hbar/T = h/(2\pi T)$. From (5.1) it follows that $\Delta \nu = \Delta E/h$. In fact, the uncertainty of the energy depends on the lifetimes of both the initial and final states. The *natural width* of a line is defined as

$$\gamma = \frac{\Delta E_\mathrm{i} + \Delta E_\mathrm{f}}{\hbar} = \frac{1}{T_\mathrm{i}} + \frac{1}{T_\mathrm{f}} \, . \tag{5.9}$$

It can be shown that the corresponding line profile is

$$I_\nu = \frac{\gamma}{2\pi} \frac{I_0}{(\nu - \nu_0)^2 + \gamma^2/4} \, , \tag{5.10}$$

where ν_0 is the frequency at the centre of the line and I_0 the total intensity of the line. At the centre of the line the intensity per frequency unit is

$$I_{\nu_0} = \frac{2}{\pi\gamma} I_0 \, ,$$

and at the frequency $\nu = \nu_0 + \gamma/2$,

$$I_{\nu_0 + \gamma/2} = \frac{1}{\pi\gamma} I_0 = \frac{1}{2} I_{\nu_0} \, .$$

Thus the width γ is the width of the line profile at a depth where the intensity is half of the maximum. This is called the *full width at half maximum* (FWHM).

Doppler Broadening. Atoms of a gas are moving the faster the higher the temperature of the gas. Thus spectral lines arising from individual atoms are shifted by the Doppler effect. The observed line consists of a collection of lines with different Doppler shifts, and the shape of the line depends on the number of atoms with different velocities.

Each Doppler shifted line has its characteristic natural width. The resulting line profile is obtained by giving each Doppler shifted line a weight proportional to the number of atoms given by the velocity distribution and integrating over all velocities. This gives rise to the *Voigt profile* (Fig. 5.5), which already describes most spectral lines quite well. The shapes of different profiles don't

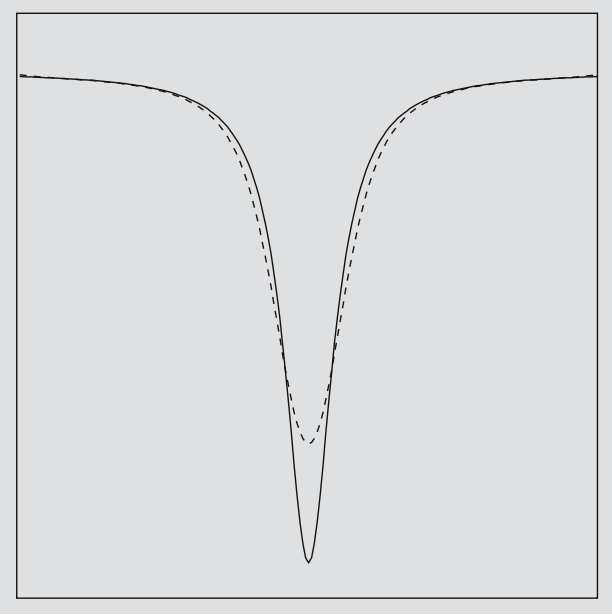


Fig. 5.5. Each spectral line has its characteristic natural width (*solid line*). Motions of particles broaden the line further due to the Doppler effect, resulting in the Voigt profile (*dashed line*). Both profiles have the same area

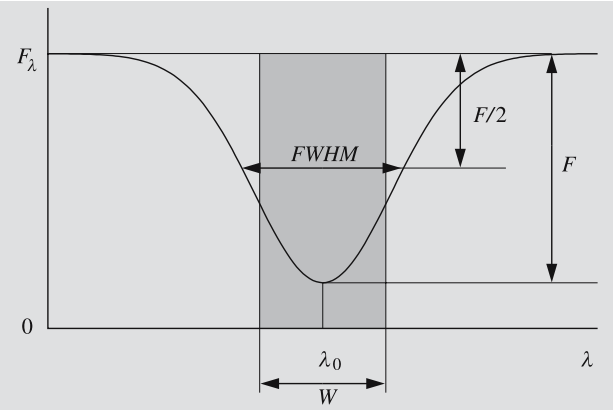


Fig. 5.6. The full width at half maximum (FWHM) of a spectral line is the width at the depth where the intensity is half of the maximum. The equivalent width W is defined so that the line and the shaded rectangle have the same area in the picture. The two measures are not generally the same, although they usually are close to each other

seem very different; the most obvious consequence of the broadening is that the maximum depth decreases.

One way to describe the width of a line is to give its full width at half maximum (Fig. 5.6). Due to Doppler broadening this is usually greater than the natural width. The *equivalent width* is another measure of a line strength. It is the area of a rectangular line that has the same area as the line profile and that emits no light at all. The equivalent width can be used to describe the energy corresponding to a line independently of the shape of the line profile.

5.4 Quantum Numbers, Selection Rules, Population Numbers

Quantum Numbers. The Bohr model needs only one quantum number, n, to describe all the energy levels of the electron. This can explain only the coarse features of an atom with a single electron.

Quantum mechanics describes the electron as a three dimensional wave, which only gives the probability of finding the electron in a certain place. Quantum mechanics has accurately predicted all the energy levels of hydrogen atoms. The energy levels of heavier atoms and molecules can also be computed; however, such calculations are very complicated. Also the existence of quantum numbers can be understood from the quantum mechanical point of view.

The quantum mechanical description involves four quantum numbers, one of which is our n, the *principal quantum number*. The principal quantum number describes the quantized energy levels of the electron. The classical interpretation of discrete energy levels allows only certain orbits given by (5.6). The orbital angular momentum of the electron is also quantized. This is described by the *angular momentum quantum number* l. The angular momentum corresponding to a quantum number l is

$$L = \sqrt{l(l+1)}\hbar \, .$$

The classical analogy would be to allow some elliptic orbits. The quantum number l can take only the values

$$l = 0, 1, \ldots , n-1 \, .$$

For historical reasons, these are often denoted by the letters s, p, d, f, g, h, i, j.

Although l determines the magnitude of the angular momentum, it does not give its direction. In a magnetic field this direction is important, since the orbiting electron also generates a tiny magnetic field. In any experiment, only one component of the angular momentum can be measured at a time. In a given direction z (e. g. in the direction of the applied magnetic field), the projection of the angular momentum can have only the values

$$L_z = m_l \hbar \, ,$$

where m_l is the *magnetic quantum number*

$$m_l = 0, \pm 1, \pm 2, \ldots , \pm l \, .$$

The magnetic quantum number is responsible for the splitting of spectral lines in strong magnetic fields, known as the *Zeeman effect*. For example, if $l = 1$, m_l can have $2l + 1 = 3$ different values. Thus, the line arising from the transition $l = 1 \rightarrow l = 0$ will split into three components in a magnetic field (Fig. 5.7).

The fourth quantum number is the *spin* describing the intrinsic angular momentum of the electron. The spin of the electron is

$$S = \sqrt{s(s+1)}\hbar \, ,$$

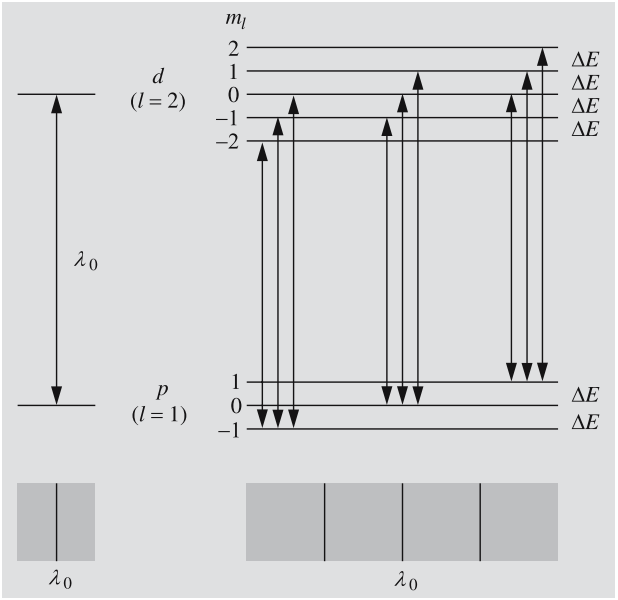


Fig. 5.7. The Zeeman effect. In strong magnetic fields each energy level of a hydrogen atom splits into $(2l+1)$ separate levels, which correspond to different values of the magnetic quantum number $m_l = l, l-1, \ldots, -l$. The energy differences of the successive levels have the same constant value ΔE. For example the p state ($l = 1$) splits into three and the d state ($l = 2$) into five sublevels. The selection rules require that in electric dipole transitions Δm_l equals 0 or ± 1, and only nine different transitions between p and d states are possible. Moreover, the transitions with the same Δm_l have the same energy difference. Thus the spectrum has only three separate lines

where the spin quantum number is $s = \frac{1}{2}$. In a given direction z, the spin is

$$S_z = m_s \hbar \,,$$

where m_s can have one of the two values:

$$m_s = \pm \frac{1}{2} \,.$$

All particles have a spin quantum number. Particles with an integral spin are called *bosons* (photon, mesons); particles with a half-integral spin are *fermions* (proton, neutron, electron, neutrino etc.).

Classically, spin can be interpreted as the rotation of a particle; this analogy, however, should not be taken too literally.

The total angular momentum $\boldsymbol{J}$ of an electron is the sum of its orbital and spin angular momentum:

$$\boldsymbol{J} = \boldsymbol{L} + \boldsymbol{S} \,.$$

Depending on the mutual orientation of the vectors $\boldsymbol{L}$ and $\boldsymbol{S}$ the quantum number j of total angular momentum can have one of two possible values,

$$j = l \pm \frac{1}{2} \,,$$

(except if $l = 0$, when $j = \frac{1}{2}$). The z component of the total angular momentum can have the values

$$m_j = 0, \pm 1, \pm 2, \ldots \pm j \,.$$

Spin also gives rise to the fine structure of spectral lines. Lines appear as close pairs or doublets.

Selection Rules. The state of an electron cannot change arbitrarily; transitions are restricted by selection rules, which follow from certain conservation laws. The selection rules express how the quantum numbers must change in a transition. Most probable are the *electric dipole transitions*, which make the atom behave like an oscillating dipole. The conservation laws require that in a transition we have

$$\begin{aligned} \Delta l &= \pm 1 \,, \\ \Delta m_l &= 0, \pm 1 \,. \end{aligned}$$

In terms of the total angular momentum the selection rules are

$$\begin{aligned} \Delta l &= \pm 1 \,, \\ \Delta j &= 0, \pm 1 \,, \\ \Delta m_j &= 0, \pm 1 \,. \end{aligned}$$

The probabilities of all other transitions are much smaller, and they are called *forbidden transitions*; examples are magnetic dipole transitions and all quadrupole and higher multipole transitions.

Spectral lines originating in forbidden transitions are called *forbidden lines*. The probability of such a transition is so low that under normal circumstances, the transition cannot take place before collisions force the electron to change state. Forbidden lines are possible only if the gas is extremely rarified (like in auroras and planetary nebulae).

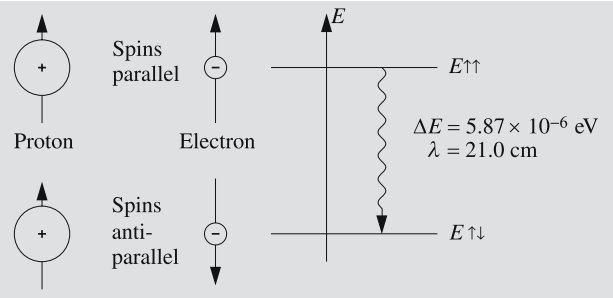


Fig. 5.8. The origin of the hydrogen 21 cm line. The spins of the electron and the proton may be either parallel or opposite. The energy of the former state is slightly larger. The wavelength of a photon corresponding to a transition between these states is 21 cm

The spins of an electron and nucleus of a hydrogen atom can be either parallel or antiparallel (Fig. 5.8). The energy of the former state is 0.0000059 eV higher. But the selection rules make an electric dipole transition between these states impossible. The transition, which is a magnetic dipole transition, has a very low probability, $A = 2.8 \times 10^{-15}\ \mathrm{s}^{-1}$. This means that the average lifetime of the higher state is $T = 1/A = 11 \times 10^6$ years. Usually collisions change the state of the electron well before this period of time has elapsed. But in interstellar space the density of hydrogen is so low and the total amount of hydrogen so great that a considerable number of these transitions can take place.

The wavelength of the radiation emitted by this transition is 21 cm, which is in the radio band of the spectrum. Extinction at radio wavelengths is very small, and we can observe more distant objects than by using optical wavelengths. The 21 cm radiation has been of crucial importance for surveys of interstellar hydrogen.

Population Numbers. The population number n_i of an energy state i means the number of atoms in that state per unit volume. In thermal equilibrium, the population numbers obey the *Boltzmann distribution*:

$$\frac{n_i}{n_0} = \frac{g_i}{g_0} \mathrm{e}^{-\Delta E/(kT)}\,, \tag{5.11}$$

where T is the temperature, k is the Boltzmann constant, $\Delta E = E_i - E_0 = h\nu$ is the energy difference between the excited and ground state, and g_i is the statistical weight of the level i (it is the number of different states with the same energy E_i). The subscript 0 always refers to the ground state. Often the population numbers differ from the values given by (5.11), but still we can define an *excitation temperature* T_{exc} in such a way that (5.11) gives correct population numbers, when T is replaced by T_{exc}. The excitation temperature may be different for different energy levels.

5.5 Molecular Spectra

The energy levels of an atom are determined by its electrons. In the case of a molecule, there are many more possibilities: atoms can vibrate around their equilibria and the molecule can rotate around some axis. Both vibrational and rotational states are quantized. Transitions between successive vibrational states typically involve photons in the infrared band, while transitions between rotational states involve photons in the microwave band. These combined with transitions of electrons produce a band spectrum, characteristic for molecules (Fig. 5.9). The spectrum has several narrow bands composed of a great number of lines.

5.6 Continuous Spectra

We have already mentioned some processes that produce continuous spectra. Continuous emission spectra can originate in recombinations and free–free transitions. In recombination, an atom captures a free

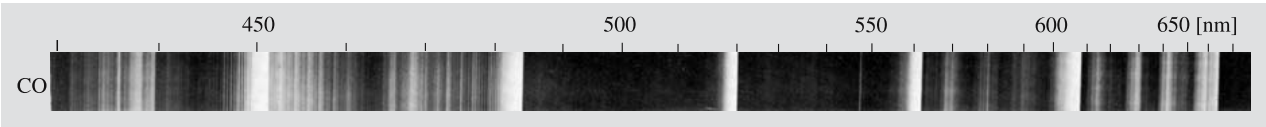


Fig. 5.9. Spectrum of carbon monoxide CO from 430 nm to 670 nm. The various bands correspond to different vibrational transitions. Each band is composed of numerous rotational lines. Near the right edge of each band the lines are so closely packed that they overlap and at this resolution, the spectrum looks continuous. (R.W.B. Pearse, A.G. Gaydon: *The Identification of Molecular Spectra* (Chapman & Hall Ltd., London 1976) p. 394)

electron whose energy is not quantized; in free–free transitions, both initial and final states are unquantized. Thus the emission line can have any frequency whatsoever. Similarly, ionizations and free–free transitions can give rise to a continuous absorption spectrum.

Each spectrum contains a continuous component, or *continuum*, and spectral lines. Sometimes, however, the lines are so closely packed and so broad that they seem to form a nearly continuous spectrum.

When the pressure of hot gas is increased, the spectral lines begin to broaden. At high pressure, atoms bump into each other more frequently, and the close neighbors disturb the energy levels. When the pressure is high enough, the lines begin to overlap. Thus the spectrum of hot gas at high pressure is continuous. Electric fields also broaden spectral lines (the Stark effect).

In liquids and solids the atoms are more densely packed than in gaseous substances. Their mutual perturbations broaden the energy levels, producing a continuous spectrum.

5.7 Blackbody Radiation

A *blackbody* is defined as an object that does not reflect or scatter radiation shining upon it, but absorbs and re-emits the radiation completely. A blackbody is a kind of an ideal radiator, which cannot exist in the real world. Yet many objects behave very much as if they were blackbodies.

The radiation of a blackbody depends only on its temperature, being perfectly independent of its shape, material and internal constitution. The wavelength distribution of the radiation follows *Planck's law*, which is a function of temperature only. The intensity at a frequency ν of a blackbody at temperature T is

$$B_\nu(T) = B(\nu; T) = \frac{2h\nu^3}{c^2}\frac{1}{\mathrm{e}^{h\nu/(kT)} - 1} , \qquad (5.12)$$

where

h = the Planck constant $= 6.63 \times 10^{-34}$ J s ,
c = the speed of light $\approx 3 \times 10^8$ m s^{-1} ,
k = the Boltzmann constant $= 1.38 \times 10^{-23}$ J K^{-1} .

By definition of the intensity, the dimension of B_ν is W m^{-2} Hz^{-1} sterad^{-1}.

Blackbody radiation can be produced in a closed cavity whose walls absorb all radiation incident upon them (and coming from inside the cavity). The walls and the radiation in the cavity are in equilibrium; both are at the same temperature, and the walls emit all the energy they receive. Since radiation energy is constantly transformed into thermal energy of the atoms of the walls and back to radiation, the blackbody radiation is also called *thermal radiation*.

The spectrum of a blackbody given by Planck's law (5.12) is continuous. This is true if the size of the radiator is very large compared with the dominant wavelengths. In the case of the cavity, this can be understood by considering the radiation as standing waves trapped in the cavity. The number of different wavelengths is larger, the shorter the wavelengths are compared with the size of the cavity. We already mentioned that spectra of solid bodies are continuous; very often such spectra can be quite well approximated by Planck's law.

We can also write Planck's law as a function of the wavelength. We require that $B_\nu \,\mathrm{d}\nu = -B_\lambda \,\mathrm{d}\lambda$. The wavelength decreases with increasing frequency; hence the minus sign. Since $\nu = c/\lambda$, we have

$$\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = -\frac{c}{\lambda^2} , \qquad (5.13)$$

whence

$$B_\lambda = -B_\nu\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = B_\nu\frac{c}{\lambda^2} , \qquad (5.14)$$

or

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5}\frac{1}{\mathrm{e}^{hc/(\lambda kT)} - 1} , \qquad (5.15)$$

$$[B_\lambda] = \mathrm{W\,m^{-2}\,m^{-1}\,sterad^{-1}} .$$

The functions B_ν and B_λ are defined in such a way that the total intensity can be obtained in the same way using either of them:

$$B(T) = \int_0^\infty B_\nu \,\mathrm{d}\nu = \int_0^\infty B_\lambda \,\mathrm{d}\lambda .$$

Let us now try to find the total intensity using the first of these integrals:

$$B(T) = \int_0^\infty B_\nu(T) \,\mathrm{d}\nu = \frac{2h}{c^2}\int_0^\infty \frac{\nu^3 \,\mathrm{d}\nu}{\mathrm{e}^{h\nu/(kT)} - 1} .$$

We now change the integration variable to $x = h\nu/(kT)$, whence $\mathrm{d}\nu = (kT/h)\mathrm{d}x$:

$$B(T) = \frac{2h}{c^2}\frac{k^4}{h^4}T^4\int_0^\infty \frac{x^3\,\mathrm{d}x}{\mathrm{e}^x - 1}\,.$$

The definite integral in this expression is just a real number, independent of the temperature. Thus we find that

$$B(T) = AT^4\,, \tag{5.16}$$

where the constant A has the value

$$A = \frac{2k^4}{c^2h^3}\frac{\pi^4}{15}\,. \tag{5.17}$$

(In order to get the value of A we have to evaluate the integral. There is no elementary way to do that. We can tell those who are familiar with all the exotic functions so beloved by theoretical physicists, that the integral can rather easily be expressed as $\Gamma(4)\zeta(4)$, where ζ is the Riemann zeta function and Γ is the gamma function. For integral values, $\Gamma(n)$ is simply the factorial $(n-1)!$. The difficult part is showing that $\zeta(4) = \pi^4/90$. This can be done by expanding $x^4 - x^2$ as a Fourier-series and evaluating the series at $x = \pi$.)

The flux density F for isotropic radiation of intensity B is (Sect. 4.1):

$$F = \pi B$$

or

$$F = \sigma T^4\,. \tag{5.18}$$

This is the *Stefan-Boltzmann law*, and the constant σ $(= \pi A)$ is the *Stefan-Boltzmann constant*,

$$\sigma = 5.67\times10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}\,.$$

From the Stefan-Boltzmann law we get a relation between the luminosity and temperature of a star. If the radius of the star is R, its surface area is $4\pi R^2$, and if the flux density on the surface is F, we have

$$L = 4\pi R^2 F\,.$$

If the star is assumed to radiate like a blackbody, we have $F = \sigma T^4$, which gives

$$L = 4\pi\sigma R^2 T^4\,. \tag{5.19}$$

In fact this defines the *effective temperature* of the star, discussed in more detail in the next section.

The luminosity, radius and temperature of a star are interdependent quantities, as we can see from (5.19). They are also related to the absolute bolometric magnitude of the star. Equation (4.13) gives the difference of the absolute bolometric magnitude of the star and the Sun:

$$M_{\mathrm{bol}} - M_{\mathrm{bol},\odot} = -2.5\lg\frac{L}{L_\odot}\,. \tag{5.20}$$

But we can now use (5.19) to express the luminosities in terms of the radii and temperatures:

$$M_{\mathrm{bol}} - M_{\mathrm{bol},\odot} = -5\lg\frac{R}{R_\odot} - 10\lg\frac{T}{T_\odot}\,. \tag{5.21}$$

As we can see in Fig. 5.10, the wavelength of the maximum intensity decreases with increasing total intensity (equal to the area below the curve). We can find the wavelength $\lambda_{\max}$ corresponding to the maximum intensity by differentiating Planck's function $B_\lambda(T)$ with

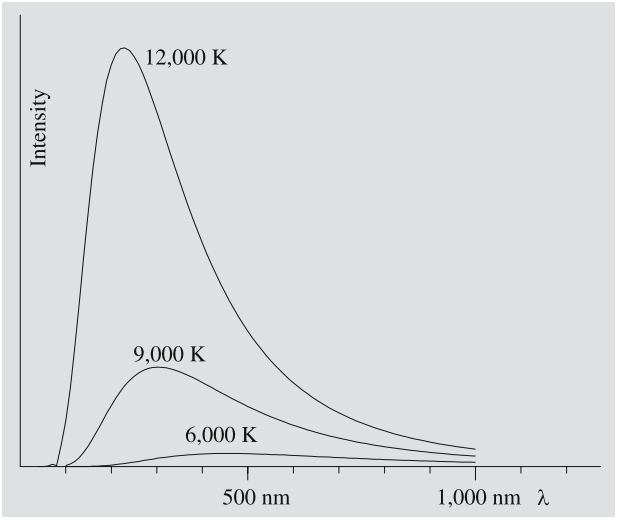


Fig. 5.10. Intensity distributions of blackbodies at temperature 12,000 K, 9000 K and 6000 K. Since the ratios of the temperatures are 4:3:2, the wavelengths of intensity maxima given by the Wien displacement law are in the proportions 1:4, 1:3 and 1:2, or 3, 4 and 6. The actual wavelengths of the maxima are 241.5 nm, 322 nm and 483 nm. The total intensities or the areas below the curves are proportional to 4^4, 3^4 and 2^4

respect to λ and finding zero of the derivative. The result is the *Wien displacement law*:

$$\lambda_{\max} T = b = \text{const} , \tag{5.22}$$

where the *Wien displacement constant* b is

$$b = 0.0028978 \text{ K m} .$$

We can use the same procedure to find the maximum of B_ν. But the frequency $\nu_{\max}$ thus obtained is different from $\nu_{\max} = c/\lambda_{\max}$ given by (5.22). The reason for this is the fact that the intensities are given per unit frequency or unit wavelength, and the dependence of frequency on wavelength is nonlinear.

