T.L. Wilson
K. Rohlfs
S. Hüttemeister

Tools of
Radio Astronomy

Fifth Edition

Springer

ASTRONOMY AND ASTROPHYSICS LIBRARY

Thomas L. Wilson · Kristen Rohlfs ·
Susanne Hüttemeister

Tools of Radio Astronomy

Fifth Edition

Thomas L. Wilson
European Southern
Observatory (ESO)
Karl-Schwarzschild-Str. 2
85748 Garching
Germany
twilson@eso.org

Kristen Rohlfs
Ruhr-University-Bochum
Inst. Astrophysik
44721 Bochum
Germany
Kristen.Rohlfs@t-online.de

Susanne Hüttemeister
Zeiss Planetarium Bochum
Castroper Str. 68
44777 Bochum
Germany
huettemeister@planetarium-bochum.de

Cover image: The Japanese MELCO 12 meter antennas on the ALMA Operation Support Facility. Credit: (c) ALMA / ESO / NAOJ / NRAO.

Cover design: eStudio Calamar S.L.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface to the Fifth Edition

Four significant factors have led us to update this text. The first is the breathtaking progress in technology, especially in receiver and digital techniques. The second is the advance of radio astronomy to shorter wavelengths, and the increased availability of astronomical satellites. The third is a need to reorganize some of the chapters in order to separate the basic theory, that seldom changes, from practical aspects that change often. Finally, it is our desire to enhance the text by including problem sets for each chapter. In view of this ambitious plan, we have expanded the number of authors.

In the reorganization of this edition, we have divided Chap. 4 of the 4th edition into two Chaps. 4 and 5. The first remains Chap. 4, with a slightly different title, *Signal Processing and Receivers: Theory*. This was expanded to include digital processing and components including samplers and digitizers. In Chap. 5, *Practical Receiver Systems*. we have relegated the presentations of maser and parametric amplifier front ends, which are no longer commonly used as microwave receivers in radio astronomy, to a short section on "historical developments" and We have retained and improved the presentations of current state-of-the-art devices, cooled transistor and superconducting front ends. We have also included descriptions of local oscillators and phase lock loops. Chapters 5 and 6 in the 4th edition has now become Chap. 6, *Fundamentals of Antenna Theory* and Chap. 7, *Practical Aspects of Filled Aperture Antennas*. Our goal is to have an exposition of the rather mathematical theory, in Chap. 6 followed by a treatment of the practical aspects of antennas. Chapter 7 in the 4th edition is now Chap. 8, titled *Single Dish Observational Methods*. Chapter 9 deals with *Interferometers and Aperture Synthesis*. Aperture synthesis has become the most important imaging technique in radio astronomy; this provides the only general method available for obtaining images of extremely high resolution and quality, so the discussion has been extended and improved with material pertenant to interferometers such as the Atacama Large Millimeter Array (ALMA) and the Square Kilometer Array (SKA). Chapters 10 to 14 of this edition have been updated to include recent observational results. Chapter 15 of the 4th edition, *Molecules in Interstellar Space*, has been divided into two Chapters, *Overview of Molecular Basics* and Chap. 16, *Molecules in the Interstellar Medium*. Chapters 15 and 16 have been updated to take new developments into account.

The existing facilities are providing new results on a daily basis. The increased number of ground based radio single dish telescopes, especially in the millimeter and sub-mm wavelength range, such as ASTE, APEX, and NANTEN2, and the availability of astronomical satellites starting with IRAS, and then ISO, ODIN, MSX, CHANDRA and SPITZER have increased the number of discoveries. Somewhat more specialized are the radio telescopes dedicated to the study of the 3 K microwave background: these include the satellite missions COBE and WMAP and the balloon mission Boomerang, as well as numerous additional ground based facilities. Taken together, these have changed our concepts of astronomy. A sample of these results have been included. This trend is expected to continue with the launch of the *Herschel Satellite Observatory* and the start of scientific measurements with the *Stratospheric Observatory for Infrared Astronomy*, SOFIA.

We believe that this text is of interest for communications engineers as well as radio astronomers. We hope this new edition will serve a useful purpose as radio astronomy enters the era of Herschel, SOFIA, ALMA, SKA, SKA precursors.

The Table of interstellar molecules was provided by T. Millar (Queen's University Belfast) & E. Herbst (Ohio State University). Advice from G. H. Tan, H. Rudolf, R. Laing (all ESO) and A. Veronig (Graz University), W. Alef (MPIfR, Bonn), A. Clegg (NSF), D. Boboltz (USNO) and A. Fey (USNO) is greatfully acknowledged. We thank E. Janssen, J. Howard and M. Martins (ESO) who provided new or updated figures for this edition. As in previous editions, we have corrected a number of errors in the text. Most of these were kindly provided by J. J. Condon (NRAO), A. Guzmann (Chile) and Biwei Jiang (Peking).

Web sites are a new mode of communicating recent results. However we have limited our references to these as much as possible since the addresses change often. A remark about nomenclature: in the index, we have (with some arbitrariness) ordered single radio telescopes under *antennas*, arrays of antennas with coupled outputs under *interferometers* and facilities such as Herschel and SOFIA under their names.

Munich, Bonn and Bochum   *T. L. Wilson*
September 2008   *Kristen Rohlfs*
                                                                      S. Hüttemeister

Excerpts from the Prefaces of Previous Editions

This book describes the tools radio astronomers need to pursue their goals. These tools consist of: (1) descriptions of the properties and use of radio telescopes and various types of receivers needed to analyze cosmic radio signals, and (2) descriptions of radiation mechanisms responsible for broadband and spectral line radiation. This book developed from a one-year graduate course that was given repeatedly at the Ruhr-Universität at Bochum. We hope that this text will be useful for all who use results obtained from radio astronomy. Our aim is to help them to understand the origin of well known results particularly the underlying assumptions and this book may occasionally save some scientists working in the field of radio astronomy from long searches in the literature when questions concerning tools occur.

The students to whom this course was addressed have had a rather thorough background knowledge of physics. However, difficulties often arose when the instrumental tools were discussed. Clearly there is a difference between how such a subject is treated in general physics books and the way it is presented in texts intended for engineers. Our explanations are meant to use concepts familiar to astrophysicists and physicists.

For each chapter, a list of references is given. Usually this list has two parts: general references give a list of papers and books that cover the general aspects and which often give a more thorough treatment of the subjects covered, and special references document the sources for specific topics. However, these references do not give a complete review of the relevant literature. The papers cited are those that present the subject in a convenient way.

The basic concepts used in the first edition have remained unchanged. This book gives an outline of the methods and tools of radio astronomy. Results are given to illustrate aspects of the theories or to make the approach used plausible. The book is intended to be of help in applying radio astronomy, but it is *not* a description of the many results. This book is not intended to be a review of the entire field of radio astronomy in the literature but describes only the basic and undisputed concepts and results.

Another problem encountered when writing a textbook is that of consistent designations, symbols, and units. Since the astronomical community prefers their traditional mixed set of units, we use the *Gaussian CGS system*, augmented when

necessary with other units. Where needed, we give the relations in their respective units in the equations.

References to the current literature have been updated. We do not attempt to give a complete review and we chose those references that are the most recent or cover the subject most comprehensively.

Contents

Chapter 1
Radio Astronomical Fundamentals

1.1 On the Role of Radio Astronomy in Astrophysics

Almost everything that we know about distant sources, that is, stars and the interstellar medium, has been obtained from electromagnetic radiation. This includes spatial distributions, kinematics and composition. Only a very small part of our knowledge stems from material information carriers, such as meteorites that impact the earth, cosmic ray particles or samples of material collected by manned or unmanned space probes.

For many thousands of years, mankind was restricted to measurements of visible light; only since the time of Herschel was this wavelength range slightly expanded into the near Infrared; in 1930, it extended from the near ultraviolet to the near infrared: $0.35\,\mu m \leq \lambda \leq 1\,\mu m$. At other wavelengths, investigations were limited either because the terrestrial atmosphere blocks radiation or because no detectors for this radiation were available. In 1931 this situation changed dramatically when Jansky showed that radiation at a wavelength of 14.6 m (=20.5 MHz) received with a direction sensitive antenna array, must be emitted by an extraterrestrial source which was not the sun. Jansky continued his observations over several years without achieving much scientific impact. His observations were first taken up and improved after 1937 by another radio engineer, Grote Reber, who carried out measurements at a shorter wavelength, $\lambda = 1.87$ m (=160 MHz). These observations were published in a professional astronomical journal. Later, after the end of World War II, improved receivers allowed the new radio window to develop. Radio physics had made great progress during the war years, mainly due to efforts directed towards the development of sensitive and efficient radar equipment. After the war, some researchers turned their attention towards the radio "noise" from extraterrestrial sources.