When the wavelength is near the maximum or much longer than $\lambda_{\max}$ Planck's function can be approximated by simpler expressions. When $\lambda \approx \lambda_{\max}$ (or $hc/(\lambda kT) \gg 1$), we have

$$\mathrm{e}^{hc/(\lambda kT)} \gg 1 .$$

In this case we get the *Wien approximation*

$$B_\lambda(T) \approx \frac{2hc^2}{\lambda^5} \mathrm{e}^{-hc/(\lambda kT)} . \tag{5.23}$$

When $hc/(\lambda kT) \ll 1$ ($\lambda \gg \lambda_{\max}$), we have

$$\mathrm{e}^{hc/\lambda kT} \approx 1 + hc/(\lambda kT) ,$$

which gives the *Rayleigh–Jeans approximation*

$$B_\lambda(T) \approx \frac{2hc^2}{\lambda^5} \frac{\lambda kT}{hc} = \frac{2ckT}{\lambda^4} . \tag{5.24}$$

This is particularly useful in radio astronomy.

Classical physics predicted only the Rayleigh–Jeans approximation. Were (5.24) true for all wavelengths, the intensity would grow beyond all limits when the wavelength approaches zero, contrary to observations. This contradiction was known as the ultraviolet catastrophe.

5.8 Temperatures

Temperatures of astronomical objects range from almost absolute zero to millions of degrees. Temperature can be defined in a variety of ways, and its numerical value depends on the specific definition used. All these different temperatures are needed to describe different physical phenomena, and often there is no unique 'true' temperature.

Often the temperature is determined by comparing the object, a star for instance, with a blackbody. Although real stars do not radiate exactly like blackbodies, their spectra can usually be approximated by blackbody spectra after the effect of spectral lines has been eliminated. The resulting temperature depends on the exact criterion used to fit Planck's function to observations.

The most important quantity describing the surface temperature of a star is the *effective temperature* T_e. It is defined as the temperature of a blackbody which radiates with the same total flux density as the star. Since the effective temperature depends only on the total radiation power integrated over all frequencies, it is well defined for all energy distributions even if they deviate far from Planck's law.

In the previous section we derived the Stefan-Boltzmann law, which gives the total flux density as a function of the temperature. If we now find a value T_e of the temperature such that the Stefan-Boltzmann law gives the correct flux density F on the surface of the star, we have found the effective temperature. The flux density on the surface is

$$F = \sigma T_\mathrm{e}^4 . \tag{5.25}$$

The total flux is $L = 4\pi R^2 F$, where R is the radius of the star, and the flux density at a distance r is

$$F' = \frac{L}{4\pi r^2} = \frac{R^2}{r^2} F = \left(\frac{\alpha}{2}\right)^2 \sigma T_\mathrm{e}^4 , \tag{5.26}$$

where $\alpha = 2R/r$ is the observed angular diameter of the star. For direct determination of the effective temperature, we have to measure the total flux density and the angular diameter of the star. This is possible only in the few cases in which the diameter has been found by interferometry.

If we assume that at some wavelength λ the flux density F_λ on the surface of the star is obtained from Planck's law, we get the *brightness temperature* T_b. In the isotropic case we have then $F_\lambda = \pi B_\lambda(T_\mathrm{b})$. If the radius of the star is R and distance from the Earth r, the observed flux density is

$$F'_\lambda = \frac{R^2}{r^2} F_\lambda .$$

Again F_λ can be determined only if the angular diameter α is known. The brightness temperature T_b can then be solved from

$$F'_\lambda = \left(\frac{\alpha}{2}\right)^2 \pi B_\lambda(T_\mathrm{b}) \,. \tag{5.27}$$

Since the star does not radiate like a blackbody, its brightness temperature depends on the particular wavelength used in (5.27).

In radio astronomy, brightness temperature is used to express the intensity (or surface brightness) of the source. If the intensity at frequency ν is I_ν, the brightness temperature is obtained from

$$I_\nu = B_\nu(T_\mathrm{b}) \,.$$

T_b gives the temperature of a blackbody with the same surface brightness as the observed source.

Since radio wavelengths are very long, the condition $h\nu \ll kT$ of the Rayleigh–Jeans approximation is usually satisfied (except for millimetre and submillimetre bands), and we can write Planck's law as

$$\begin{aligned} B_\nu(T_\mathrm{b}) &= \frac{2h\nu^3}{c^2} \frac{1}{\mathrm{e}^{h\nu/(kT_\mathrm{b})} - 1} \\ &= \frac{2h\nu^3}{c^2} \frac{1}{1 + h\nu/(kT_\mathrm{b}) + \ldots - 1} \\ &\approx \frac{2k\nu^2}{c^2} T_\mathrm{b} \,. \end{aligned}$$

Thus we get the following expression for the radio astronomical brightness temperature:

$$T_\mathrm{b} = \frac{c^2}{2k\nu^2} I_\nu = \frac{\lambda^2}{2k} I_\nu \,. \tag{5.28}$$

A measure of the signal registered by a radio telescope is the *antenna temperature* T_A. After the antenna temperature is measured, we get the brightness temperature from

$$T_\mathrm{A} = \eta T_\mathrm{b} \,, \tag{5.29}$$

where η is the *beam efficiency* of the antenna (typically $0.4 \lesssim \eta \lesssim 0.8$). Equation (5.29) holds if the source is wide enough to cover the whole beam, i.e. the solid angle Ω_A from which the antenna receives radiation. If the solid angle subtended by the source, Ω_S, is smaller than Ω_A, the observed antenna temperature is

$$T_\mathrm{A} = \eta \frac{\Omega_\mathrm{S}}{\Omega_\mathrm{A}} T_\mathrm{b} \,, \quad (\Omega_\mathrm{S} < \Omega_\mathrm{A}) \,. \tag{5.30}$$

The *colour temperature* T_c can be determined even if the angular diameter of the source is unknown (Fig. 5.11). We only have to know the relative energy distribution in some wavelength range $[\lambda_1, \lambda_2]$; the absolute value of the flux is not needed. The observed flux density as a function of wavelength is compared with Planck's function at different temperatures. The temperature giving the best fit is the colour temperature in the interval $[\lambda_1, \lambda_2]$. The colour temperature is usually different for different wavelength intervals, since the shape of the observed energy distribution may be quite different from the blackbody spectrum.

A simple method for finding a colour temperature is the following. We measure the flux density F'_λ at two wavelengths λ_1 and λ_2. If we assume that the intensity distribution follows Planck's law, the ratio of these flux densities must be the same as the ratio obtained from Planck's law:

$$\frac{F'_{\lambda_1}(T)}{F'_{\lambda_2}(T)} = \frac{B_{\lambda_1}(T)}{B_{\lambda_2}(T)} = \frac{\lambda_2^5}{\lambda_1^5} \frac{\mathrm{e}^{hc/(\lambda_2 kT)} - 1}{\mathrm{e}^{hc/(\lambda_1 kT)} - 1} \,. \tag{5.31}$$

The temperature T solved from this equation is a colour temperature.

The observed flux densities correspond to certain magnitudes m_{λ_1} and m_{λ_2}. The definition of magnitudes gives

$$m_{\lambda_1} - m_{\lambda_2} = -2.5 \lg \frac{F'_{\lambda_1}}{F'_{\lambda_2}} + \mathrm{const} \,,$$

where the constant term is a consequence of the different zero points of the magnitude scales. If the temperature

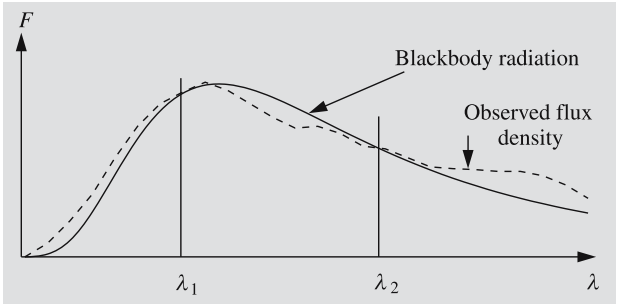


Fig. 5.11. Determination of the colour temperature. The ratio of the flux densities at wavelengths λ_1 and λ_2 gives the temperature of a blackbody with the same ratio. In general the result depends on the wavelengths chosen

is not too high, we can use the Wien approximation in the optical part of the spectrum:

$$m_{\lambda_1} - m_{\lambda_2} = -2.5 \lg \frac{B_{\lambda_1}}{B_{\lambda_2}} + \text{const}$$
$$= -2.5 \lg \left(\frac{\lambda_2}{\lambda_1}\right)^5 + 2.5 \frac{hc}{kT} \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \lg \mathrm{e} + \text{const} .$$

This can be written as

$$m_{\lambda_1} - m_{\lambda_2} = a + b/T_\mathrm{c} , \tag{5.32}$$

where a and b are constants. This shows that there is a simple relationship between the difference of two magnitudes and the colour temperature.

Strictly speaking, the magnitudes in (5.32) are monochromatic, but the same relation can be also used with broadband magnitudes like B and V. In that case, the two wavelengths are essentially the effective wavelengths of the B and V bands. The constant is chosen so that $B - V = 0$ for stars of the spectral type A0 (see Chap. 8). Thus the colour index $B - V$ also gives a colour temperature.

The *kinetic temperature* T_k, is related to the average speed of gas molecules. The kinetic energy of an ideal gas molecule as a function of temperature follows from the kinetic gas theory:

$$\text{Kinetic energy} = \frac{1}{2} m v^2 = \frac{3}{2} k T_\mathrm{k} .$$

Solving for T_k we get

$$T_\mathrm{k} = \frac{m v^2}{3k} , \tag{5.33}$$

where m is the mass of the molecule, v its average velocity (or rather its r. m. s velocity, which means that v^2 is the average of the squared velocities), and k, the Boltzmann constant. For ideal gases the pressure is directly proportional to the kinetic temperature (c. f. *Gas Pressure and Radiation Pressure, p. 238):

$$P = nkT_\mathrm{k} , \tag{5.34}$$

where n is the number density of the molecules (molecules per unit volume). We previously defined the excitation temperature T_exc as a temperature which, if substituted into the Boltzmann distribution (5.11), gives the observed population numbers. If the distribution of atoms in different levels is a result of mutual collisions of the atoms only, the excitation temperature equals the kinetic temperature, $T_\mathrm{exc} = T_\mathrm{k}$.

The *ionization temperature* T_i is found by comparing the number of atoms in different states of ionization. Since stars are not exactly blackbodies, the values of excitation and ionization temperatures usually vary, depending on the element whose spectral lines were used for temperature determination.

In *thermodynamic equilibrium* all these various temperatures are equal.

5.9 Other Radiation Mechanisms

The radiation of a gas in thermodynamic equilibrium depends on the temperature and density only. In astrophysical objects deviations from thermodynamic equilibrium are, however, quite common. Some examples of *non-thermal radiation* arising under such conditions are mentioned in the following.

Maser and Laser (Fig. 5.12). The Boltzmann distribution (5.11) shows that usually there are fewer atoms in excited states than in the ground state. There are, however, means to produce a *population inversion*, an excited state containing more atoms than the ground state. This inversion is essential for both the *maser* and the *laser* (Microwave/Light Amplification by Stimulated Emission of Radiation). If the excited atoms are now illuminated with photons having energies equal to the excitation energy, the radiation will induce downward transitions. The number of photons emitted greatly exceeds the number of absorbed photons, and radiation is amplified. Typically the excited state is a *metastable state*, a state with a very long average lifetime, which means that the contribution of spontaneous emission is negligible. Therefore the resulting radiation is coherent and monochromatic. Several maser sources have been found in interstellar molecular clouds and dust envelopes around stars.

Synchrotron Radiation. A free charge in accelerated motion will emit electromagnetic radiation. Charged particles moving in a magnetic field follow helices around the field lines. As seen from the direction of

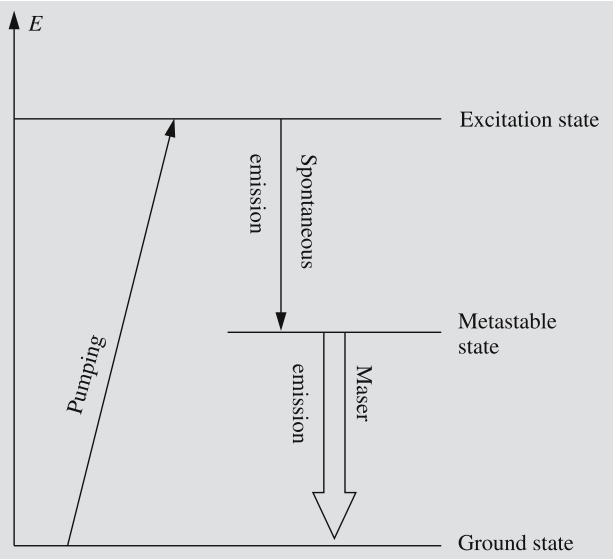


Fig. 5.12. The operational principle of the maser and the laser. A metastable state (a state with a relatively long average lifetime) stores atoms where they accumulate; there are more atoms in the metastable state than in the ground state. This population inversion is maintained by radiatively exciting atoms to a higher excitation state ("pumping"), from which they spontaneously jump down to the metastable state. When the atoms are illuminated by photons with energies equal to the excitation energy of the metastable state, the photons will induce more radiation of the same wavelength, and the radiation is amplified in geometric progression

the field, the motion is circular and therefore accelerated. The moving charge will radiate in the direction of its velocity vector. Such radiation is called *synchrotron radiation*. It will be further discussed in Chap. 15.

5.10 Radiative Transfer

Propagation of radiation in a medium, also called radiative transfer, is one of the basic problems of astrophysics. The subject is too complicated to be discussed here in any detail. The fundamental equation of radiative transfer is, however, easily derived.

Assume we have a small cylinder, the bottom of which has an area $\mathrm{d}A$ and the length of which is $\mathrm{d}r$. Let I_ν be the intensity of radiation perpendicular to the bottom surface going into a solid angle $\mathrm{d}\omega$ ($[I_\nu] = \mathrm{W\,m^{-2}\,Hz^{-1}\,sterad^{-1}}$). If the intensity changes by an amount $\mathrm{d}I_\nu$ in the distance $\mathrm{d}r$, the energy changes by

$$\mathrm{d}E = \mathrm{d}I_\nu\,\mathrm{d}A\,\mathrm{d}\nu\,\mathrm{d}\omega\,\mathrm{d}t$$

in the cylinder in time $\mathrm{d}t$. This equals the emission minus absorption in the cylinder. The absorbed energy is (c. f. (4.14))

$$\mathrm{d}E_{\mathrm{abs}} = \alpha_\nu I_\nu\,\mathrm{d}r\,\mathrm{d}A\,\mathrm{d}\nu\,\mathrm{d}\omega\,\mathrm{d}t\ , \tag{5.35}$$

where α_ν is the opacity of the medium at frequency ν. Let the amount of energy emitted per hertz at frequency ν into unit solid angle from unit volume and per unit time be j_ν ($[j_\nu] = \mathrm{W\,m^{-3}\,Hz^{-1}\,sterad^{-1}}$). This is called the *emission coefficient* of the medium. The energy emitted into solid angle $\mathrm{d}\omega$ from the cylinder is then

$$\mathrm{d}E_{\mathrm{em}} = j_\nu\,\mathrm{d}r\,\mathrm{d}A\,\mathrm{d}\nu\,\mathrm{d}\omega\,\mathrm{d}t\ . \tag{5.36}$$

The equation

$$\mathrm{d}E = -\mathrm{d}E_{\mathrm{abs}} + \mathrm{d}E_{\mathrm{em}}$$

gives then

$$\mathrm{d}I_\nu = -\alpha_\nu I_\nu \mathrm{d}r + j_\nu \mathrm{d}r$$

or

$$\frac{\mathrm{d}I_\nu}{\alpha_\nu\,\mathrm{d}r} = -I_\nu + \frac{j_\nu}{\alpha_\nu}\ . \tag{5.37}$$

We shall denote the ratio of the emission coefficient j_ν to the absorption coefficient or opacity α_ν by S_ν:

$$S_\nu = \frac{j_\nu}{\alpha_\nu}\ . \tag{5.38}$$

S_ν is called the *source function*. Because $\alpha_\nu\,\mathrm{d}r = \mathrm{d}\tau_\nu$, where τ_ν is the optical thickness at frequency ν, (5.37) can be written as

$$\frac{\mathrm{d}I_\nu}{\mathrm{d}\tau_\nu} = -I_\nu + S_\nu\ . \tag{5.39}$$

Equation (5.39) is the basic equation of radiative transfer. Without solving the equation, we see that if $I_\nu < S_\nu$, then $\mathrm{d}I_\nu/\mathrm{d}\tau_\nu > 0$, and the intensity tends to increase in the direction of propagation. And, if $I_\nu > S_\nu$, then $\mathrm{d}I_\nu/\mathrm{d}\tau_\nu < 0$, and I_ν will decrease. In an equilibrium the emitted and absorbed energies are equal, in which case we find from (5.35) and (5.36)

$$I_\nu = j_\nu/\alpha_\nu = S_\nu\ . \tag{5.40}$$

Substituting this into (5.39), we see that $\mathrm{d}I_\nu/\mathrm{d}\tau_\nu = 0$. In thermodynamic equilibrium the radiation of the medium is blackbody radiation, and the source function is given by Planck's law:

$$S_\nu = B_\nu(T) = \frac{2h\nu^3}{c^2}\frac{1}{\mathrm{e}^{h\nu/(kT)}-1} .$$

Even if the system is not in thermodynamic equilibrium, it may be possible to find an excitation temperature T_{exc} such that $B_\nu(T_{\mathrm{exc}}) = S_\nu$. This temperature may depend on frequency.

A formal solution of (5.39) is

$$I_\nu(\tau_\nu) = I_\nu(0)\,\mathrm{e}^{-\tau_\nu} + \int_0^{\tau_\nu} \mathrm{e}^{-(\tau_\nu - t)} S_\nu(t)\,\mathrm{d}t . \qquad (5.41)$$

Here $I_\nu(0)$ is the intensity of the background radiation, coming through the medium (e. g. an interstellar cloud) and decaying exponentially in the medium. The second term gives the emission in the medium. The solution is only formal, since in general, the source function S_ν is unknown and must be solved simultaneously with the intensity. If $S_\nu(\tau_\nu)$ is constant in the cloud and the background radiation is ignored, we get

$$I_\nu(\tau_\nu) = S_\nu \int_0^{\tau_\nu} \mathrm{e}^{-(\tau_\nu - t)}\mathrm{d}t = S_\nu(1-\mathrm{e}^{-\tau_\nu}) . \qquad (5.42)$$

If the cloud is optically thick ($\tau_\nu \gg 1$), we have

$$I_\nu = S_\nu , \qquad (5.43)$$

i. e. the intensity equals the source function, and the emission and absorption processes are in equilibrium.

An important field of application of the theory of radiative transfer is in the study of planetary and stellar atmospheres. In this case, to a good approximation, the properties of the medium only vary in one direction, say along the z axis. The intensity will then depend only on z and θ, where θ is the angle between the z axis and the direction of propagation of the radiation.

In applications to atmospheres it is customary to define the optical depth τ_ν in the vertical direction as

$$\mathrm{d}\tau_\nu = -\alpha_\nu\,\mathrm{d}z .$$

Conventionally z increases upwards and the optical depth inwards in the atmosphere. The vertical line element $\mathrm{d}z$ is related to that along the light ray, $\mathrm{d}r$, according to

$$\mathrm{d}z = \mathrm{d}r\cos\theta .$$

With these notational conventions, (5.39) now yields

$$\cos\theta\frac{\mathrm{d}I_\nu(z,\theta)}{\mathrm{d}\tau_\nu} = I_\nu - S_\nu . \qquad (5.44)$$

This is the form of the equation of radiative transfer usually encountered in the study of stellar and planetary atmospheres.

A formal expression for the intensity emerging from an atmosphere can be obtained by integrating (5.44) from $\tau_\nu = \infty$ (we assume that the bottom of the atmosphere is at infinite optical depth) to $\tau_\nu = 0$ (corresponding to the top of the atmosphere). This yields

$$I_\nu(0,\theta) = \int_0^{\infty} S_\nu\,\mathrm{e}^{-\tau_\nu\sec\theta}\sec\theta\,\mathrm{d}\tau_\nu . \qquad (5.45)$$

This expression will be used later in Chap. 8 on the interpretation of stellar spectra.

5.11 Examples

Example 5.1 Find the wavelength of the photon emitted in the transition of a hydrogen atom from $n_2 = 110$ to $n_1 = 109$.

Equation (5.8) gives

$$\begin{aligned}\frac{1}{\lambda} &= R\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right)\\ &= 1.097\times10^7\ \mathrm{m}^{-1}\left(\frac{1}{109^2}-\frac{1}{110^2}\right)\\ &= 16.71\ \mathrm{m}^{-1} ,\end{aligned}$$

whence

$$\lambda = 0.060\ \mathrm{m} .$$

This is in the radio band. Such radiation was observed for the first time in 1965 by an NRAO radio telescope.

Example 5.2 The effective temperature of a star is 12,000 K and the absolute bolometric magnitude 0.0. Find the radius of the star, when the effective temperature of the Sun is 5000 K and the absolute bolometric magnitude 4.7.

We can apply (5.21):

$$M_{\rm bol} - M_{\rm bol,\odot} = -5\lg\frac{R}{R_\odot} - 10\lg\frac{T}{T_\odot}$$

$$\Rightarrow \frac{R}{R_\odot} = \left(\frac{T_{\rm e\odot}}{T_{\rm e}}\right)^2 10^{-0.2(-M_{\rm bol,\odot})}$$

$$= \left(\frac{5800}{12000}\right)^2 10^{-0.2(0.0-4.7)}$$

$$= 2.0\,.$$

Thus the radius is twice the Solar radius.

Example 5.3 Derive the Wien displacement laws.

Let us denote $x = hc/(\lambda kT)$. Planck's law then becomes

$$B_\lambda(T) = \frac{2k^5T^5}{h^4c^3}\,\frac{x^5}{{\rm e}^x-1}\,.$$

For a given temperature, the first factor is constant. Thus, it is sufficient to find the maximum of the function $f(x) = x^5/({\rm e}^x-1)$.

First we must evaluate the derivative of f:

$$f'(x) = \frac{5x^4({\rm e}^x-1) - x^5{\rm e}^x}{({\rm e}^x-1)^2}$$

$$= \frac{x^4{\rm e}^x}{({\rm e}^x-1)^2}\,(5-5{\rm e}^{-x}-x)\,.$$

By definition, x is always strictly positive. Hence $f'(x)$ can be zero only if the factor $5-5{\rm e}^{-x}-x$ is zero. This equation cannot be solved analytically. Instead we write the equation as $x = 5-5{\rm e}^{-x}$ and solve it by iteration:

$$x_0 = 5\,, \quad \text{(this is just a guess)}$$

$$x_1 = 5-5{\rm e}^{-x_0} = 4.96631\,,$$

$$\vdots$$

$$x_5 = 4.96511\,.$$

Thus the result is $x = 4.965$. The Wien displacement law is then

$$\lambda_{\rm max}T = \frac{hc}{xk} = b = 2.898\times10^{-3}\,{\rm K\,m}\,.$$

In terms of frequency Planck's law is

$$B_\nu(T) = \frac{2h\nu^3}{c^2}\,\frac{1}{{\rm e}^{h\nu/(kT)}-1}\,.$$

Substituting $x = h\nu/(kT)$ we get

$$B_\nu(T) = \frac{2k^3T^3}{h^2c^2}\,\frac{x^3}{{\rm e}^x-1}\,.$$

Now we study the function $f(x) = x^3/({\rm e}^x-1)$:

$$f'(x) = \frac{3x^2({\rm e}^x-1) - x^3{\rm e}^x}{({\rm e}^x-1)^2}$$

$$= \frac{x^2{\rm e}^x}{({\rm e}^x-1)^2}\,(3-3{\rm e}^{-x}-x)\,.$$

This vanishes, when $3-3{\rm e}^{-x}-x = 0$. The solution of this equation is $x = 2.821$. Hence

$$\frac{cT}{\nu_{\rm max}} = \frac{hc}{kx} = b' = 5.100\times10^{-3}\,{\rm K\,m}$$

or

$$\frac{T}{\nu_{\rm max}} = 1.701\times10^{-11}\,{\rm K\,s}\,.$$

Note that the wavelength corresponding to $\nu_{\rm max}$ is different from $\lambda_{\rm max}$. The reason is that we have used two different forms of Planck's function, one giving the intensity per unit wavelength, the other per unit frequency.

Example 5.4 a) Find the fraction of radiation that a blackbody emits in the range $[\lambda_1, \lambda_2]$, where λ_1 and $\lambda_2 \gg \lambda_{\rm max}$. b) How much energy does a 100 W incandescent light bulb radiate in the radio wavelengths, $\lambda \geq 1$ cm? Assume the temperature is 2500 K.

Since the wavelengths are much longer than $\lambda_{\rm max}$ we can use the Rayleigh–Jeans approximation $B_\lambda(T) \approx 2ckT/\lambda^4$. Then

$$B' = \int_{\lambda_1}^{\lambda_2} B_\lambda(T)\,{\rm d}\lambda \approx 2ckT\int_{\lambda_1}^{\lambda_2}\frac{d\lambda}{\lambda^4}$$

$$= \frac{2ckT}{3}\left(\frac{1}{\lambda_1^3}-\frac{1}{\lambda_2^3}\right)\,,$$

and hence

$$\frac{B'}{B_{\rm tot}} = \frac{5c^3h^3}{k^3\pi^4}\,\frac{1}{T^3}\left(\frac{1}{\lambda_1^3}-\frac{1}{\lambda_2^3}\right)\,.$$

Now the temperature is $T = 2500$ K and the wavelength range [0.01 m, ∞), and so

$$B' = 100\,{\rm W}\times1.529\times10^{-7}\,\frac{1}{2500^3}\,\frac{1}{0.01^3}$$

$$= 9.8\times10^{-10}\,{\rm W}\,.$$

It is quite difficult to listen to the radio emission of a light bulb with an ordinary radio receiver.

Example 5.5 *Determination of Effective Temperature*

The observed flux density of Arcturus is

$$F' = 4.5 \times 10^{-8}\ \mathrm{W\,m^{-2}}\ .$$

Interferometric measurements give an angular diameter of $\alpha = 0.020''$. Thus, $\alpha/2 = 4.85 \times 10^{-8}$ radians. From (5.26) we get

$$T_\mathrm{e} = \left(\frac{4.5 \times 10^{-8}}{(4.85 \times 10^{-8})^2 \times 5.669 \times 10^{-8}} \right)^{1/4} \mathrm{K}$$
$$= 4300\ \mathrm{K}\ .$$

Example 5.6 Flux densities at the wavelengths 440 nm and 550 nm are 1.30 and $1.00\ \mathrm{W\,m^{-2}\,m^{-1}}$, respectively. Find the colour temperature.

If the flux densities at the wavelengths λ_1 and λ_2 are F_1 and F_2, respectively, the colour temperature can be solved from the equation

$$\frac{F_1}{F_2} = \frac{B_{\lambda_1}(T_\mathrm{c})}{B_{\lambda_2}(T_\mathrm{c})} = \left(\frac{\lambda_2}{\lambda_1}\right)^5 \frac{\mathrm{e}^{hc/(\lambda_2 k T_\mathrm{c})} - 1}{\mathrm{e}^{hc/(\lambda_1 k T_\mathrm{c})} - 1}\ .$$

If we denote

$$A = \frac{F_1}{F_2}\left(\frac{\lambda_1}{\lambda_2}\right)^5\ ,$$

$$B_1 = \frac{hc}{\lambda_1 k}\ ,$$

$$B_2 = \frac{hc}{\lambda_2 k}\ ,$$

we get the equation

$$A = \frac{\mathrm{e}^{B_2/T_\mathrm{c}} - 1}{\mathrm{e}^{B_1/T_\mathrm{c}} - 1}$$

for the colour temperature T_c. This equation must be solved numerically.

In our example the constants have the following values:

$$A = \frac{1.00}{1.30}\left(\frac{550}{440}\right)^5 = 2.348\ ,$$
$$B_1 = 32{,}700\ \mathrm{K}\ , \quad B_2 = 26{,}160\ \mathrm{K}\ .$$

By substituting different values for T_c, we find that $T_\mathrm{c} = 7545$ K satisfies our equation.

5.12 Exercises

Exercise 5.1 Show that in the Wien approximation the relative error of B_λ is

$$\frac{\Delta B_\lambda}{B_\lambda} = -\mathrm{e}^{-hc/(\lambda k T)}\ .$$

Exercise 5.2 If the transition of the hydrogen atom $n+1 \to n$ were to correspond to the wavelength 21.05 cm, what would the quantum number n be? The interstellar medium emits strong radiation at this wavelength. Can this radiation be due to such transitions?

Exercise 5.3 The space is filled with background radiation, remnant of the early age of the universe. Currently the distribution of this radiation is similar to the radiation of a blackbody at the temperature of 2.7 K. What is λ_max corresponding to this radiation? What is its total intensity? Compare the intensity of the background radiation to the intensity of the Sun at the visual wavelengths.

Exercise 5.4 The temperature of a red giant is $T = 2500$ K and radius 100 times the solar radius.

a) Find the total luminosity of the star, and the luminosity in the visual band $400\ \mathrm{nm} \le \lambda \le 700\ \mathrm{nm}$.
b) Compare the star with a 100 W lamp that radiates 5% of its energy in the visual band. What is the distance of the lamp if it looks as bright as the star?

Exercise 5.5 The effective temperature of Sirius is 10,000 K, apparent visual magnitude -1.5, distance 2.67 kpc and bolometric correction 0.5. What is the radius of Sirius?

Exercise 5.6 The observed flux density of the Sun at $\lambda = 300$ nm is $0.59\ \mathrm{W\,m^{-2}\,nm^{-1}}$. Find the brightness temperature of the Sun at this wavelength.

Exercise 5.7 The colour temperature can be determined from two magnitudes corresponding to two

different wavelengths. Show that

$$T_{\mathrm{c}} = \frac{7000\ \mathrm{K}}{(B-V)+0.47}\ .$$

The wavelengths of the B and V bands are 440 nm and 548 nm, respectively, and we assume that $B = V$ for stars of the spectral class A0, the colour temperature of which is about 15,000 K.

Exercise 5.8 The kinetic temperature of the plasma in the solar corona can reach 10^6 K. Find the average speed of the electrons in such a plasma.

6. Celestial Mechanics

Celestial mechanics, the study of motions of celestial bodies, together with spherical astronomy, was the main branch of astronomy until the end of the 19th century, when astrophysics began to evolve rapidly. The primary task of classical celestial mechanics was to explain and predict the motions of planets and their satellites. Several empirical models, like epicycles and Kepler's laws, were employed to describe these motions. But none of these models explained why the planets moved the way they did. It was only in the 1680's that a simple explanation was found for all these motions – Newton's law of universal gravitation. In this chapter, we will derive some properties of orbital motion. The physics we need for this is simple indeed, just Newton's laws. (For a review, see *Newton's Laws, p. 126)

This chapter is mathematically slightly more involved than the rest of the book. We shall use some vector calculus to derive our results, which, however, can be easily understood with very elementary mathematics. A summary of the basic facts of vector calculus is given in Appendix A.4.

6.1 Equations of Motion

We shall concentrate on the systems of only two bodies. In fact, this is the most complicated case that allows a neat analytical solution. For simplicity, let us call the bodies the Sun and a planet, although they could quite as well be a planet and its moon, or the two components of a binary star.

Let the masses of the two bodies be m_1 and m_2 and the radius vectors in some fixed inertial coordinate frame $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ (Fig. 6.1). The position of the planet relative to the Sun is denoted by $\boldsymbol{r} = \boldsymbol{r}_2 - \boldsymbol{r}_1$. According to Newton's law of gravitation the planet feels a gravitational pull proportional to the masses m_1 and m_2 and inversely proportional to the square of the distance r. Since the force is directed towards the Sun, it can be expressed as

$$\boldsymbol{F} = \frac{Gm_1m_2}{r^2}\frac{-\boldsymbol{r}}{r} = -Gm_1m_2\frac{\boldsymbol{r}}{r^3} , \tag{6.1}$$

where G is the *gravitational constant*. (More about this in Sect. 6.5.)

Newton's second law tells us that the acceleration $\ddot{\boldsymbol{r}}_2$ of the planet is proportional to the applied force:

$$\boldsymbol{F} = m_2\ddot{\boldsymbol{r}}_2 . \tag{6.2}$$

Combining (6.1) and (6.2), we get the *equation of motion* of the planet

$$m_2\ddot{\boldsymbol{r}}_2 = -Gm_1m_2\frac{\boldsymbol{r}}{r^3} . \tag{6.3}$$

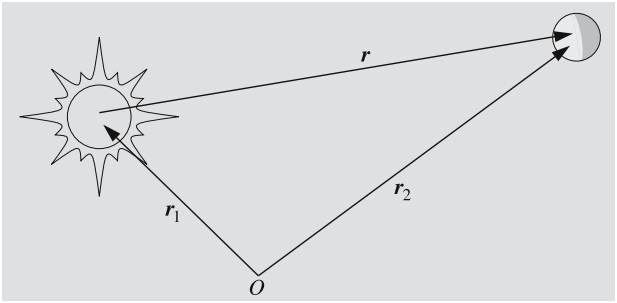


Fig. 6.1. The radius vectors of the Sun and a planet in an arbitrary inertial frame are $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, and $\boldsymbol{r} = \boldsymbol{r}_2 - \boldsymbol{r}_1$ is the position of the planet relative to the Sun

Since the Sun feels the same gravitational pull, but in the opposite direction, we can immediately write the equation of motion of the Sun:

$$m_1\ddot{\boldsymbol{r}}_1 = +Gm_1m_2\frac{\boldsymbol{r}}{r^3} . \tag{6.4}$$

We are mainly interested in the relative motion of the planet with respect to the Sun. To find the equation of the relative orbit, we cancel the masses appearing on both sides of (6.3) and (6.4), and subtract (6.4) from (6.3) to get

$$\ddot{\boldsymbol{r}} = -\mu\frac{\boldsymbol{r}}{r^3} , \tag{6.5}$$

where we have denoted

$$\mu = G(m_1 + m_2) . \tag{6.6}$$

Hannu Karttunen et al. (Eds.), *Celestial Mechanics*.
In: Hannu Karttunen et al. (Eds.), Fundamental Astronomy, 5th Edition. pp. 113–130 (2007)
DOI: 11685739_6 © Springer-Verlag Berlin Heidelberg 2007

The solution of (6.5) now gives the relative orbit of the planet. The equation involves the radius vector and its second time derivative. In principle, the solution should yield the radius vector as a function of time, $\boldsymbol{r} = \boldsymbol{r}(t)$. Unfortunately things are not this simple in practice; in fact, there is no way to express the radius vector as a function of time in a closed form (i.e. as a finite expression involving familiar elementary functions). Although there are several ways to solve the equation of motion, we must resort to mathematical manipulation in one form or another to figure out the essential properties of the orbit. Next we shall study one possible method.

6.2 Solution of the Equation of Motion

The equation of motion (6.5) is a second-order (i.e. contains second derivatives) vector valued differential equation. Therefore we need six integration constants or *integrals* for the complete solution. The solution is an infinite family of orbits with different sizes, shapes and orientations. A particular solution (e.g. the orbit of Jupiter) is selected by fixing the values of the six integrals. The fate of a planet is unambiguously determined by its position and velocity at any given moment; thus we could take the position and velocity vectors at some moment as our integrals. Although they do not tell us anything about the geometry of the orbit, they can be used as initial values when integrating the orbit numerically with a computer. Another set of integrals, the *orbital elements*, contains geometric quantities describing the orbit in a very clear and concrete way. We shall return to these later. A third possible set involves certain physical quantities, which we shall derive next.

We begin by showing that the angular momentum remains constant. The angular momentum of the planet in the heliocentric frame is

$$\boldsymbol{L} = m_2 \boldsymbol{r} \times \dot{\boldsymbol{r}} \,. \tag{6.7}$$

Celestial mechanicians usually prefer to use the angular momentum divided by the planet's mass

$$\boldsymbol{k} = \boldsymbol{r} \times \dot{\boldsymbol{r}} \,. \tag{6.8}$$

Let us find the time derivative of this:

$$\dot{\boldsymbol{k}} = \boldsymbol{r} \times \ddot{\boldsymbol{r}} + \dot{\boldsymbol{r}} \times \dot{\boldsymbol{r}} \,.$$

The latter term vanishes as a vector product of two parallel vectors. The former term contains $\ddot{\boldsymbol{r}}$, which is given by the equation of motion:

$$\dot{\boldsymbol{k}} = \boldsymbol{r} \times (-\mu \boldsymbol{r}/r^3) = -(\mu/r^3)\boldsymbol{r} \times \boldsymbol{r} = 0 \,.$$

Thus $\boldsymbol{k}$ is a constant vector independent of time (as is $\boldsymbol{L}$, of course).

Since the angular momentum vector is always perpendicular to the motion (this follows from (6.8)), the motion is at all times restricted to the invariable plane perpendicular to $\boldsymbol{k}$ (Fig. 6.2).

To find another constant vector, we compute the vector product $\boldsymbol{k} \times \ddot{\boldsymbol{r}}$:

$$\begin{aligned} \boldsymbol{k} \times \ddot{\boldsymbol{r}} &= (\boldsymbol{r} \times \dot{\boldsymbol{r}}) \times (-\mu \boldsymbol{r}/r^3) \\ &= -\frac{\mu}{r^3}[(\boldsymbol{r} \cdot \boldsymbol{r})\dot{\boldsymbol{r}} - (\boldsymbol{r} \cdot \dot{\boldsymbol{r}})\boldsymbol{r}] \,. \end{aligned}$$

The time derivative of the distance r is equal to the projection of $\dot{\boldsymbol{r}}$ in the direction of $\boldsymbol{r}$ (Fig. 6.3); thus, using the properties of the scalar product, we get $\dot{r} = \boldsymbol{r} \cdot \dot{\boldsymbol{r}}/r$, which gives

$$\boldsymbol{r} \cdot \dot{\boldsymbol{r}} = r\dot{r} \,. \tag{6.9}$$

Hence,

$$\boldsymbol{k} \times \ddot{\boldsymbol{r}} = -\mu(\dot{\boldsymbol{r}}/r - \boldsymbol{r}\dot{r}/r^2) = \frac{\mathrm{d}}{\mathrm{d}t}(-\mu \boldsymbol{r}/r) \,.$$

The vector product can also be expressed as

$$\boldsymbol{k} \times \ddot{\boldsymbol{r}} = \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{k} \times \dot{\boldsymbol{r}}) \,,$$

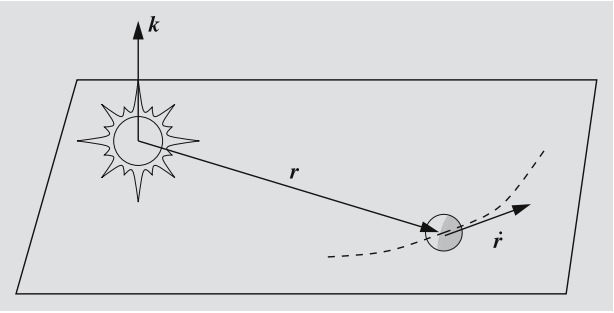


Fig. 6.2. The angular momentum vector $\boldsymbol{k}$ is perpendicular to the radius and velocity vectors of the planet. Since $\boldsymbol{k}$ is a constant vector, the motion of the planet is restricted to the plane perpendicular to $\boldsymbol{k}$

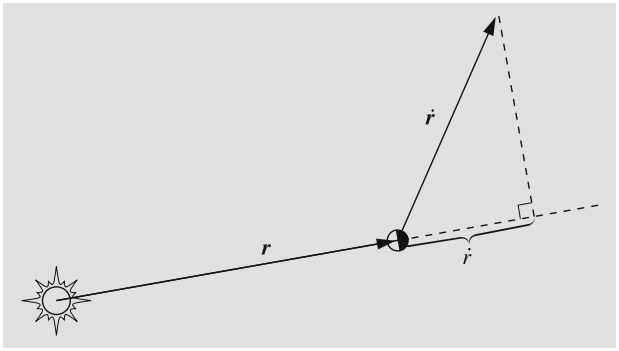


Fig. 6.3. The radial velocity $\dot{r}$ is the projection of the velocity vector $\dot{\boldsymbol{r}}$ in the direction of the radius vector $\boldsymbol{r}$

since $\boldsymbol{k}$ is a constant vector. Combining this with the previous equation, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{k}\times\dot{\boldsymbol{r}}+\mu\boldsymbol{r}/r)=0$$

and

$$\boldsymbol{k}\times\dot{\boldsymbol{r}}+\mu\boldsymbol{r}/r=\text{const}=-\mu\boldsymbol{e}\,. \tag{6.10}$$

Since $\boldsymbol{k}$ is perpendicular to the orbital plane, $\boldsymbol{k}\times\dot{\boldsymbol{r}}$ must lie in that plane. Thus, $\boldsymbol{e}$ is a linear combination of two vectors in the orbital plane; so $\boldsymbol{e}$ itself must be in the orbital plane (Fig. 6.4). Later we shall see that it points to the direction where the planet is closest to the Sun in its orbit. This point is called the *perihelion*.

One more constant is found by computing $\dot{\boldsymbol{r}}\cdot\ddot{\boldsymbol{r}}$:

$$\begin{aligned}\dot{\boldsymbol{r}}\cdot\ddot{\boldsymbol{r}}&=-\mu\dot{\boldsymbol{r}}\cdot\boldsymbol{r}/r^3=-\mu r\dot{r}/r^3\\&=-\mu\dot{r}/r^2=\frac{\mathrm{d}}{\mathrm{d}t}(\mu/r)\,.\end{aligned}$$

Since we also have

$$\dot{\boldsymbol{r}}\cdot\ddot{\boldsymbol{r}}=\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\dot{\boldsymbol{r}}\cdot\dot{\boldsymbol{r}}\right)\,,$$

we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\dot{\boldsymbol{r}}\cdot\dot{\boldsymbol{r}}-\frac{\mu}{r}\right)=0$$

or

$$\frac{1}{2}v^2-\mu/r=\text{const}=h\,. \tag{6.11}$$

Here v is the speed of the planet relative to the Sun. The constant h is called the *energy integral*; the total energy of the planet is m_2h. We must not forget that energy and angular momentum depend on the coordinate frame used. Here we have used a heliocentric frame, which in fact is in accelerated motion.

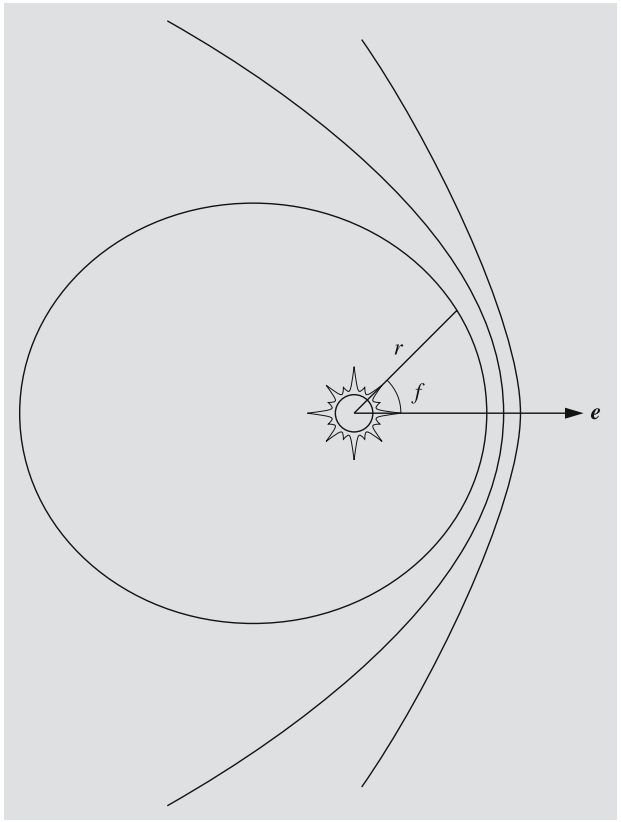


Fig. 6.4. The orbit of an object in the gravitational field of another object is a conic section: ellipse, parabola or hyperbola. Vector $\boldsymbol{e}$ points to the direction of the pericentre, where the orbiting object is closest to central body. If the central body is the Sun, this direction is called the perihelion; if some other star, periastron; if the Earth, perigee, etc. The true anomaly f is measured from the pericentre

So far, we have found two constant vectors and one constant scalar. It looks as though we already have seven integrals, i. e. one too many. But not all of these constants are independent; specifically, the following two relations hold:

$$\boldsymbol{k}\cdot\boldsymbol{e}=0\,, \tag{6.12}$$

$$\mu^2(e^2-1)=2hk^2\,, \tag{6.13}$$

where e and k are the lengths of $\boldsymbol{e}$ and $\boldsymbol{k}$. The first equation is obvious from the definitions of $\boldsymbol{e}$ and $\boldsymbol{k}$. To

prove (6.13), we square both sides of (6.10) to get

$$\mu^2 e^2 = (\boldsymbol{k}\times\dot{\boldsymbol{r}})\cdot(\boldsymbol{k}\times\dot{\boldsymbol{r}}) + \mu^2\frac{\boldsymbol{r}\cdot\boldsymbol{r}}{r^2} + 2(\boldsymbol{k}\times\dot{\boldsymbol{r}})\cdot\frac{\mu\boldsymbol{r}}{r}\ .$$

Since $\boldsymbol{k}$ is perpendicular to $\dot{\boldsymbol{r}}$, the length of $\boldsymbol{k}\times\dot{\boldsymbol{r}}$ is $|\boldsymbol{k}||\dot{\boldsymbol{r}}| = kv$ and $(\boldsymbol{k}\times\dot{\boldsymbol{r}})\cdot(\boldsymbol{k}\times\dot{\boldsymbol{r}}) = k^2v^2$. Thus, we have

$$\mu^2 e^2 = k^2v^2 + \mu^2 + \frac{2\mu}{r}(\boldsymbol{k}\times\dot{\boldsymbol{r}}\cdot\boldsymbol{r})\ .$$

The last term contains a scalar triple product, where we can exchange the dot and cross to get $\boldsymbol{k}\cdot\dot{\boldsymbol{r}}\times\boldsymbol{r}$. Next we reverse the order of the two last factors. Because the vector product is anticommutative, we have to change the sign of the product:

$$\begin{aligned}\mu^2(e^2-1) &= k^2v^2 - \frac{2\mu}{r}(\boldsymbol{k}\cdot\boldsymbol{r}\times\dot{\boldsymbol{r}}) = k^2v^2 - \frac{2\mu}{r}k^2\\ &= 2k^2\left(\frac{1}{2}v^2 - \frac{\mu}{r}\right) = 2k^2h\ .\end{aligned}$$

This completes the proof of (6.13).

The relations (6.12) and (6.13) reduce the number of independent integrals by two, so we still need one more. The constants we have describe the size, shape and orientation of the orbit completely, but we do not yet know where the planet is! To fix its position in the orbit, we have to determine where the planet is at some given instant of time $t = t_0$, or alternatively, at what time it is in some given direction. We use the latter method by specifying the time of perihelion passage, the *time of perihelion* τ.

6.3 Equation of the Orbit and Kepler's First Law

In order to find the geometric shape of the orbit, we now derive the equation of the orbit. Since $\boldsymbol{e}$ is a constant vector lying in the orbital plane, we choose it as the reference direction. We denote the angle between the radius vector $\boldsymbol{r}$ and $\boldsymbol{e}$ by f. The angle f is called the *true anomaly*. (There is nothing false or anomalous in this and other anomalies we shall meet later. Angles measured from the perihelion point are called anomalies to distinguish them from longitudes measured from some other reference point, usually the vernal equinox.) Using the properties of the scalar product we get

$$\boldsymbol{r}\cdot\boldsymbol{e} = re\cos f\ .$$

But the product $\boldsymbol{r}\cdot\boldsymbol{e}$ can also be evaluated using the definition of $\boldsymbol{e}$:

$$\begin{aligned}\boldsymbol{r}\cdot\boldsymbol{e} &= -\frac{1}{\mu}(\boldsymbol{r}\cdot\boldsymbol{k}\times\dot{\boldsymbol{r}} + \mu\boldsymbol{r}\cdot\boldsymbol{r}/r)\\ &= -\frac{1}{\mu}(\boldsymbol{k}\cdot\dot{\boldsymbol{r}}\times\boldsymbol{r} + \mu r) = -\frac{1}{\mu}(-k^2 + \mu r)\\ &= \frac{k^2}{\mu} - r\ .\end{aligned}$$

Equating the two expressions of $\boldsymbol{r}\cdot\boldsymbol{e}$ we get

$$r = \frac{k^2/\mu}{1 + e\cos f}\ . \tag{6.14}$$

This is the general equation of a *conic section* in polar coordinates (Fig. 6.4; see Appendix A.2 for a brief summary of conic sections). The magnitude of $\boldsymbol{e}$ gives the *eccentricity* of the conic:

$e = 0$	circle ,
$0 < e < 1$	ellipse ,
$e = 1$	parabola ,
$e > 1$	hyperbola .

Inspecting (6.14), we find that r attains its minimum when $f = 0$, i. e. in the direction of the vector $\boldsymbol{e}$. Thus, $\boldsymbol{e}$ indeed points to the direction of the perihelion.

Starting with Newton's laws, we have thus managed to prove Kepler's first law:

> The orbit of a planet is an ellipse, one focus of which is in the Sun.

Without any extra effort, we have shown that also other conic sections, the parabola and hyperbola, are possible orbits.

6.4 Orbital Elements

We have derived a set of integrals convenient for studying the dynamics of orbital motion. We now turn to another collection of constants more appropriate for describing the geometry of the orbit. The following six quantities are called the *orbital elements* (Fig. 6.5):

- semimajor axis a,
- eccentricity e,
- inclination i (or ι),

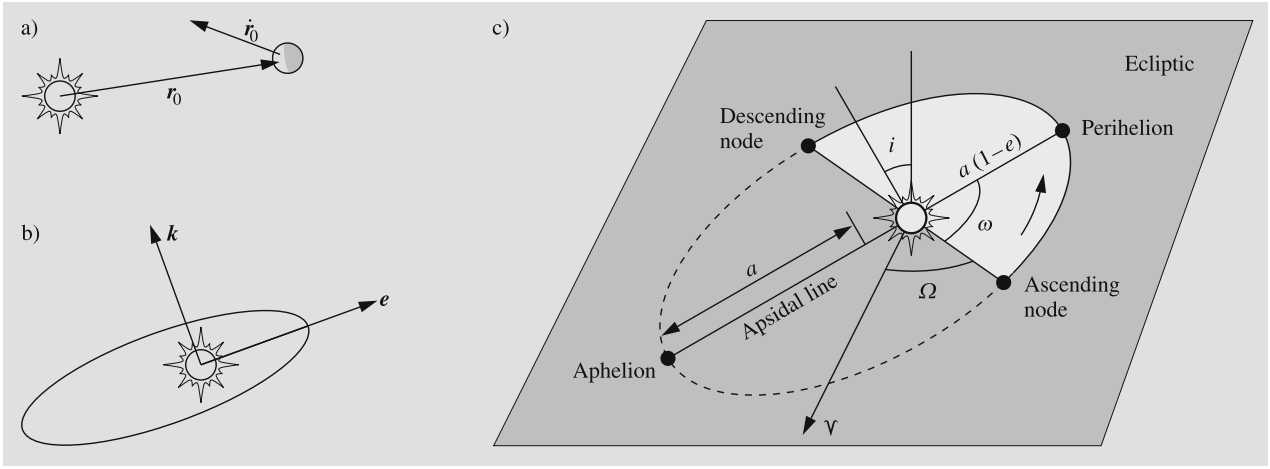


Fig. 6.5a–c. Six integration constants are needed to describe a planet's orbit. These constants can be chosen in various ways. (**a**) If the orbit is to be computed numerically, the simplest choice is to use the initial values of the radius and velocity vectors. (**b**) Another possibility is to use the angular momentum $\boldsymbol{k}$, the direction of the perihelion $\boldsymbol{e}$ (the length of which gives the eccentricity), and the perihelion time τ. (**c**) The third method best describes the geometry of the orbit. The constants are the longitude of the ascending node Ω, the argument of perihelion ω, the inclination i, the semimajor axis a, the eccentricity e and the time of perihelion τ

– longitude of the ascending node Ω,
– argument of the perihelion ω,
– time of the perihelion τ.

The eccentricity is obtained readily as the length of the vector $\boldsymbol{e}$. From the equation of the orbit (6.14), we see that the *parameter* (or semilatus rectum) of the orbit is $p = k^2/\mu$. But the parameter of a conic section is always $a|1-e^2|$, which gives the semimajor axis, if e and k are known:

$$a = \frac{k^2/\mu}{|1-e^2|} \,. \tag{6.15}$$

By applying (6.13), we get an important relation between the size of the orbit and the energy integral h:

$$a = \begin{cases} -\mu/2h \,, & \text{if the orbit is an ellipse} \,, \\ \mu/2h \,, & \text{if the orbit is a hyperbola} \,. \end{cases} \tag{6.16}$$

For a bound system (elliptical orbit), the total energy and the energy integral are negative. For a hyperbolic orbit h is positive; the kinetic energy is so high that the particle can escape the system (or more correctly, recede without any limit). The parabola, with $h = 0$, is a limiting case between elliptical and hyperbolic orbits. In reality parabolic orbits do not exist, since hardly any object can have an energy integral exactly zero. However, if the eccentricity is very close to one (as with many comets), the orbit is usually considered parabolic to simplify calculations.

The orientation of the orbit is determined by the directions of the two vectors $\boldsymbol{k}$ (perpendicular to the orbital plane) and $\boldsymbol{e}$ (pointing towards the perihelion). The three angles i, Ω and ω contain the same information.

The inclination i gives the obliquity of the orbital plane relative to some fixed reference plane. For bodies in the solar system, the reference plane is usually the ecliptic. For objects moving in the usual fashion, i. e. counterclockwise, the inclination is in the interval $[0°, 90°]$; for retrograde orbits (clockwise motion), the inclination is in the range $(90°, 180°]$. For example, the inclination of Halley's comet is 162°, which means that the motion is retrograde and the angle between its orbital plane and the ecliptic is $180° - 162° = 18°$.

The longitude of the ascending node, Ω, indicates where the object crosses the ecliptic from south to north. It is measured counterclockwise from the vernal equinox. The orbital elements i and Ω together determine the orientation of the orbital plane, and they correspond to the direction of $\boldsymbol{k}$, i. e. the ratios of its components.

The argument of the perihelion ω gives the direction of the perihelion, measured from the ascending node

in the direction of motion. The same information is contained in the direction of $\boldsymbol{e}$. Very often another angle, the *longitude of the perihelion* ϖ (pronounced as pi), is used instead of ω. It is defined as

$$\varpi = \Omega + \omega \,. \tag{6.17}$$

This is a rather peculiar angle, as it is measured partly along the ecliptic, partly along the orbital plane. However, it is often more practical than the argument of perihelion, since it is well defined even when the inclination is close to zero in which case the direction of the ascending node becomes indeterminate.

We have assumed up to this point that each planet forms a separate two-body system with the Sun. In reality planets interfere with each other by disturbing each other's orbits. Still their motions do not deviate very far from the shape of conic sections, and we can use orbital elements to describe the orbits. But the elements are no longer constant; they vary slowly with time. Moreover, their geometric interpretation is no longer quite as obvious as before. Such elements are *osculating elements* that would describe the orbit if all perturbations were to suddenly disappear. They can be used to find the positions and velocities of the planets exactly as if the elements were constants. The only difference is that we have to use different elements for each moment of time.

Table C.12 (at the end of the book) gives the mean orbital elements for the nine planets for the epoch J2000.0 as well as their first time derivatives. In addition to these secular variations the orbital elements suffer from periodic disturbations, which are not included in the table. Thus only approximate positions can be calculated with these elements. Instead of the time of perihelion the table gives the *mean longitude*

$$L = M + \omega + \Omega \,, \tag{6.18}$$

which gives directly the mean anomaly M (which will be defined in Sect. 6.7).

6.5 Kepler's Second and Third Law

The radius vector of a planet in polar coordinates is simply

$$\boldsymbol{r} = r\hat{\boldsymbol{e}}_r \,, \tag{6.19}$$

where $\hat{\boldsymbol{e}}_r$ is a unit vector parallel with $\boldsymbol{r}$ (Fig. 6.6). If the planet moves with angular velocity $\dot{f}$, the direction of this unit vector also changes at the same rate:

$$\dot{\hat{\boldsymbol{e}}}_r = \dot{f}\hat{\boldsymbol{e}}_f \,, \tag{6.20}$$

where $\hat{\boldsymbol{e}}_f$ is a unit vector perpendicular to $\hat{\boldsymbol{e}}_r$. The velocity of the planet is found by taking the time derivative of (6.19):

$$\dot{\boldsymbol{r}} = \dot{r}\hat{\boldsymbol{e}}_r + r\dot{\hat{\boldsymbol{e}}}_r = \dot{r}\hat{\boldsymbol{e}}_r + r\dot{f}\hat{\boldsymbol{e}}_f \,. \tag{6.21}$$

The angular momentum $\boldsymbol{k}$ can now be evaluated using (6.19) and (6.21):

$$\boldsymbol{k} = \boldsymbol{r} \times \dot{\boldsymbol{r}} = r^2 \dot{f}\hat{\boldsymbol{e}}_z \,, \tag{6.22}$$

where $\hat{\boldsymbol{e}}_z$ is a unit vector perpendicular to the orbital plane. The magnitude of $\boldsymbol{k}$ is

$$k = r^2 \dot{f} \,. \tag{6.23}$$

The *surface velocity* of a planet means the area swept by the radius vector per unit of time. This is obviously the time derivative of some area, so let us call it $\dot{A}$. In terms of the distance r and true anomaly f, the surface velocity is

$$\dot{A} = \frac{1}{2} r^2 \dot{f} \,. \tag{6.24}$$

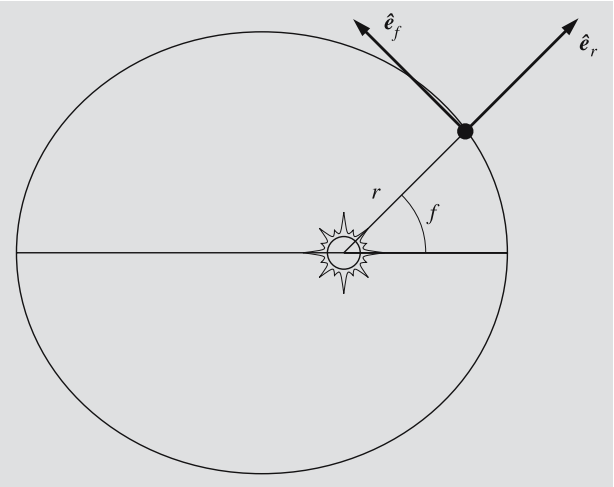


Fig. 6.6. Unit vectors $\hat{\boldsymbol{e}}_r$ and $\hat{\boldsymbol{e}}_f$ of the polar coordinate frame. The directions of these change while the planet moves along its orbit

By comparing this with the length of $\boldsymbol{k}$ (6.23), we find that

$$\dot{A} = \frac{1}{2}k \,. \tag{6.25}$$

Since k is constant, so is the surface velocity. Hence we have Kepler's second law:

> The radius vector of a planet sweeps equal areas in equal amounts of time.

Since the Sun–planet distance varies, the orbital velocity must also vary (Fig. 6.7). From Kepler's second law it follows that a planet must move fastest when it is closest to the Sun (near perihelion). Motion is slowest when the planet is farthest from the Sun at *aphelion*.

We can write (6.25) in the form

$$\mathrm{d}A = \frac{1}{2}k\,\mathrm{d}t \,, \tag{6.26}$$

and integrate over one complete period:

$$\int\limits_{\text{orbital ellipse}} \mathrm{d}A = \frac{1}{2}k \int\limits_{0}^{P} \mathrm{d}t \,, \tag{6.27}$$

where P is the orbital period. Since the area of the ellipse is

$$\pi ab = \pi a^2\sqrt{1-e^2} \,, \tag{6.28}$$

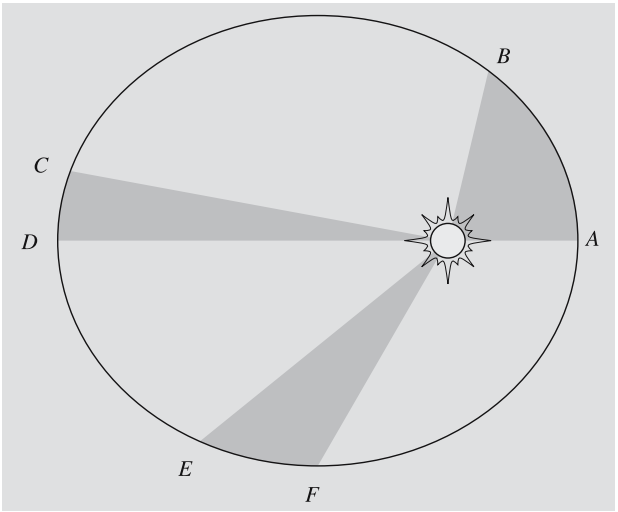


Fig. 6.7. The areas of the shaded sectors of the ellipse are equal. According to Kepler's second law, it takes equal times to travel distances AB, CD and EF

where a and b are the semimajor and semiminor axes and e the eccentricity, we get

$$\pi a^2\sqrt{1-e^2} = \frac{1}{2}kP \,. \tag{6.29}$$

To find the length of $\boldsymbol{k}$, we substitute the energy integral h as a function of semimajor axis (6.16) into (6.13) to get

$$k = \sqrt{G(m_1+m_2)a\,(1-e^2)} \,. \tag{6.30}$$

When this is substituted into (6.29) we have

$$P^2 = \frac{4\pi^2}{G(m_1+m_2)}\,a^3 \,. \tag{6.31}$$

This is the exact form of Kepler's third law as derived from Newton's laws. The original version was

> The ratio of the cubes of the semimajor axes of the orbits of two planets is equal to the ratio of the squares of their orbital periods.

In this form the law is not exactly valid, even for planets of the solar system, since their own masses influence their periods. The errors due to ignoring this effect are very small, however.

Kepler's third law becomes remarkably simple if we express distances in astronomical units (AU), times in sidereal years (the abbreviation is unfortunately a, not to be confused with the semimajor axis, denoted by a somewhat similar symbol a) and masses in solar masses ($M_\odot$). Then $G = 4\pi^2$ and

$$a^3 = (m_1+m_2)P^2 \,. \tag{6.32}$$

The masses of objects orbiting around the Sun can safely be ignored (except for the largest planets), and we have the original law $P^2 = a^3$. This is very useful for determining distances of various objects whose periods have been observed. For absolute distances we have to measure at least one distance in metres to find the length of one AU. Earlier, triangulation was used to measure the parallax of the Sun or a minor planet, such as Eros, that comes very close to the Earth. Nowadays, radiotelescopes are used as radar to very accurately measure, for example, the distance to Venus. Since changes in the value of one AU also change all other distances, the International Astronomical Union decided in 1968 to adopt the value $1\ \mathrm{AU} = 1.496000 \times 10^{11}$ m. The semimajor axis of Earth's orbit is then slightly over one AU.

But constants tend to change. And so, after 1984, the astronomical unit has a new value,

$$1\ \mathrm{AU} = 1.49597870 \times 10^{11}\ \mathrm{m}\ .$$

Another important application of Kepler's third law is the determination of masses. By observing the period of a natural or artificial satellite, the mass of the central body can be obtained immediately. The same method is used to determine masses of binary stars (more about this subject in Chap. 9).

Although the values of the AU and year are accurately known in SI-units, the gravitational constant is known only approximately. Astronomical observations give the product $G(m_1 + m_2)$, but there is no way to distinguish between the contributions of the gravitational constant and those of the masses. The gravitational constant must be measured in the laboratory; this is very difficult because of the weakness of gravitation. Therefore, if a precision higher than 2–3 significant digits is required, the SI-units cannot be used. Instead we have to use the solar mass as a unit of mass (or, for example, the Earth's mass after $Gm_\oplus$ has been determined from observations of satellite orbits).

6.6 Systems of Several Bodies

This far we have discussed systems consisting of only two bodies. In fact it is the most complex system for which a complete solution is known. The equations of motion are easily generalized, though. As in (6.5) we get the equation of motion for the body k, $k = 1, \ldots, n$:

$$\ddot{\boldsymbol{r}}_k = \sum_{i=1,\, i\neq k}^{i=n} Gm_i \frac{\boldsymbol{r}_i - \boldsymbol{r}_k}{|\boldsymbol{r}_i - \boldsymbol{r}_k|^3}\ , \tag{6.33}$$

where m_i is the mass of the ith body and $\boldsymbol{r}_i$ its radius vector. On the right hand side of the equation we now have the total gravitational force due to all other objects, instead of the force of just one body. If there are more than two bodies, these equations cannot be solved analytically in a closed form. The only integrals that can be easily derived in the general case are the total energy, total momentum, and total angular momentum.

If the radius and velocity vectors of all bodies are known for a certain instant of time, the positions at some other time can easily be calculated numerically from the equations of motion. For example, the planetary positions needed for astronomical yearbooks are computed by integrating the equations numerically.

Another method can be applied if the gravity of one body dominates like in the solar system. Planetary orbits can then be calculated as in a two-body system, and the effects of other planets taken into account as small perturbations. For these perturbations several series expansions have been derived.

The *restricted three-body problem* is an extensively studied special case. It consists of two massive bodies or *primaries*, moving on circular orbits around each other, and a third, massless body, moving in the same plane with the primaries. This small object does in no way disturb the motion of the primaries. Thus the orbits of the massive bodies are as simple as possible, and their positions are easily computed for all times. The problem is to find the orbit of the third body. It turns out that there is no finite expression for this orbit.

The Finnish astronomer *Karl Frithiof Sundman* (1873–1949) managed to show that a solution exists and derive a series expansion for the orbit. The series converges so slowly that it has no practical use, but as a mathematical result it was remarkable, since many mathematicians had for a long time tried to attack the problem without success.

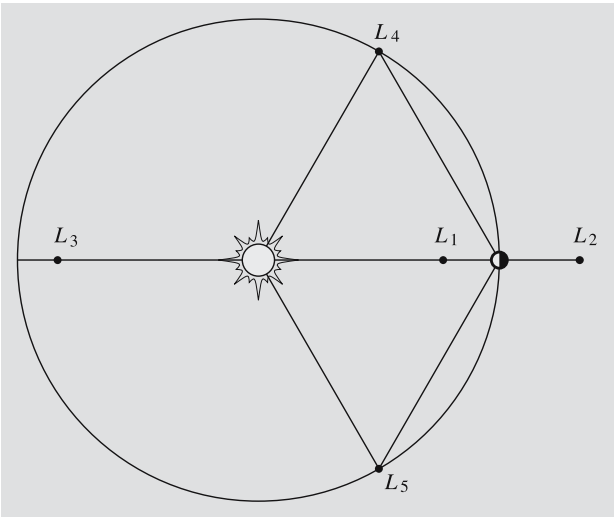


Fig. 6.8. The Lagrangian points of the restricted three-body problem. The points L_1, L_2 and L_3 are on the same line with the primaries, but the numbering may vary. The points L_4 and L_5 form equilateral triangles with the primaries

The three-body problem has some interesting special solutions. It can be shown that in certain points the third body can remain at rest with respect to the primaries. There are five such points, known as the *Lagrangian points* $L_1, \ldots, L_5$ (Fig. 6.8). Three of them are on the straight line determined by the primaries. These points are unstable: if a body in any of these points is disturbed, it will escape. The two other points, on the other hand, are stable. These points together with the primaries form equilateral triangles. For example, some *asteroids* have been found around the Lagrangian points L_4 and L_5 of Jupiter and Mars. The first of them were named after heroes of the Trojan war, and so they are called *Trojan asteroids*. They move around the Lagrangian points and can actually travel quite far from them, but they cannot escape. Fig. 7.56 shows two distinct condensations around the Lagrangian points of Jupiter.

6.7 Orbit Determination

Celestial mechanics has two very practical tasks: to determine orbital elements from observations and to predict positions of celestial bodies with known elements. Planetary orbits are already known very accurately, but new comets and minor planets are found frequently, requiring orbit determination.

The first practical methods for orbit determination were developed by *Johann Karl Friedrich Gauss* (1777–1855) at the beginning of the 19th century. By that time the first minor planets had been discovered, and thanks to Gauss's orbit determinations, they could be found and observed at any time.

At least three observations are needed for computing the orbital elements. The directions are usually measured from pictures taken a few nights apart. Using these directions, it is possible to find the corresponding absolute positions (the rectangular components of the radius vector). To be able to do this, we need some additional constraints on the orbit; we must assume that the object moves along a conic section lying in a plane that passes through the Sun. When the three radius vectors are known, the ellipse (or some other conic section) going through these three points can be determined. In practice, more observations are used. The elements determined are more accurate if there are more observations and if they cover the orbit more completely.

Although the calculations for orbit determination are not too involved mathematically, they are relatively long and laborious. Several methods can be found in textbooks of celestial mechanics.

6.8 Position in the Orbit

Although we already know everything about the geometry of the orbit, we still cannot find the planet at a given time, since we do not know the radius vector $\boldsymbol{r}$ as a function of time. The variable in the equation of the orbit is an angle, the true anomaly f, measured from the perihelion. From Kepler's second law it follows that f cannot increase at a constant rate with time. Therefore we need some preparations before we can find the radius vector at a given instant.

The radius vector can be expressed as

$$\boldsymbol{r} = a(\cos E - e)\hat{\boldsymbol{i}} + b \sin E \hat{\boldsymbol{j}}\,, \tag{6.34}$$

where $\hat{\boldsymbol{i}}$ and $\hat{\boldsymbol{j}}$ are unit vectors parallel with the major and minor axes, respectively. The angle E is the *eccentric anomaly*; its slightly eccentric definition is shown in Fig. 6.9. Many formulas of elliptical motion become very simple if either time or true anomaly is replaced

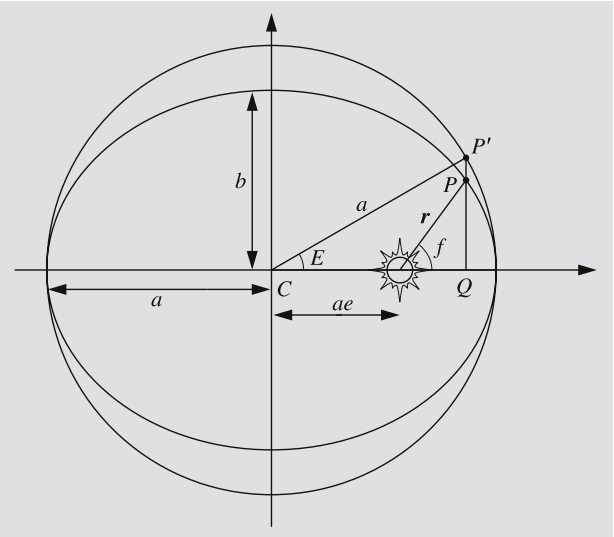


Fig. 6.9. Definition of the eccentric anomaly E. The planet is at P, and $\boldsymbol{r}$ is its radius vector

by the eccentric anomaly. As an example, we take the square of (6.34) to find the distance from the Sun:

$$\begin{aligned} r^2 &= \boldsymbol{r}\cdot\boldsymbol{r} \\ &= a^2(\cos E - e)^2 + b^2 \sin^2 E \\ &= a^2[(\cos E - e)^2 + (1-e^2)(1-\cos^2 E)] \\ &= a^2[1 - 2e\cos E + e^2\cos^2 E] \,, \end{aligned}$$

whence

$$r = a\,(1 - e\cos E)\,. \tag{6.35}$$

Our next problem is to find how to calculate E for a given moment of time. According to Kepler's second law, the surface velocity is constant. Thus the area of the shaded sector in Fig. 6.10 is

$$A = \pi ab\frac{t-\tau}{P}\,, \tag{6.36}$$

where $t-\tau$ is the time elapsed since the perihelion, and P is the orbital period. But the area of a part of an ellipse is obtained by reducing the area of the corresponding part of the circumscribed circle by the axial ratio b/a. (As the mathematicians say, an ellipse is an affine transformation of a circle.) Hence the area of SPX is

$$\begin{aligned} A &= \frac{b}{a}(\text{area of } SP'X) \\ &= \frac{b}{a}(\text{area of the sector } CP'X \\ &\quad - \text{area of the triangle } CP'S) \\ &= \frac{b}{a}\left(\frac{1}{2}a\cdot aE - \frac{1}{2}ae\cdot a\sin E\right) \\ &= \frac{1}{2}ab(E - e\sin E)\,. \end{aligned}$$

By equating these two expressions for the area A, we get the famous *Kepler's equation*,

$$E - e\sin e = M\,, \tag{6.37}$$

where

$$M = \frac{2\pi}{P}(t-\tau) \tag{6.38}$$

is the *mean anomaly* of the planet at time t. The mean anomaly increases at a constant rate with time. It indicates where the planet would be if it moved in a circular orbit of radius a. For circular orbits all three anomalies f, E, and M are always equal.

If we know the period and the time elapsed after the perihelion, we can use (6.38) to find the mean anomaly. Next we must solve for the eccentric anomaly from Kepler's equation (6.37). Finally the radius vector is given by (6.35). Since the components of $\boldsymbol{r}$ expressed in terms of the true anomaly are $r\cos f$ and $r\sin f$, we find

$$\begin{aligned} \cos f &= \frac{a(\cos E - e)}{r} = \frac{\cos E - e}{1 - e\cos E}\,, \\ \sin f &= \frac{b\sin E}{r} = \sqrt{1-e^2}\,\frac{\sin E}{1-e\cos E}\,. \end{aligned} \tag{6.39}$$

These determine the true anomaly, should it be of interest.

Now we know the position in the orbital plane. This must usually be transformed to some other previously selected reference frame. For example, we may want to know the ecliptic longitude and latitude, which can later be used to find the right ascension and declination. These transformations belong to the realm of spherical astronomy and are briefly discussed in Examples 6.5–6.7.

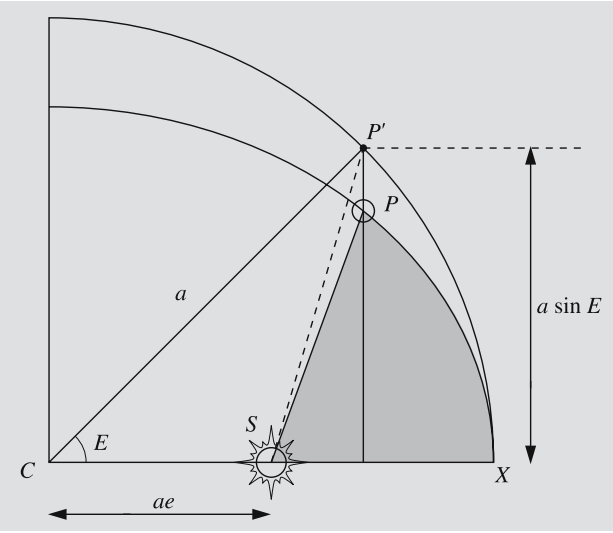


Fig. 6.10. The area of the shaded sector equals b/a times the area $SP'X$. S = the Sun, P = the planet, X = the perihelion

6.9 Escape Velocity

If an object moves fast enough, it can escape from the gravitational field of the central body (to be precise: the field extends to infinity, so the object never really escapes, but is able to recede without any limit). If the escaping object has the minimum velocity allowing escape, it will have lost all its velocity at infinity (Fig. 6.11). There its kinetic energy is zero, since $v = 0$, and the potential energy is also zero, since the distance r is infinite. At infinite distance the total energy as well as the energy integral h are zero. The law of conservation of energy gives, then:

$$\frac{1}{2}v^2 - \frac{\mu}{R} = 0\,, \tag{6.40}$$

where R is the initial distance at which the object is moving with velocity v. From this we can solve the *escape velocity*:

$$v_\mathrm{e} = \sqrt{\frac{2\mu}{R}} = \sqrt{\frac{2G(m_1+m_2)}{R}}\,. \tag{6.41}$$

For example on the surface of the Earth, v_e is about 11 km/s (if $m_2 \ll m_\oplus$).

The escape velocity can also be expressed using the orbital velocity of a circular orbit. The orbital period P as a function of the radius R of the orbit and the orbital velocity v_c is

$$P = \frac{2\pi R}{v_\mathrm{c}}\,.$$

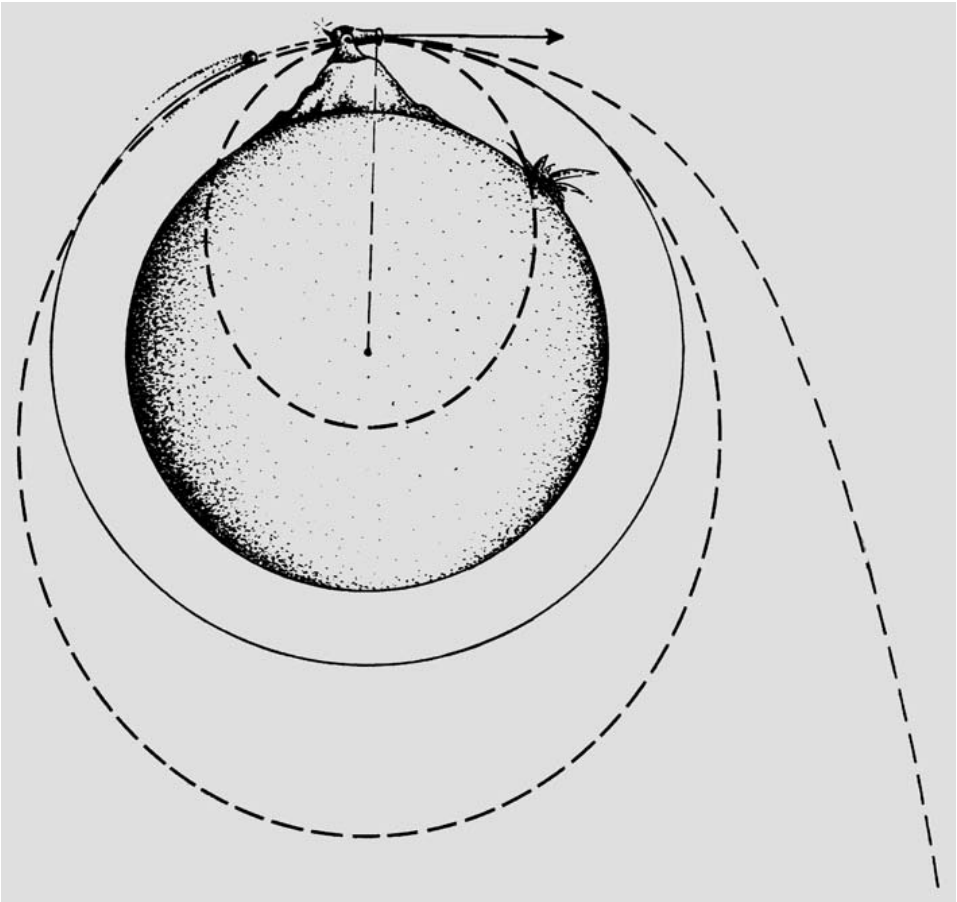


Fig. 6.11. A projectile is shot horizontally from a mountain on an atmosphereless planet. If the initial velocity is small, the orbit is an ellipse whose pericentre is inside the planet, and the projectile will hit the surface of the planet. When the velocity is increased, the pericentre moves outside the planet. When the initial velocity is v_c, the orbit is circular. If the velocity is increased further, the eccentricity of the orbit grows again and the pericentre is at the height of the cannon. The apocentre moves further away until the orbit becomes parabolic when the initial velocity is v_e. With even higher velocities, the orbit becomes hyperbolic

Substitution into Kepler's third law yields

$$\frac{4\pi^2 R^2}{v_c^2} = \frac{4\pi^2 R^3}{G(m_1+m_2)} .$$

From this we can solve the velocity v_c in a circular orbit of radius R:

$$v_c = \sqrt{\frac{G(m_1+m_2)}{R}} . \tag{6.42}$$

Comparing this with the expression (6.41) of the escape velocity, we see that

$$v_e = \sqrt{2} v_c . \tag{6.43}$$

6.10 Virial Theorem

If a system consists of more than two objects, the equations of motion cannot in general be solved analytically (Fig. 6.12). Given some initial values, the orbits can, of course, be found by numerical integration, but this does not tell us anything about the general properties of all possible orbits. The only integration constants available for an arbitrary system are the total momentum, angular momentum and energy. In addition to these, it is possible to derive certain statistical results, like the virial theorem. It concerns time averages only, but does not say anything about the actual state of the system at some specified moment.

Suppose we have a system of n point masses m_i with radius vectors $\boldsymbol{r}_i$ and velocities $\dot{\boldsymbol{r}}_i$. We define a quantity A (the "virial" of the system) as follows:

$$A = \sum_{i=1}^{n} m_i \dot{\boldsymbol{r}}_i \cdot \boldsymbol{r}_i . \tag{6.44}$$

The time derivative of this is

$$\dot{A} = \sum_{i=1}^{n} (m_i \dot{\boldsymbol{r}}_i \cdot \dot{\boldsymbol{r}}_i + m_i \ddot{\boldsymbol{r}}_i \cdot \boldsymbol{r}_i) . \tag{6.45}$$

The first term equals twice the kinetic energy of the ith particle, and the second term contains a factor $m_i \ddot{\boldsymbol{r}}_i$ which, according to Newton's laws, equals the force applied to the ith particle. Thus we have

$$\dot{A} = 2T + \sum_{i=1}^{n} \boldsymbol{F}_i \cdot \boldsymbol{r}_i , \tag{6.46}$$

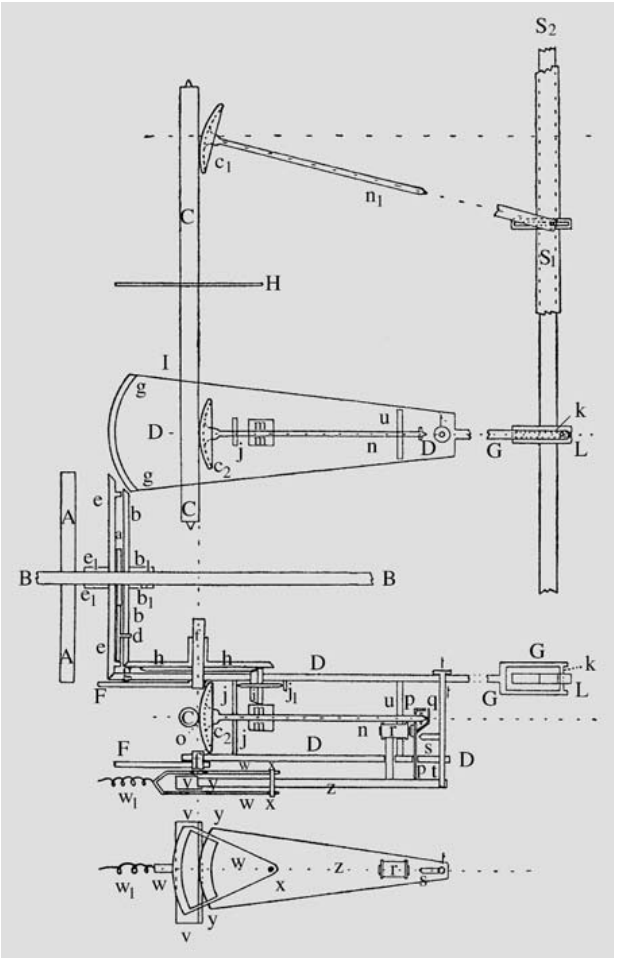


Fig. 6.12. When a system consists of more than two bodies, the equations of motion cannot be solved analytically. In the solar system the mutual disturbances of the planets are usually small and can be taken into account as small perturbations in the orbital elements. K.F. Sundman designed a machine to carry out the tedious integration of the perturbation equations. This machine, called the perturbograph, is one of the earliest analogue computers; unfortunately it was never built. Shown is a design for one component that evaluates a certain integral occurring in the equations. (The picture appeared in K.F. Sundman's paper in *Festskrift tillegnad Anders Donner* in 1915.)

where T is the total kinetic energy of the system. If $\langle x \rangle$ denotes the time average of x in the time interval $[0, \tau]$, we have

$$\langle \dot{A} \rangle = \frac{1}{\tau} \int_0^\tau \dot{A} \, \mathrm{d}t = \langle 2T \rangle + \left\langle \sum_{i=1}^{n} \boldsymbol{F}_i \cdot \boldsymbol{r}_i \right\rangle . \tag{6.47}$$

If the system remains bounded, i. e. none of the particles escapes, all $\boldsymbol{r}_i$'s as well as all velocities will remain bounded. In such a case, A does not grow without limit, and the integral of the previous equation remains finite. When the time interval becomes longer ($\tau \to \infty$), $\langle \dot{A} \rangle$ approaches zero, and we get

$$\langle 2T \rangle + \left\langle \sum_{i=1}^{n} \boldsymbol{F}_i \cdot \boldsymbol{r}_i \right\rangle = 0 \,. \tag{6.48}$$

This is the general form of the virial theorem. If the forces are due to mutual gravitation only, they have the expressions

$$\boldsymbol{F}_i = -Gm_i \sum_{j=1,\, j\neq i}^{n} m_j \frac{\boldsymbol{r}_i - \boldsymbol{r}_j}{r_{ij}^3} \,, \tag{6.49}$$

where $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$. The latter term in the virial theorem is now

$$\begin{aligned} \sum_{i=1}^{n} \boldsymbol{F}_i \cdot \boldsymbol{r}_i &= -G \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} m_i m_j \frac{\boldsymbol{r}_i - \boldsymbol{r}_j}{r_{ij}^3} \cdot \boldsymbol{r}_i \\ &= -G \sum_{i=1}^{n} \sum_{j=i+1}^{n} m_i m_j \frac{\boldsymbol{r}_i - \boldsymbol{r}_j}{r_{ij}^3} \cdot (\boldsymbol{r}_i - \boldsymbol{r}_j) \,, \end{aligned}$$

where the latter form is obtained by rearranging the double sum, combining the terms

$$m_i m_j \frac{\boldsymbol{r}_i - \boldsymbol{r}_j}{r_{ij}^3} \cdot \boldsymbol{r}_i$$

and

$$m_j m_i \frac{\boldsymbol{r}_j - \boldsymbol{r}_i}{r_{ji}^3} \cdot \boldsymbol{r}_j = m_i m_j \frac{\boldsymbol{r}_i - \boldsymbol{r}_j}{r_{ij}^3} \cdot (-\boldsymbol{r}_j) \,.$$

Since $(\boldsymbol{r}_i - \boldsymbol{r}_j) \cdot (\boldsymbol{r}_i - \boldsymbol{r}_j) = r_{ij}^2$ the sum reduces to

$$-G \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{m_i m_j}{r_{ij}} = U \,,$$

where U is the potential energy of the system. Thus, the virial theorem becomes simply

$$\langle T \rangle = -\frac{1}{2} \langle U \rangle \,. \tag{6.50}$$

6.11 The Jeans Limit

We shall later study the birth of stars and galaxies. The initial stage is, roughly speaking, a gas cloud that begins to collapse due to its own gravitation. If the mass of the cloud is high enough, its potential energy exceeds the kinetic energy and the cloud collapses. From the virial theorem we can deduce that the potential energy must be at least twice the kinetic energy. This provides a criterion for the critical mass necessary for the cloud of collapse. This criterion was first suggested by *Sir James Jeans* in 1902.

The critical mass will obviously depend on the pressure P and density ρ. Since gravitation is the compressing force, the gravitational constant G will probably also enter our expression. Thus the critical mass is of the form

$$M = CP^a G^b \rho^c \,, \tag{6.51}$$

where C is a dimensionless constant, and the constants a, b and c are determined so that the right-hand side has the dimension of mass. The dimension of pressure is $\mathrm{kg\,m^{-1}\,s^{-2}}$, of gravitational constant $\mathrm{kg^{-1}\,m^3\,s^{-2}}$ and of density $\mathrm{kg\,m^{-3}}$. Thus the dimension of the right-hand side is

$$\mathrm{kg}^{(a-b+c)}\,\mathrm{m}^{(-a+3b-3c)}\,\mathrm{s}^{(-2a-2b)} \,.$$

Since this must be kilograms ultimately, we get the following set of equations:

$$\begin{aligned} &a - b + c = 1 \,, \quad -a + 3b - 3c = 0 \\ &-2a - 2b = 0 \,. \end{aligned}$$

The solution of this is $a = 3/2$, $b = -3/2$ and $c = -2$. Hence the critical mass is

$$M_{\mathrm{J}} = C \frac{P^{3/2}}{G^{3/2}\rho^2} \,. \tag{6.52}$$

This is called the *Jeans mass*. In order to determine the constant C, we naturally must calculate both kinetic and potential energy. Another method based on the propagation of waves determines the diameter of the cloud, the *Jeans length* λ_{J}, by requiring that a disturbance of size λ_{J} grow unbounded. The value of the constant C depends on the exact form of the perturbation, but its typical values are in the range $[1/\pi, 2\pi]$. We can take $C = 1$ as well, in which case (6.52) gives a correct order of magnitude for the critical mass. If the mass of

a cloud is much higher than M_J, it will collapse by its own gravitation.

In (6.52) the pressure can be replaced by the kinetic temperature T_k of the gas (see Sect. 5.8 for a definition). According to the kinetic gas theory, the pressure is

$$P = nkT_k \,, \tag{6.53}$$

where n is the number density (particles per unit volume) and k is Boltzmann's constant. The number density is obtained by dividing the density of the gas ρ by the average molecular weight μ:

$$n = \rho/\mu \,,$$

whence

$$P = \rho k T_k/\mu \,.$$

By substituting this into (6.52) we get

$$M_J = C \left(\frac{kT_k}{\mu G}\right)^{3/2} \frac{1}{\sqrt{\rho}} \,. \tag{6.54}$$

* Newton's Laws

1. In the absence of external forces, a particle will remain at rest or move along a straight line with constant speed.
2. The rate of change of the momentum of a particle is equal to the applied force $\boldsymbol{F}$:
$$\dot{\boldsymbol{p}} = \frac{\mathrm{d}}{\mathrm{d}t}(m\boldsymbol{v}) = \boldsymbol{F} \,.$$
3. If particle A exerts a force $\boldsymbol{F}$ on another particle B, B will exert an equal but opposite force $-\boldsymbol{F}$ on A.

If several forces $\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots$ are applied on a particle, the effect is equal to that caused by one force $\boldsymbol{F}$ which is the vector sum of the individual forces ($\boldsymbol{F} = \boldsymbol{F}_1 + \boldsymbol{F}_2 + \ldots$).

Law of gravitation: If the masses of particles A and B are m_A and m_B and their mutual distance r, the force exerted on A by B is directed towards B and has the magnitude $Gm_A\, m_B/r^2$, where G is a constant depending on the units chosen.

Newton denoted the derivative of a function f by $\dot{f}$ and the integral function by f'. The corresponding notations used by Leibniz were $\mathrm{d}f/\mathrm{d}t$ and $\int f\,\mathrm{d}x$. Of Newton's notations, only the dot is still used, always signifying the time derivative: $\dot{f} \equiv \mathrm{d}f/\mathrm{d}t$. For example, the velocity $\dot{\boldsymbol{r}}$ is the time derivative of $\boldsymbol{r}$, the acceleration $\ddot{\boldsymbol{r}}$ its second derivative, etc.

6.12 Examples

Example 6.1 Find the orbital elements of Jupiter on August 23, 1996.

The Julian date is 2,450,319, hence from (6.17), $T = -0.0336$. By substituting this into the expressions of Table C.12, we get

$$\begin{aligned} a &= 5.2033 \,, \\ e &= 0.0484 \,, \\ i &= 1.3053^\circ \,, \\ \Omega &= 100.5448^\circ \,, \\ \varpi &= 14.7460^\circ \,, \\ L &= -67.460^\circ = 292.540^\circ \,. \end{aligned}$$

From these we can compute the argument of perihelion and mean anomaly:

$$\begin{aligned} \omega &= \varpi - \Omega = -85.7988^\circ = 274.201^\circ \,, \\ M &= L - \varpi = -82.2060^\circ = 277.794^\circ \,. \end{aligned}$$

Example 6.2 *Orbital Velocity*

a) Comet Austin (1982g) moves in a parabolic orbit. Find its velocity on October 8, 1982, when the distance from the Sun was 1.10 AU.

The energy integral for a parabola is $h = 0$. Thus (6.11) gives the velocity v:

$$\begin{aligned} v &= \sqrt{\frac{2\mu}{r}} = \sqrt{\frac{2GM_\odot}{r}} \\ &= \sqrt{\frac{2 \times 4\pi^2 \times 1}{1.10}} = 8.47722 \text{ AU/a} \\ &= \frac{8.47722 \times 1.496 \times 10^{11} \text{ m}}{365.2564 \times 24 \times 3600 \text{ s}} \approx 40 \text{ km/s} \,. \end{aligned}$$

b) The semimajor axis of the minor planet 1982 RA is 1.568 AU and the distance from the Sun on October 8, 1982, was 1.17 AU. Find its velocity.

The energy integral (6.16) is now

$$h = -\mu/2a \, .$$

Hence

$$\frac{1}{2}v^2 - \frac{\mu}{r} = -\frac{\mu}{2a} \, ,$$

which gives

$$\begin{aligned} v &= \sqrt{\mu\left(\frac{2}{r} - \frac{1}{a}\right)} \\ &= \sqrt{4\pi^2\left(\frac{2}{1.17} - \frac{1}{1.568}\right)} \\ &= 6.5044 \text{ AU/a} \approx 31 \text{ km/s} \, . \end{aligned}$$

Example 6.3 In an otherwise empty universe, two rocks of 5 kg each orbit each other at a distance of 1 m. What is the orbital period?

The period is obtained from Kepler's third law:

$$\begin{aligned} P^2 &= \frac{4\pi^2 a^3}{G(m_1 + m_2)} \\ &= \frac{4\pi^2 1}{6.67 \times 10^{-11}(5+5)} \text{ s}^2 \\ &= 5.9 \times 10^{10} \text{ s}^2 \, , \end{aligned}$$

whence

$$P = 243{,}000 \text{ s} = 2.8 \text{ d} \, .$$

Example 6.4 The period of the Martian moon Phobos is 0.3189 d and the radius of the orbit 9370 km. What is the mass of Mars?

First we change to more appropriate units:

$$P = 0.3189 \text{ d} = 0.0008731 \text{ sidereal years} \, ,$$

$$a = 9370 \text{ km} = 6.2634 \times 10^{-5} \text{ AU} \, .$$

Equation (6.32) gives (it is safe to assume that $m_{\text{Phobos}} \ll m_{\text{Mars}}$)

$$\begin{aligned} m_{\text{Mars}} &= a^3/P^2 = 0.000000322 \, M_\odot \\ &\qquad (\approx 0.107 \, M_\oplus) \, . \end{aligned}$$

Example 6.5 Derive formulas for a planet's heliocentric longitude and latitude, given its orbital elements and true anomaly.

We apply the sine formula to the spherical triangle of the figure:

$$\frac{\sin\beta}{\sin i} = \frac{\sin(\omega + f)}{\sin(\pi/2)}$$

or

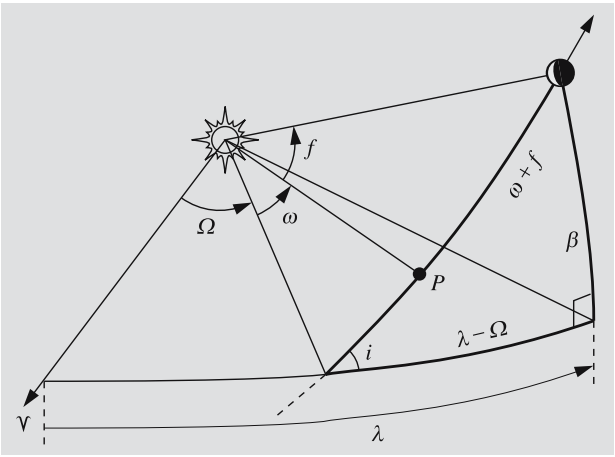
$$\sin\beta = \sin i \sin(\omega + f) \, .$$

The sine-cosine formula gives

$$\begin{aligned} &\cos(\pi/2)\sin\beta \\ &\quad = -\cos i \sin(\omega + f)\cos(\lambda - \Omega) \\ &\qquad + \cos(\omega + f)\sin(\lambda - \Omega) \, , \end{aligned}$$

whence

$$\tan(\lambda - \Omega) = \cos i \tan(\omega + f) \, .$$



Example 6.6 Find the radius vector and heliocentric longitude and latitude of Jupiter on August 23, 1996.

The orbital elements were computed in Example 6.1:

$$a = 5.2033 \text{ AU} \, ,$$
$$e = 0.0484 \, ,$$
$$i = 1.3053^\circ \, ,$$
$$\Omega = 100.5448^\circ \, ,$$
$$\omega = 274.2012^\circ \, ,$$
$$M = 277.7940^\circ = 4.8484 \text{ rad} \, .$$

Since the mean anomaly was obtained directly, we need not compute the time elapsed since perihelion.

Now we have to solve Kepler's equation. It cannot be solved analytically, and we are obliged to take the brute force approach (also called numerical analysis) in the form of iteration. For iteration, we write the equation as

$$E_{n+1} = M + e \sin E_n \ ,$$

where E_n is the value found in the nth iteration. The mean anomaly is a reasonable initial guess, E_0. (N.B.: Here, all angles must be in radians; otherwise, nonsense results!) The iteration proceeds as follows:

$$\begin{aligned} E_0 &= M = 4.8484 \ , \\ E_1 &= M + e \sin E_0 = 4.8004 \ , \\ E_2 &= M + e \sin E_1 = 4.8002 \ , \\ E_3 &= M + e \sin E_2 = 4.8002 \ , \end{aligned}$$

after which successive approximations no longer change, which means that the solution, accurate to four decimal places, is

$$E = 4.8002 = 275.0^\circ \ .$$

The radius vector is

$$\begin{aligned} \boldsymbol{r} &= a(\cos E - e)\,\hat{\boldsymbol{i}} + a\sqrt{1-e^2}\sin E\,\hat{\boldsymbol{j}} \\ &= 0.2045\,\hat{\boldsymbol{i}} - 5.1772\,\hat{\boldsymbol{j}} \end{aligned}$$

and the distance from the Sun,

$$r = a(1 - e\cos E) = 5.1813 \text{ AU} \ .$$

The signs of the components of the radius vector show that the planet is in the fourth quadrant. The true anomaly is

$$f = \arctan \frac{-5.1772}{0.2045} = 272.3^\circ \ .$$

Applying the results of the previous example, we find the latitude and longitude:

$$\begin{aligned} \sin\beta &= \sin i \sin(\omega + f) \\ &= \sin 1.3^\circ \sin(274.2^\circ + 272.3^\circ) \\ &= -0.0026 \\ \Rightarrow \beta &= -0.15^\circ \ , \end{aligned}$$

$$\begin{aligned} \tan(\lambda - \Omega) &= \cos i \tan(\omega + f) \\ &= \cos 1.3^\circ \tan(274.2^\circ + 272.3^\circ) \\ &= 0.1139 \\ \Rightarrow \lambda &= \Omega + 186.5^\circ \\ &= 100.5^\circ + 186.5^\circ \\ &= 287.0^\circ \ . \end{aligned}$$

(We must be careful here; the equation for $\tan(\lambda - \Omega)$ allows two solutions. If necessary, a figure can be drawn to decide which is the correct one.)

Example 6.7 Find Jupiter's right ascension and declination on August 23, 1996.

In Example 6.6, we found the longitude and latitude, $\lambda = 287.0^\circ$, $\beta = -0.15^\circ$. The corresponding rectangular (heliocentric) coordinates are:

$$\begin{aligned} x &= r\cos\lambda\cos\beta = 1.5154 \text{ AU} \ , \\ y &= r\sin\lambda\cos\beta = -4.9547 \text{ AU} \ , \\ z &= r\sin\beta = -0.0133 \text{ AU} \ . \end{aligned}$$

Jupiter's ecliptic coordinates must be transformed to equatorial ones by rotating them around the x-axis by an angle ε, the obliquity of the ecliptic (see *Reduction of Coordinates, p. 38):

$$\begin{aligned} X_\mathrm{J} &= x = 1.5154 \text{ AU} \ , \\ Y_\mathrm{J} &= y\cos\varepsilon - z\sin\varepsilon = -4.5405 \text{ AU} \ , \\ Z_\mathrm{J} &= y\sin\varepsilon + z\cos\varepsilon = -1.9831 \text{ AU} \ . \end{aligned}$$

To find the direction relative to the Earth, we have to find where the Earth is. In principle, we could repeat the previous procedure with the orbital elements of the Earth. Or, if we are lazy, we could pick up the nearest *Astronomical Almanac*, which lists the equatorial

coordinates of the Earth:

$$X_\oplus = 0.8815\ \text{AU}\ ,$$
$$Y_\oplus = -0.4543\ \text{AU}\ ,$$
$$Z_\oplus = -0.1970\ \text{AU}\ .$$

Then the position relative to the Earth is

$$X_0 = X_\text{J} - X_\oplus = 0.6339\ \text{AU}\ ,$$
$$Y_0 = Y_\text{J} - Y_\oplus = -4.0862\ \text{AU}\ ,$$
$$Z_0 = Z_\text{J} - Z_\oplus = -1.7861\ \text{AU}\ .$$

And finally, the right ascension and declination are

$$\alpha = \arctan(Y_0/X_0) = 278.82^\circ = 18\,\text{h}\ 35\,\text{min}\ ,$$
$$\delta = \arctan\frac{Z_0}{\sqrt{X_0^2+Y_0^2}} = -23.4^\circ\ .$$

If the values given by the *Astronomical Almanac* are rounded to the same accuracy, the same result is obtained. We should not expect a very precise position since we have neglected all short-period perturbations in Jupiter's orbital elements.

Example 6.8 Which is easier, to send a probe to the Sun or away from the Solar system?

The orbital velocity of the Earth is about 30 km/s. Thus the escape velocity from the Solar system is $\sqrt{2}\times 30 \approx 42$ km/s. A probe that is sent from the Earth already has a velocity equal to the orbital velocity of the Earth. Hence an extra velocity of only 12 km/s is needed. In addition, the probe has to escape from the Earth, which requires 11 km/s. Thus the total velocity changes are about 23 km/s.

If the probe has to fall to the Sun it has to get rid of the orbital velocity of the Earth 30 km/s. In this case, too, the probe has first to be lifted from the Earth. Thus the total velocity change needed is 41 km/s. This is nearly impossible with current technology. Therefore a probe to be sent to the Sun is first directed close to some planet, and the gravitational field of the planet is used to accelerate the probe towards its final destination.

Example 6.9 An interstellar hydrogen cloud contains 10 atoms per cm^3. How big must the cloud be to collapse due to its own gravitation? The temperature of the cloud is 100 K.

The mass of one hydrogen atom is 1.67×10^{-27} kg, which gives a density

$$\rho = n\mu = 10^7\ \text{m}^{-3}\times 1.67\times10^{-27}\ \text{kg}$$
$$= 1.67\times10^{-20}\ \text{kg/m}^3\ .$$

The critical mass is

$$M_\text{J} = \left(\frac{1.38\times10^{-23}\ \text{J/K}\times 100\ \text{K}}{1.67\times10^{-27}\ \text{kg}\times 6.67\times10^{-11}\ \text{N m}^2\,\text{kg}^{-2}}\right)^{3/2}$$
$$\times\ \frac{1}{\sqrt{1.67\times10^{-20}\ \text{kg/m}^3}}$$
$$\approx 1\times10^{34}\ \text{kg} \approx 5000\,M_\odot\ .$$

The radius of the cloud is

$$R = \sqrt[3]{\frac{3}{4\pi}\frac{M}{\rho}} \approx 5\times10^{17}\ \text{m} \approx 20\ \text{pc}\ .$$

6.13 Exercises

Exercise 6.1 Find the ratio of the orbital velocities at aphelion and perihelion v_a/v_p. What is this ratio for the Earth?

Exercise 6.2 The perihelion and aphelion of the orbit of Eros are 1.1084 and 1.8078 astronomical units from the Sun. What is the velocity of Eros when its distance from the Sun equals the mean distance of Mars?

Exercise 6.3 Find the radius of the orbit of a geostationary satellite; such a satellite remains always over the same point of the equator of the Earth. Are there areas on the surface of the Earth that cannot be seen from any geostationary satellite? If so, what fraction of the total surface area?

Exercise 6.4 From the angular diameter of the Sun and the length of the year, derive the mean density of the Sun.

Exercise 6.5 Find the mean, eccentric and true anomalies of the Earth one quarter of a year after the perihelion.

Exercise 6.6 The velocity of a comet is 5 m/s, when it is very far from the Sun. If it moved along a straight line, it would pass the Sun at a distance of 1 AU. Find the eccentricity, semimajor axis and perihelion distance of the orbit. What will happen to the comet?

Exercise 6.7 a) Find the ecliptic geocentric radius vector of the Sun on May 1, 1997 ($J = 2450570$).
b) What are the declination and right ascension of the Sun then?

7. The Solar System

The solar system consists of a central star, called *the Sun*, eight *planets*, several *dwarf planets*, dozens of *moons* or *satellites*, millions of *asteroids* and *Trans-Neptunian Objects* (TNOs), and myriads of *comets* and *meteoroids*.

Borders between the categories are not clear. Discoveries of new Solar System bodies caused that in 2006 the *International Astronomical Union* (IAU) in its General Assembly defined three distinct categories to clarify the situation:

(1) A *planet* is a celestial body that: (a) is in orbit around the Sun, (b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium (nearly round) shape, and (c) has cleared the neighbourhood around its orbit.

(2) A *dwarf planet* or a *planetoid* is a celestial body that: (a) is in orbit around the Sun, (b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium (nearly round) shape, (c) has not cleared the neighbourhood around its orbit, and (d) is not a satellite.

(3) All other objects orbiting the Sun shall be referred to collectively as *Small Solar System Bodies*. These include most of the asteroids, Trans-Neptunian Objects, comets, and other small bodies.

A *satellite* is a body which orbits the primary body so that the centre of mass (barycentre) is inside the primary. If this is not the case, then the system is called a *binary system*. For example, in the case of the Earth and Moon the barycentre of the system is inside the Earth, and the Moon is Earth's satellite. In the Pluto-Charon system the centre of mass is outside Pluto, and therefore they are called a binary system.

The planets in order from the Sun are: *Mercury*, *Venus*, *Earth*, *Mars*, *Jupiter*, *Saturn*, *Uranus*, and *Neptune*.

According to the IAU 2006 definition, *Pluto* is a *dwarf planet* and the prototype of a new category of Trans-Neptunian objects.

The planets from Mercury to Saturn are bright and well visible with a naked eye. Uranus and Neptune can be seen with a pair of binoculars. In addition to the bright planets, only the brightest comets are visible with a naked eye.

Distances in the solar system are often measured in *astronomical units* (AU), the mean distance of the Sun and Earth. The semimajor axis of the orbit of Mercury is 0.39 AU, and the distance of Neptune is 30 AU. Beyond the orbit of Neptune there is a huge population of small icy bodies extending out to tens of thousands AUs. The Solar System has no obvious outer edge. The distance to the nearest star, *Proxima Centauri* is over 270,000 AU.

Gravitation controls the motion of the solar system bodies. The planetary orbits around the Sun (Fig. 7.1) are almost coplanar ellipses which deviate only slightly from circles. The orbital planes of *asteroids*, minor bodies that circle the Sun mainly between the orbits of Mars and Jupiter, are often more tilted than the planes of the planetary orbits. Asteroids and distant Trans-Neptunian Objects revolve in the same direction as the major planets; comets, however, may move in the opposite direction. Cometary orbits can be very elongated, even hyperbolic. Most of the satellites circle their parent planets in the same direction as the planet moves around the Sun. Only the motions of the smallest particles, gas and dust are affected by the *solar wind*, *radiation pressure* and *magnetic fields*.

The planets can be divided into physically different groups (see Fig. Fig. 7.2). Mercury, Venus, Earth, and Mars are called *terrestrial* (Earth-like) planets; they have a solid surface, are of almost equal size (diameters from 5000 to 12,000 km), and have quite a high mean density ($4000-5000\ \mathrm{kg\,m^{-3}}$; the density of water is $1000\ \mathrm{kg\,m^{-3}}$). The planets from Jupiter to Neptune are called *Jovian* (Jupiter-like) or *giant planets*. The densities of the giant planets are about $1000-2000\ \mathrm{kg\ m^{-3}}$, and most of their volume is liquid. Diameters are ten times greater than those of the terrestrial planets.

Dwarf planet Pluto is falling outside this classification. Pluto is the prototype to the family of icy bodies orbiting the Sun at the outer edges of the solar system. The discovery of large objects since early 1990's beyond the orbit of Neptune raised the question of the status of Pluto. The discussion culminated in the General Assembly of the IAU in 2006 when a new planetary definition was accepted. This reduced the number of major planets to eight.

Most and most accurate solar system data are today collected by *spacecraft*. Many methods used in geosciences are nowadays applied in planetary studies. Landers have been sent to Moon, Venus, Mars, and Saturnian moon *Titan* and all major planets, their satellites, and many asteroids and comets have been studied with spacecraft.

Hannu Karttunen et al. (Eds.), *The Solar System*.
In: Hannu Karttunen et al. (Eds.), Fundamental Astronomy, 5th Edition. pp. 131–205 (2007)
DOI: 11685739_7 © Springer-Verlag Berlin Heidelberg 2007

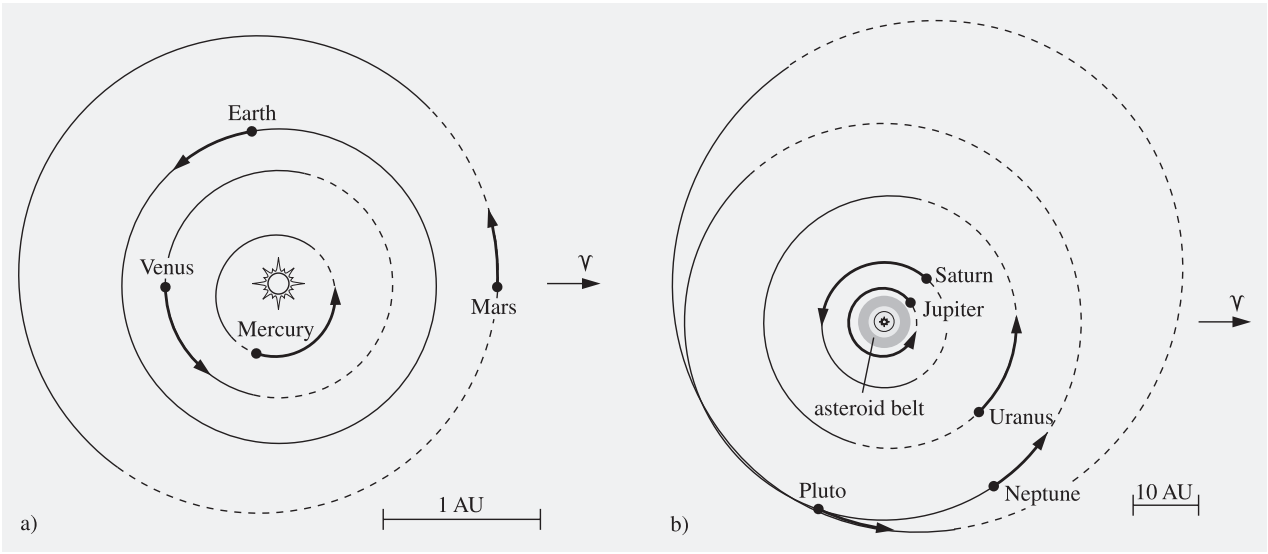


Fig. 7.1. (**a**) Planetary orbits from Mercury to Mars. The dashed line represents the part of the orbit below the ecliptic; the arrows show the distances travelled by the planets during one month (January 2000). (**b**) Planets from Jupiter to Neptune and the dwarf planet Pluto. The arrows indicate the distances travelled by the planets during the 10 year interval 2000–2010.

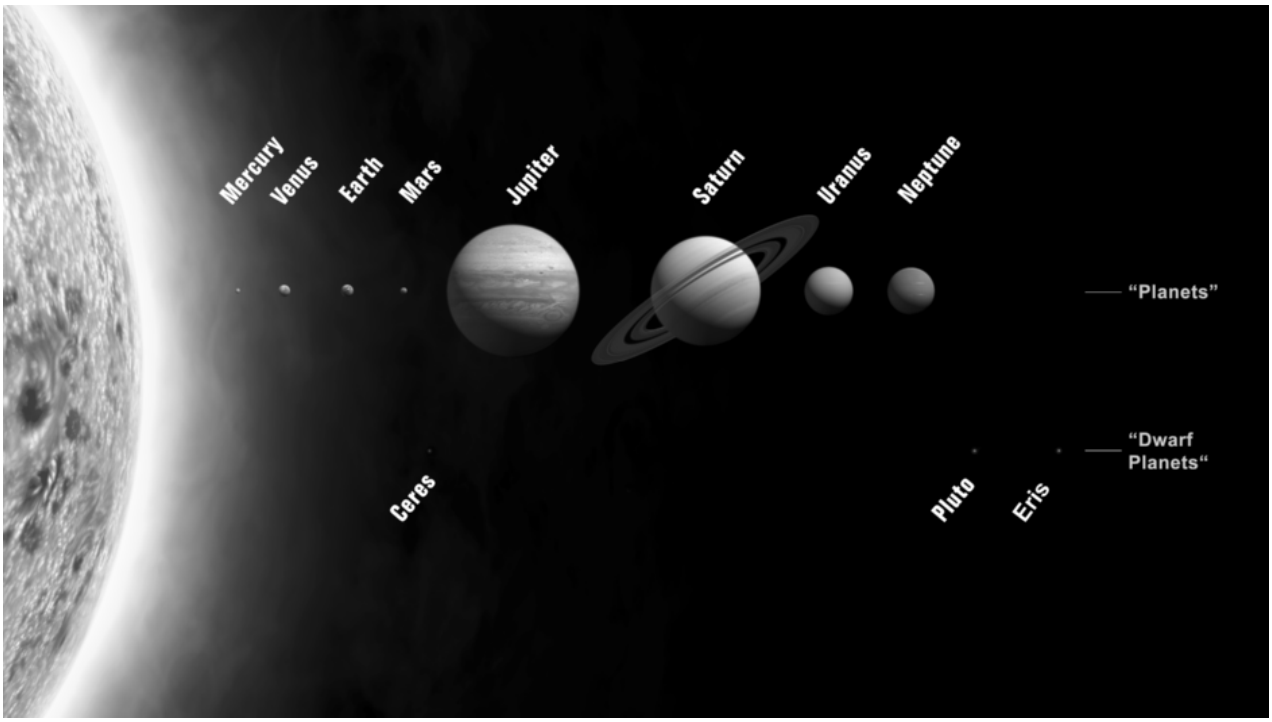


Fig. 7.2. Major planets from Mercury to Neptune. Four innermost planets are called terrestrial planets and four outermost ones are giant planets. Three dwarf planets are also shown. Relative size of the Sun is shown at left. Planetary distances to the Sun are not in scale. (The International Astronomical Union/Martin Kornmesser)

7.1 Planetary Configurations

The *apparent motions* of the planets are quite complicated, partly because they reflect the motion of the Earth around the Sun (Fig. 7.3). Normally the planets move eastward (*direct motion*, counterclockwise as seen from the Northern hemisphere) when compared with the stars. Sometimes the motion reverses to the opposite or *retrograde* direction. After a few weeks of retrograde motion, the direction is changed again, and the planet continues in the original direction. It is quite understandable that the ancient astronomers had great difficulties in explaining and modelling such complicated turns and loops.

Figure 7.4 explains some basic planetary configurations. A *superior planet* (planet outside the orbit of the Earth) is said to be in *opposition* when it is exactly opposite the Sun, i. e. when the Earth is between the planet and the Sun. When the planet is behind the Sun, it is in *conjunction*. In practise, the planet may not be exactly opposite or behind the Sun because the orbits of the planet and the Earth are not in the same plane. In astronomical almanacs oppositions and conjunctions are defined in terms of ecliptic longitudes. The longitudes of a body and the Sun differ by 180° at the moment of opposition; in conjunction the longitudes are equal. However, the right ascension is used if the other body is not the Sun. Those points at which the apparent motion of a planet turns toward the opposite direction are called *stationary points*. Opposition occurs in the middle of the retrograde loop.

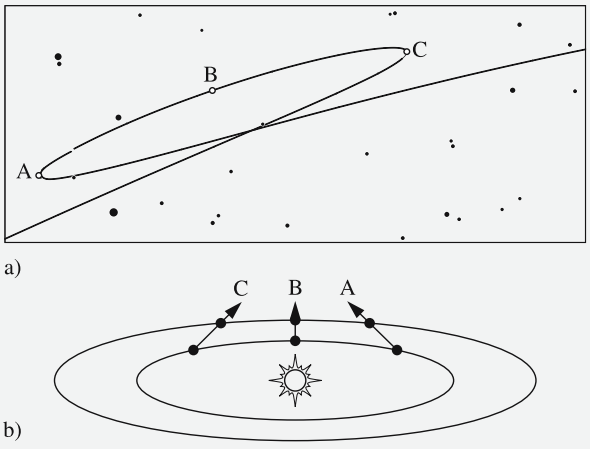


Fig. 7.3. (**a**) Apparent motion of Mars during the 1995 opposition. (**b**) Relative positions of the Earth and Mars. The projection of the Earth–Mars direction on the infinitely distant celestial sphere results in (**a**)

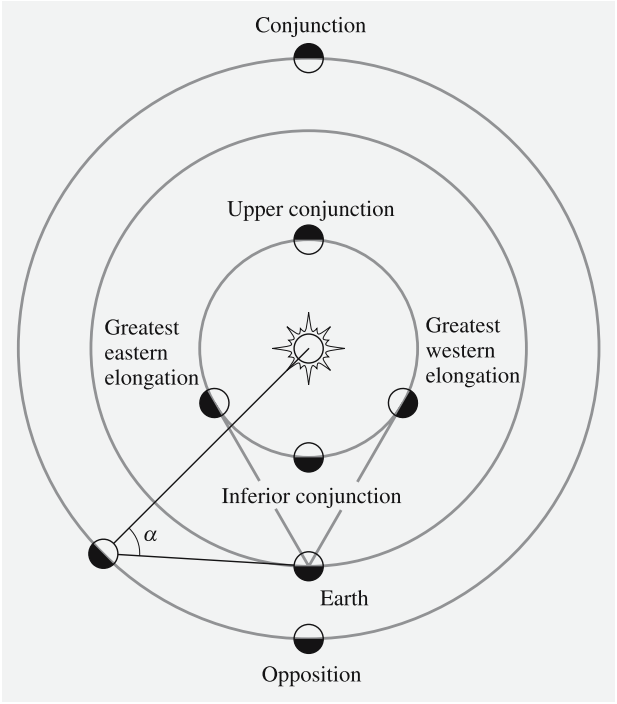


Fig. 7.4. Planetary configurations. The angle α (Sun–object–Earth) is the phase angle and ε (Sun–Earth–object) is the elongation

Inferior planets (Mercury and Venus) are never in opposition. The configuration occurring when either of these planets is between the Earth and the Sun is called *inferior conjunction*. The conjunction corresponding to that of a superior planet is called *upper conjunction* or *superior conjunction*. The maximum (eastern or western) *elongation*, i. e. the angular distance of the planet from the Sun is 28° for Mercury and 47° for Venus. Elongations are called eastern or western, depending on which side of the Sun the planet is seen. The planet is an "evening star" and sets after the Sun when it is in eastern elongation; in western elongation the planet is seen in the morning sky as a "morning star".

The *synodic period* is the time interval between two successive events (e. g. oppositions). The period which we used in the previous chapters is the *sidereal period*, the true time of revolution around the Sun, unique

for each object. The synodic period depends on the difference of the sidereal periods of two bodies.

Let the sidereal periods of two planets be P_1 and P_2 (assume that $P_1 < P_2$). Their mean angular velocities (mean motions) are $2\pi/P_1$ and $2\pi/P_2$. After one synodic period $P_{1,2}$, the inner planet has made one full revolution more than the outer planet:

$$P_{1,2}\frac{2\pi}{P_1} = 2\pi + P_{1,2}\frac{2\pi}{P_2} ,$$

or

$$\frac{1}{P_{1,2}} = \frac{1}{P_1} - \frac{1}{P_2} . \tag{7.1}$$

The angle Sun–planet–Earth is called the *phase angle*, often denoted by the Greek letter α. The phase angle is between 0° and 180° in the case of Mercury and Venus. This means that we can see "full Venus", "half Venus", and so on, exactly as in the phases of the Moon. The phase angle range for the superior planets is more limited. For Mars the maximum phase is 41°, for Jupiter 11°, and for Neptune only 2°.

7.2 Orbit of the Earth and Visibility of the Sun

The *sidereal year* is the real orbital period of the Earth around the Sun. After one sidereal year, the Sun is seen at the same position relative to the stars. The length of the sidereal year is 365.256363051 days of 86,400 SI seconds at the epoch J2000.0 = 2000 January 1 12:00:00 TT.

We noted earlier that, owing to precession, the direction of the vernal equinox moves along the ecliptic at about 50″ per year. This means that the Sun returns to the vernal equinox before one complete sidereal year has elapsed. This time interval, called the *tropical year*, is 365.24218967 days.

A third definition of the year is based on the perihelion passages of the Earth. Planetary perturbations cause a gradual change in the direction of the Earth's perihelion. The time interval between two perihelion passages is called the *anomalistic year*, the length of which is 365.259635864 days, a little longer than the sidereal year. It takes about 21,000 years for the perihelion to revolve 360° relative to the vernal equinox.

The equator of the Earth is tilted about 23.5° with respect to the ecliptic. Owing to perturbations, this angle changes with time. If periodic terms are neglected, the *obliquity of the ecliptic* ε can be expressed as:

$$\varepsilon = 23°26'21.448'' - 46.8150''T - 0.00059''T^2 + 0.001813''T^3 , \tag{7.2}$$

where T is the time elapsed since the epoch 2000.0 in Julian centuries (see Sect. 2.14). The expression is valid for a few centuries before and after the year 2000. The obliquity varies between 22.1° and 24.5° with a 41,000 year periodicity. At present the tilt is decreasing. There are also small short term variations, the *nutation*.

The declination of the Sun varies between $-\varepsilon$ and $+\varepsilon$ during the year. At any given time, the Sun is seen at zenith from one point on the surface of the Earth. The latitude of this point is the same as the declination of the Sun. At the latitudes $-\varepsilon$ (the *Tropic of Capricorn*) and $+\varepsilon$ (the *Tropic of Cancer*), the Sun is seen at zenith once every year, and between these latitudes twice a year. In the Northern hemisphere the Sun will not set if the latitude is greater than $90° - \delta$, where δ is the declination of the Sun.

The southernmost latitude where *the midnight Sun* can be seen is thus $90° - \varepsilon = 66.55°$. This is called the *Arctic Circle*. (The same holds true in the Southern hemisphere.) The Arctic Circle is the southernmost place where the Sun is (in theory) below the horizon during the whole day at the winter solstice. The sunless time lasts longer and longer when one goes north (south in the Southern hemisphere). At the poles, day and night last half a year each. In practise, refraction and location of the observing site will have a large influence on the visibility of the midnight Sun and the number of sunless days. Because refraction raises objects seen at the horizon, the midnight Sun can be seen a little further south than at the Arctic Circle. For the same reason the Sun can be seen simultaneously at both poles around the time of vernal and autumnal equinox.

The eccentricity of the Earth's orbit is about 0.0167. The distance from the Sun varies between 147–152 million km. The flux density of solar radiation varies somewhat at different parts of the Earth's orbit, but this has practically no effect on the seasons. In fact the Earth is at perihelion in the beginning of January, in the middle of the northern hemisphere's winter. The seasons are due to the obliquity of the ecliptic.

The energy received from the Sun depends on three factors. First the flux per unit area is proportional to

$\sin a$, where a is the altitude of the Sun. In summer the altitude can have greater values than in winter, giving more energy per unit area. Another effect is due to the atmosphere: When the Sun is near the horizon, the radiation must penetrate thick atmospheric layers. This means large extinction and less radiation at the surface. The third factor is the length of the time the Sun is above the horizon. This is important at high latitudes, where the low altitude of the Sun is compensated by the long daylight time in summer. These effects are discussed in detail in Example 7.2.

There are also long-term variations in the annual Solar flux. Serbian geophysicist Milutin Milanković (1879–1958) published in the 1930's and 1940's his theory of ice ages. During last 2–3 million years, large ice ages have been repeated approximately every 100,000 years. He proposed that variations of the Earth's orbit cause long-term periodic climate change, now known as *Milanković cycles*. Milanković claimed that the cycles in eccentricity, direction of the perigee, obliquity, and precession result in 100,000 year ice age cycle. The cycle of precession is 26,000 years, direction of the perigee relative to the equinoxes is 22,000 years, and the obliquity of the ecliptic has a 41,000 year cycle. Changes in orbital eccentricity are not fully periodic but some periods above 100,000 years can be found. The eccentricity varies between 0.005–0.058 and is currently 0.0167.

The annual incoming Solar flux varies with these orbital changes and the effect is largest at high latitudes. If, for example, the eccentricity is high, and the Earth is near the apogee during the hemisphere's winter, then winters are long and cold and summers are short. However, the theory is controversial, orbital forcing on the climate change is not well understood, and probably not enough to trigger glaciation. There exist also positive feedback loops, like the effect of low albedo of snow and ice. It means that ice reflects more radiation back into space, thus cooling the climate. The system is highly chaotic so that even minor changes in the primary conditions will result in large differences in the outcome. There are also other effects causing climate change, including emerging gases from large lava flows and eruptions of volcanos and, nowadays, anthropogenic reasons.

The future is also uncertain. Some theories predict that the warm period will continue next 50,000 years, whereas others conclude that the climate is already cooling. Anthropogenic reasons, like ever increasing fraction of green house gases, e.g. carbon dioxide, will change the short-term predictions.

7.3 The Orbit of the Moon

The Earth's satellite, *the Moon*, circles the Earth counterclockwise. One revolution, the *sidereal month*, takes about 27.322 days. In practise, a more important period is the *synodic month*, the duration of the Lunar phases (e. g. from full moon to full moon). In the course of one sidereal month the Earth has travelled almost $1/12$ of its orbit around the Sun. The Moon still has about $1/12$ of its orbit to go before the Earth–Moon–Sun configuration is again the same. This takes about 2 days, so the phases of the Moon are repeated every 29 days. More exactly, the length of the synodic month is 29.531 days.

The *new moon* is that instant when the Moon is in conjunction with the Sun. Almanacs define the phases of the Moon in terms of ecliptic longitudes; the longitudes of the new moon and the Sun are equal. Usually the new moon is slightly north or south of the Sun because the lunar orbit is tilted 5° with respect to the ecliptic.

About 2 days after the new moon, the waxing crescent moon can be seen in the western evening sky. About 1 week after the new moon, the *first quarter* follows, when the longitudes of the Moon and the Sun differ by 90°. The right half of the Moon is seen lit (left half when seen from the Southern hemisphere). The *full moon* appears a fortnight after the new moon, and 1 week after this the *last quarter*. Finally the waning crescent moon disappears in the glory of the morning sky.

The orbit of the Moon is approximately elliptic. The length of the semimajor axis is 384,400 km and the eccentricity 0.055. Owing to perturbations caused mainly by the Sun, the orbital elements vary with time. The minimum distance of the Moon from the centre of the Earth is 356,400 km, and the maximum distance 406,700 km. This range is larger than the one calculated from the semimajor axis and the eccentricity. The apparent angular diameter is in the range $29.4'$–$33.5'$.

The rotation time of the Moon is equal to the sidereal month, so the same side of the Moon always faces the Earth. Such *synchronous rotation* is common among

Fig. 7.5. Librations of the Moon can be seen in this pair of photographs taken when the Moon was close to the perigee and the apogee, respectively. (Helsinki University Observatory)

the satellites of the solar system: almost all large moons rotate synchronously.

The orbital speed of the Moon varies according to Kepler's second law. The rotation period, however, remains constant. This means that, at different phases of the lunar orbit, we can see slightly different parts of the surface. When the Moon is close to its perigee, its speed is greater than average (and thus greater than the mean rotation rate), and we can see more of the right-hand edge of the Moon's limb (as seen from the Northern hemisphere). Correspondingly, at the apogee we see "behind" the left edge. Owing to the *libration*, a total of 59% of the surface area can be seen from the Earth (Fig. 7.5). The libration is quite easy to see if one follows some detail at the edge of the lunar limb.

The orbital plane of the Moon is tilted only about 5° to the ecliptic. However, the orbital plane changes gradually with time, owing mainly to the perturbations caused by the Earth and the Sun. These perturbations cause the nodal line (the intersection of the plane of the ecliptic and the orbital plane of the Moon) to make one full revolution in 18.6 years. We have already encountered the same period in the nutation. When the ascending node of the lunar orbit is close to the vernal equinox, the Moon can be $23.5° + 5° = 28.5°$ north or south of the equator. When the descending node is close to the vernal equinox, the zone where the Moon can be found extends only $23.5° - 5° = 18.5°$ north or south of the equator.

The *nodical* or *draconic month* is the time in which the Moon moves from one ascending node back to the next one. Because the line of nodes is rotating, the nodical month is 3 hours shorter than the sidereal month, i. e. 27.212 days. The orbital ellipse itself also precesses slowly. The orbital period from perigee to perigee, the *anomalistic month*, is 5.5 h longer than the sidereal month, or about 27.555 days.

Gravitational differences caused by the Moon and the Sun on different parts of the Earth's surface give rise to the *tides*. Gravitation is greatest at the sub-lunar point and smallest at the opposite side of the Earth. At

these points, the surface of the seas is highest (high tide, *flood*). About 6 h after flood, the surface is lowest (low tide, *ebb*). The tide generated by the Sun is less than half of the lunar tide. When the Sun and the Moon are in the same direction with respect to the Earth (new moon) or opposite each other (full moon), the tidal effect reaches its maximum; this is called *spring tide*.

The sea level typically varies 1 m, but in some narrow straits, the difference can be as great as 15 m. Due to the irregular shape of the oceans, the true pattern of the oceanic tide is very complicated. The solid surface of the Earth also suffers tidal effects, but the amplitude is much smaller, about 30 cm.

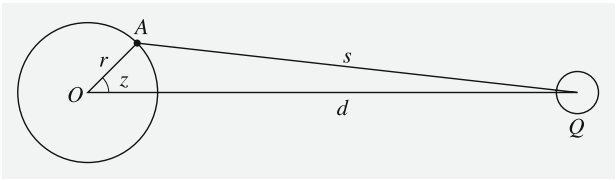
Tides generate friction, which dissipates the rotational and orbital kinetic energy of the Earth–Moon system. This energy loss induces some changes in the system. First, the rotation of the Earth slows down until the Earth also rotates synchronously, i. e. the same side of Earth will always face the Moon. Secondly, the semi-major axis of the orbit of the Moon increases, and the Moon drifts away about 3 cm per year.

* Tides

Let the tide generating body, the mass of which is M to be at point Q at a distance d from the centre of the Earth. The potential V at the point A caused by the body Q is

$$V(A) = \frac{GM}{s} \,, \tag{7.3}$$

where s is the distance of the point A from the body Q.



Applying the cosine law in the triangle OAQ, the distance s can be expressed in terms of the other sides and the angle $z = AOQ$

$$s^2 = d^2 + r^2 - 2dr\cos z \,,$$

where r is the distance of the point A from the centre of the Earth. We can now rewrite (7.3)

$$V(A) = \frac{GM}{\sqrt{d^2 + r^2 - 2dr\cos z}} \,. \tag{7.4}$$

When the denominator is expanded into a Taylor series

$$(1+x)^{-\frac{1}{2}} \approx 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \cdots$$

where

$$x = \frac{r^2}{d^2} - 2\frac{r}{d}\cos z$$

and ignoring all terms higher than or equal to $1/d^4$ one obtains

$$\begin{aligned} V(A) = {} & \frac{GM}{d} + \frac{GM}{d^2} r\cos z \\ & + \frac{GMr^2}{d^3}\frac{1}{2}(3\cos^2 z - 1) \,. \end{aligned} \tag{7.5}$$

The gradient of the potential $V(A)$ gives a force vector per mass unit. The first term of (7.5) vanishes, and the second term is a constant and independent of r. It represents the central motion. The third term of the force vector, however, depends on r. It is the main term of the tidal force. As one can see, it depends inversely on the third power of the distance d. The tidal forces are diminished very rapidly when the distance of a body increases. Therefore the tidal force caused by the Sun is less than half of that of the Moon in spite of much greater mass of the Sun.

We may rewrite the third term of (7.5) as

$$V_2 = 2D\left(\cos^2 z - \frac{1}{3}\right) \,, \tag{7.6}$$

where

$$D = \frac{3}{4}GM\frac{r^2}{d^3}$$

is called *Doodson's tidal constant*. It's value for the Moon is $2.628\ \mathrm{m^2\ s^{-2}}$ and for the Sun $1.208\ \mathrm{m^2\ s^{-2}}$. We can approximate that z is the zenith angle of the body. The zenith angle z can be expressed in terms of the hour angle h and declination δ of the body and the latitude ϕ of the observer

$$\cos z = \cos h\cos\delta\cos\phi + \sin\delta\sin\phi \,.$$

Inserting this into (7.6) we obtain after a lengthy algebraic operation

$$\begin{aligned} V_2 = D\Bigg(& \cos^2\phi \cos^2\delta \cos 2h \\ & + \sin 2\phi \cos 2\delta \cos h \\ & + (3\sin^2\phi - 1)\left(\sin^2\delta - \frac{1}{3}\right)\Bigg) \\ = D(S+T+Z)\,. & \end{aligned} \tag{7.7}$$

Equation (7.7) is the traditional basic equation of the tidal potential, the *Laplace's tidal equation*.

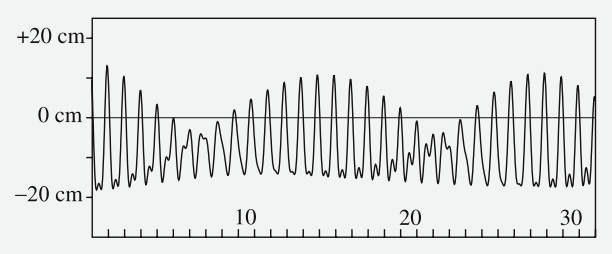
In (7.7) one can directly see several characteristics of tides. The term S causes the *semi-diurnal tide* because it depends on $\cos 2h$. It has two daily maxima and minima, separated by 12 hours, exactly as one can obtain in following the ebb and flood. It reaches its maximum at the equator and is zero at the poles ($\cos^2\phi$).

The term T expresses the *diurnal tides* ($\cos h$). It has its maximum at the latitude $\pm 45°$ and is zero at the equator and at the poles ($\sin 2\phi$). The third term Z is independent of the rotation of the Earth. It causes the *long period tides*, the period of which is half the orbital period of the body (about 14 days in the case of the Moon and 6 months for the Sun). It is zero at the latitude $\pm 35.27°$ and has its maximum at the poles. Moreover, the time average of Z is non-zero, causing a permanent deformation of the Earth. This is called the *permanent tide*. It slightly increases the *flattening of the Earth* and it is inseparable from the flattening due to the rotation.

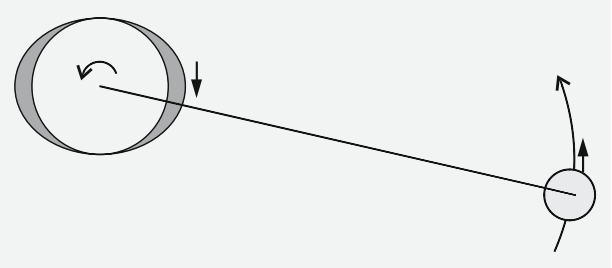
The total value of the tidal potential can be computed simply adding the potentials caused by the Moon and the Sun. Due to the tidal forces, the whole body of the Earth is deformed. The vertical motion Δr of the crust can be computed from

$$\Delta r = h\frac{V_2}{g} \approx 0.06\, V_2\,[\mathrm{m}]\,, \tag{7.8}$$

where g is the mean free fall acceleration, $g \approx 9.81\,\mathrm{m\,s^{-2}}$ and h is a dimensionless number, the *Love number*, $h \approx 0.6$, which describes the elasticity of the Earth. In the picture below, one can see the vertical motion of the crust in Helsinki, Finland ($\phi = 60°$, $\lambda = 25°$) in January 1995. The non-zero value of the temporal mean can already be seen in this picture.



The tides have other consequences, too. Because the Earth rotates faster than the Moon orbits the Earth, the tidal bulge does not lie on the Moon–Earth line but is slightly ahead (in the direction of Earth's rotation), see below.



Due to the drag, the rotation of the Earth slows down by about 1–2 ms per century. The same reason has caused the Moon's period of rotation to slow down to its orbital period and the Moon faces the same side towards the Earth. The misaligned bulge pulls the Moon forward. The acceleration causes the increase in the semimajor axis of the Moon, about 3 cm per year.

7.4 Eclipses and Occultations

An *eclipse* is an event in which a body goes through the shadow of another body. The most frequently observed eclipses are the lunar eclipses and the eclipses of the large satellites of Jupiter. An *occultation* takes place when an occulting body goes in front of another object; typical examples are stellar occultations caused by the Moon. Generally, occultations can be seen only in a narrow strip; an eclipse is visible wherever the body is above the horizon.

Solar and lunar eclipses are the most spectacular events in the sky. A *solar eclipse* occurs when the Moon is between the Earth and the Sun (Fig. 7.6). (According

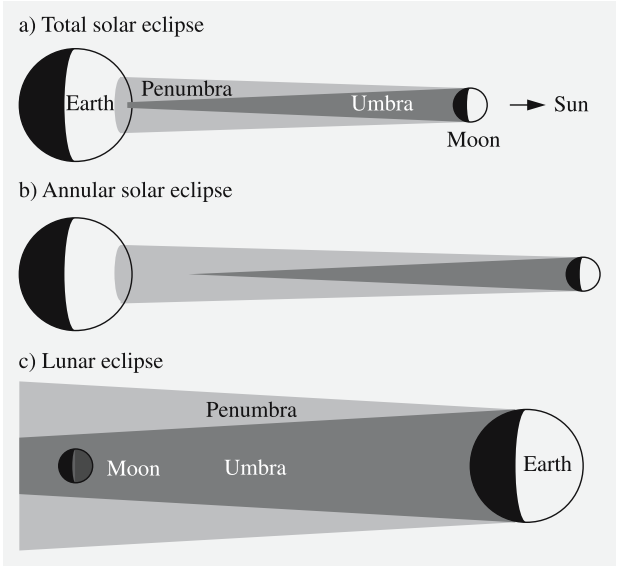


Fig. 7.6. (**a**) A total solar eclipse can be seen only inside a narrow strip; outside the zone of totality the eclipse is partial. (**b**) An eclipse is annular if the Moon is at apogee from where the shadow of the Moon does not reach the Earth. (**c**) A lunar eclipse is visible everywhere where the Moon is above the horizon

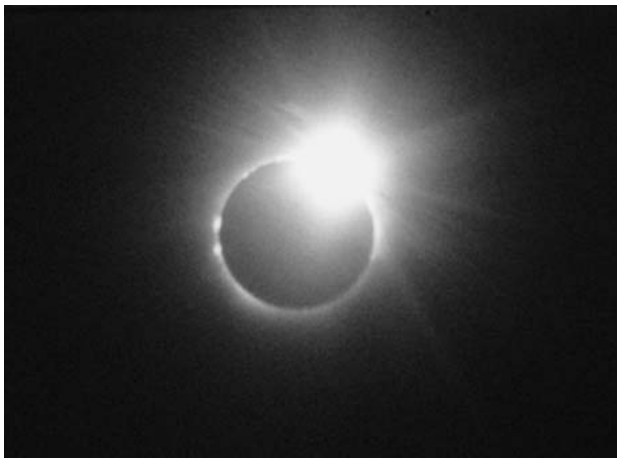


Fig. 7.7. The total eclipse of the Sun occurred in 1990 over Finland. (Photo Matti Martikainen)

to the definition, a solar eclipse is not an eclipse but an occultation!) If the whole disk of the Sun is behind the Moon, the eclipse is *total* (Fig. 7.7); otherwise, it is *partial*. If the Moon is close to its apogee, the apparent diameter of the Moon is smaller than that of the Sun, and the eclipse is *annular*.

A lunar eclipse is *total* if the Moon is entirely inside the umbral shadow of the Earth; otherwise the eclipse is *partial*. A partial eclipse is difficult to see with the unaided eye because the lunar magnitude remains almost unchanged. During the total phase the Moon is coloured deep red because some red light is refracted through the Earth's atmosphere.

If the orbital plane of the Moon coincided with the plane of the ecliptic, one solar and one lunar eclipse would occur every synodic month. However, the plane is tilted about 5°; therefore, at full moon, the Moon must be close to the nodes for an eclipse to occur. The angular distance of the Moon from the node must be smaller than 4.6° for a total lunar eclipse, and 10.3° for a total solar eclipse.

Two to seven eclipses occur annually. Usually eclipses take place in a set of 1–3 eclipses, separated by an interval of 173 days. In one set there can be just one solar eclipse or a succession of solar, lunar and another solar eclipse. In one year, eclipses belonging to 2 or 3 such sets can take place.

The Sun and the (ascending or descending) node of the lunar orbit are in the same direction once every 346.62 days. Nineteen such periods (= 6585.78 days = 18 years 11 days) are very close to the length of 223 synodic months. This means that the Sun–Moon configuration and the eclipses are repeated in the same order after this period. This *Saros* period was already known to the ancient Babylonians.

During a solar eclipse the shadow of the Moon on Earth's surface is always less than 270 km wide. The shadow moves at least 34 km/min; thus the maximum duration of an eclipse is $7\frac{1}{2}$ minutes. The maximum duration of a lunar eclipse is 3.8 h, and the duration of the total phase is always shorter than 1.7 h.

Observations of the *stellar occultations* caused by the Moon formerly served as an accurate method for determining the lunar orbit. Because the Moon has no atmosphere, the star disappears abruptly in less than 1/50 s. If a fast photometer is used for recording the event, the typical diffraction pattern can be seen. The shape of the diffraction is different for a binary star. In the first decades of radio astronomy the occultations of some radio sources were used for determining their exact positions.

The Moon moves eastwards, and stars are occulted by the dark edge of the Moon during the first quarter.

Therefore occultation is easier to observe, and photometric measurements are possible; at the same time it is much more difficult to observe the appearance of an object. There are some bright stars and planets inside the 11° wide zone where the Moon moves, but the occultation of a bright, naked-eye object is quite rare.

Occultations are also caused by planets and asteroids. Accurate predictions are complicated because such an event is visible only in a very narrow path. The Uranian rings were found during an occultation in 1977, and the shapes of some asteroids have been studied during some favourable events, timed exactly by several observers located along the predicted path.

A *transit* is an event in which Mercury or Venus moves across the Solar disk as seen from the Earth. A transit can occur only when the planet is close to its orbital node at the time of inferior conjunction. Transits of Mercury occur about 13 times per century; transits of Venus only twice. The next transits of Mercury are: May 9, 2016; Nov 11, 2019; Nov 13, 2032 and Nov 7, 2039. The next transits of Venus are: Jun 6, 2012; Dec 11, 2117; Dec 8, 2125 and Jun 11, 2247. In the 18th century the two transits of Venus (1761 and 1769) were used for determining the value of the astronomical unit.

7.5 The Structure and Surfaces of Planets

Since the 1960's a vast amount of data have been collected using spacecraft, either during a flyby, orbiting a body, or directly landing on the surface. This gives a great advantage compared to other astronomical observations. We may even speak of revolution: the solar system bodies have turned from astronomical objects to geophysical ones. Many methods traditionally used in various sibling branches of geophysics can now be applied to planetary studies.

The shape and irregularities of the gravitation field generated by a planet reflects its shape, internal structure and mass distribution. Also the surface gives certain indications on internal structure and processes.

The perturbations in the orbit of a satellite or spacecraft can be used in studying the internal structure of a planet. Any deviation from spherical symmetry is visible in the external gravitational field.

The IAU planet definition states that planets are bodies in *hydrostatic equilibrium*. Gravity of a body will pull its material inwards, but the body resist the pull if the strength of the material is greater than the pressure exerted by the overlying layers. If the diameter is larger than about 800–1000 km, gravity is able to deform rocky bodies into spherical shape. Smaller bodies than this have irregular shapes. On the other hand, e.g. icy moons of Saturn are spherical because ice is more easily deformed than rock.

Hydrostatic equilibrium means that the surface of the body approximately follows an equipotential surface of gravity. This is true e.g. on the Earth, where the sea surface very closely follows the equipotential surface called the *geoid*. Due to internal strength of rocks, continents can deviate from the geoid surface by a few kilometers but compared to the diameter of the Earth the surface topography is negligible.

A rotating planet is always *flattened*. The amount of flattening depends on the rotation rate and the strength of the material; a liquid drop is more easily deformed than a rock. The shape of a rotating body in hydrostatic equilibrium can be derived from the equations of motion. If the rotation rate is moderate, the equilibrium shape of a liquid body is an ellipsoid of revolution. The shortest axis is the axis of rotation.

If R_e and R_p are the equatorial and polar radii, respectively, the shape of the planet can be expressed as

$$\frac{x^2}{R_\mathrm{e}^2} + \frac{y^2}{R_\mathrm{e}^2} + \frac{z^2}{R_\mathrm{p}^2} = 1 \,.$$

The *dynamical flattening*, denoted by f is defined as

$$f = \frac{R_\mathrm{e} - R_\mathrm{p}}{R_\mathrm{e}} \,. \tag{7.9}$$

Because $R_\mathrm{e} > R_\mathrm{p}$, the flattening f is always positive.

The giant planets are in practise close to hydrostatic equilibrium, and their shape is determined by the rotation. The rotation period of Saturn is only 10.5 h, and its dynamical flattening is $1/10$ which is easily visible.

Asteroids and other minor bodies are so small that they are not flattened by rotation. However, there is an upper limit for a rotation rate of an asteroid before it breaks apart due to centrifugal forces. If we assume that the body is held together only by gravity, we can approximate the the maximum rotation rate by setting

the centrifugal force equal to the gravitational force:

$$\frac{GMm}{R^2} = \frac{mv^2}{R}, \tag{7.10}$$

where m is a small test mass on the surface at a distance of R from the center of the body. Substituting the rotation period P,

$$P = \frac{2\pi R}{v},$$

we get

$$\frac{GM}{R^2} = \frac{4\pi^2 R}{P^2},$$

or

$$P = 2\pi\sqrt{\frac{R^3}{GM}} = 2\pi\sqrt{\frac{3}{4\pi G\rho}} = \sqrt{\frac{3\pi}{G\rho}}. \tag{7.11}$$

If we substitute the density ρ with the mean density of terrestrial rocks, i.e. 2700 kg m^{-3}, we get for the minimum rotation period $P \approx 2$ hours.

The structure of the terrestrial planets (Fig. 7.8) can also be studied with *seismic waves*. The waves formed in an earthquake are reflected and refracted inside a planet like any other wave at the boundary of two different layers. The waves are longitudinal or transversal (*P* and *S waves*, respectively). Both can propagate in solid materials such as rock. However, only the longitudinal wave can penetrate liquids. One can determine whether a part of the interior material is in the liquid state and where the boundaries of the layers are by studying the recordings of seismometers placed on the surface of a planet. Naturally the Earth is the best-known body, but quakes of the Moon, Venus, and Mars have also been observed.

The terrestrial planets have an *iron-nickel core*. Mercury has the relatively largest core; Mars the smallest. The core of the Earth can be divided into an *inner* and an *outer core*. The outer core (2900–5150 km) is liquid but the inner core (from 5150 km to the centre) is solid.

Around the Fe–Ni core is a *mantle*, composed of *silicates* (compounds of silicon). The density of the outermost layers is about 3000 kg m^{-3}. The mean density of the terrestrial planets is 3500–5500 kg m^{-3}.

The internal structure of the giant planets (Fig. 7.9) cannot be observed with seismic waves since the planets do not have a solid surface. An alternative is to study the shape of the gravitational field by observing the orbit of a spacecraft when it passes (or orbits) the planet. This will give some information on the internal structure, but the details depend on the mathematical and physical models used for interpretation.

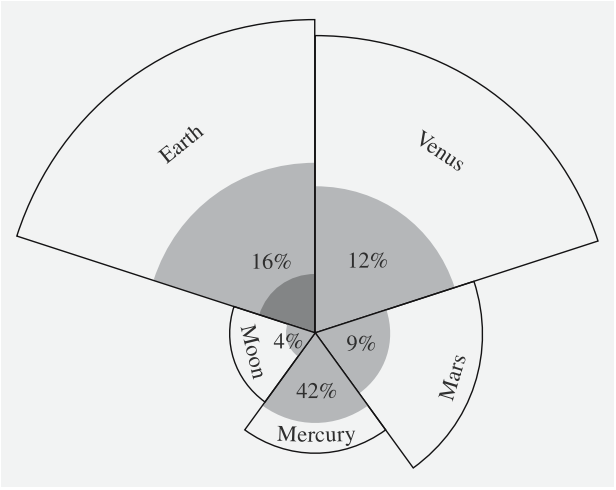


Fig. 7.8. Internal structure and relative sizes of the terrestrial planets. The percentage shows the volume of the core relative to the total volume of the planet. In the case of the Earth, the percentage includes both the outer and the inner core

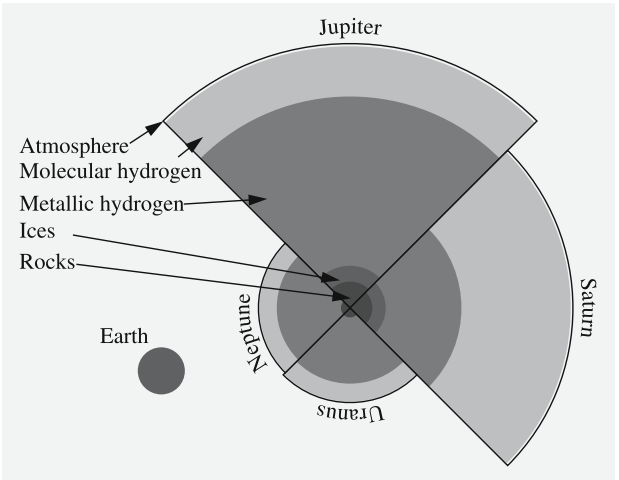


Fig. 7.9. Internal structure and relative sizes of the giant planets. Differences in size and distance from the Sun cause differences in the chemical composition and internal structure. Due to smaller size, Uranus and Neptune do not have any layer of metallic hydrogen. The Earth is shown in scale

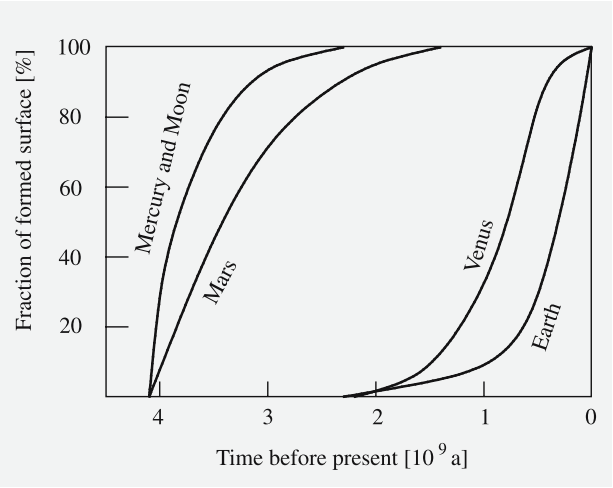


Fig. 7.10. Ages of the surfaces of Mercury, the Earth, the Moon and Mars. The curve represents the fraction of the surface which existed at a certain time. Most of the surface of the Moon, Mercury and Mars are more than 3500 million years old, whereas the surface of the Earth is mostly younger than 200 million years

The mean densities of the giant planets are quite low; the density of Saturn, for example, is only 700 kg m^{-3}. (If Saturn were put in a gigantic bathtub, it would float on the water!) Most of the volume of a giant planet is a mixture of hydrogen and helium. In the centre, there is possibly a silicate core, the mass of which is a few Earth masses. The core is surrounded by a layer of *metallic hydrogen*. Due to the extreme pressure, hydrogen is not in its normal molecular form H_2, but dissociated into atoms. In this state, hydrogen is electrically conducting. The magnetic fields of the giant planets may originate in the layer of metallic hydrogen.

Closer to the surface, the pressure is lower and hydrogen is in molecular form. The relative thickness of the layers of metallic and molecular hydrogen vary from planet to planet. Uranus and Neptune may not have any layer of metallic hydrogen because their internal pressure is too low for dissociation of the hydrogen. Instead, a layer of "ices" surround the core. This is a layer of a water-dominant mixture of water, methane and ammonia. Under the high pressure and temperature the mixture is partly dissolved into its components and it

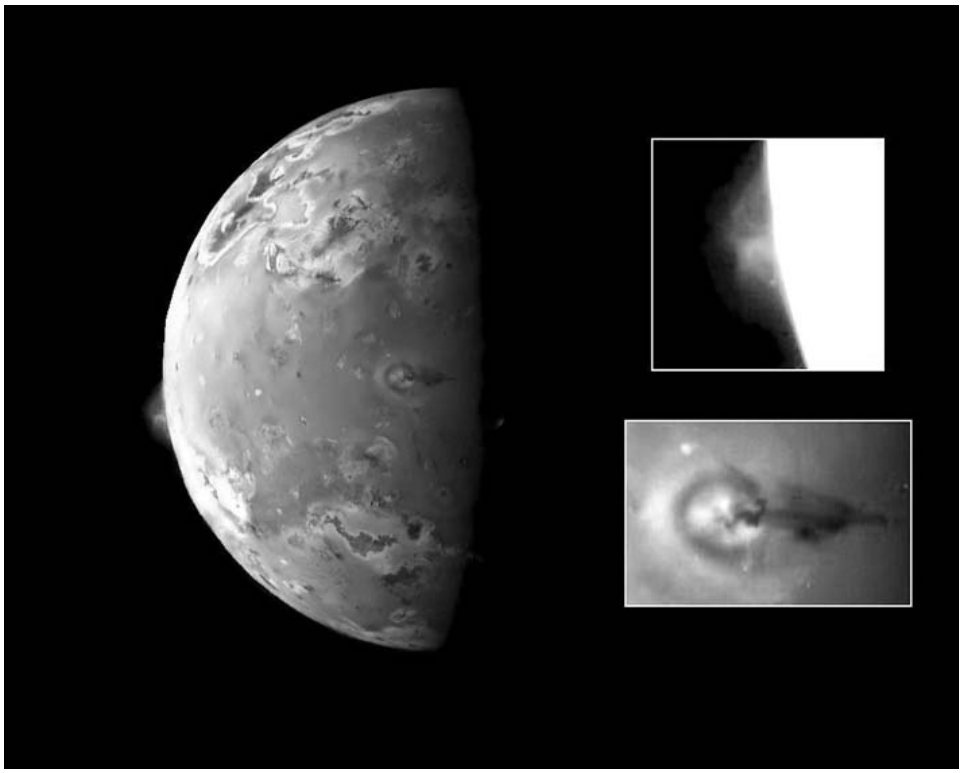


Fig. 7.11. An example of resurfacing. Two volcanic plumes on Jupiter's moon Io observed by Galileo spacecraft in 1997. One plume was captured on the bright limb or edge of the moon (inset at upper right), erupting over a caldera named Pillan Patera. The plume is 140 kilometers high. The second plume, seen near the terminator, is called Prometheus. The shadow of the 75 km high airborne plume can be seen extending to the right of the eruption vent. (NASA/JPL)