We will not follow this historical development any further, except to note that the historical development has been toward higher sensitivity, shorter wavelength, and higher angular resolution. The radio window reaches from $\lambda \cong 10\text{–}15$ m to short-ward of $\lambda \cong 0.3$ mm. Outside the near Infrared–optical window, this was the first new spectral range that became available to astronomy. The new astronomical discipline of radio astronomy has been instrumental in changing our view of astronomy. The results required mechanisms for their explanation that differed considerably

from those used previously. While the objects studied in the *optical* wavelength range usually radiate because they are hot and therefore *thermal* physics is the rule. Most often, in radio astronomy the radiation has a nonthermal origin and different physical mechanisms apply.

More recently, technological advances have opened up of additional astronomical "windows". Balloons, high-flying aircraft or satellites like IRAS, ISO and MSX permitted observations in the mid and far infrared (FIR). Other satellites such as IUE and CHANDRA permitted measurements in the ultraviolet and X-ray wavelength range. Satellite systems allow measurements over the spectral range from γ-rays to wavelengths greater than 10^4 m. Each of these spectral windows requires its own technology. The art of carrying out measurements differs for each. Astronomers have tended to view these different windows as forming different astronomies: radio astronomy, X-ray astronomy, infrared astronomy and so on. Not only does wavelength range and (to some extent) technology differ. The types of objects that emit at these wavelengths can also differ: some objects are detected only in certain spectral windows. For example, diffuse cool gas is detected only because it emits or absorbs the (first order forbidden) hyperfine structure line at $\lambda = 21$ cm; emission from this gas cannot be detected by any other means. To a lesser extent, this is true for denser, cool gas traced by allowed rotational transitions of carbon monoxide, CO. This material is detected only by molecular or atomic lines and broadband dust radiation. Although interpretations differ for each spectral window there is one single reality. An astrophysicist investigating a specific object collects information with optical, radio or other techniques. In this sense there is no such thing as a separate scientific discipline of radio astronomy.

New experimental techniques provide additional paths to attack old problems. More dramatically, when new kinds of objects are detected by these means, methods and results are often collected into a new discipline such as radio astronomy. However, when the experimental methods have become mature and both the advantages and limitations of the methods become clearer, it is appropriate to integrate the specialized field into main stream astrophysics. Radio astronomy is now in such a situation. The first, vigorous years when the pioneers worked alone or in small groups are over. Today radio astronomers rarely build their telescopes and receivers themselves. This has profound effects on the way research is done. In the pioneer days, a project usually started with an instrument collecting data; in many cases the results were unusual and exciting, so these required new explanations. Now a researcher starts with the problem and then searches for the means to attack it.

Today radio astronomy is not just a collection of the results, but also a science concerned with the instruments used to gather the data, including the instrumental properties, advantages and limitations. These instruments are usually no longer built by the astronomer. Rather, the astronomer's task is to optimize their use for a particular study. For this, the user must have a clear idea how the measurements are to be carried out. As to nomenclature, we refer to single radio telescopes as antennas and arrays of antennas with coupled outputs as interferometers. Together, either of these with *receivers* are the material tools used by radio astronomers. However there are more than only material tools: in interpreting the measurements theoretical concepts

must be applied to data. These concepts belong to a wide variety of physical fields, from plasma physics to molecular physics. All these concepts are tools, and so we have collected these in a "toolbox" that is consistent and useful.

1.2 The Radio Window

From the surface of the earth, the atmosphere is transparent to radio waves as long as none of its constituents is able to absorb this radiation to a noticeable extent. This earth-bound radio window extends roughly from a lower frequency limit of $v \cong 15\,\text{MHz}$ ($\lambda \cong 20\,\text{m}$) to a high frequency cut-off at $v \cong 1.5\,\text{THz}$ ($\lambda \cong 0.2\,\text{mm}$). These limits are not sharp (Fig. 1.1) since there are variations both with altitude, geographical position and with time.

The high-frequency cut-off occurs because the resonant absorption of the lowest rotation bands of molecules in the troposphere fall into this frequency range. Two molecules are mostly responsible for this: water vapor, H_2O and O_2. Water vapor has bands at $v = 22.2\,\text{GHz}$ ($\lambda = 1.35\,\text{cm}$) and $183\,\text{GHz}$ ($1.63\,\text{mm}$), while O_2 has an exceedingly strong band at $60\,\text{GHz}$ ($5\,\text{mm}$). Lines of O_2 consist of closely spaced rotational levels of the ground electronic state, resulting in two interleaved series of absorption lines near $60\,\text{GHz}$ ($5\,\text{mm}$) and a single line near $119\,\text{GHz}$ ($2.52\,\text{mm}$). The
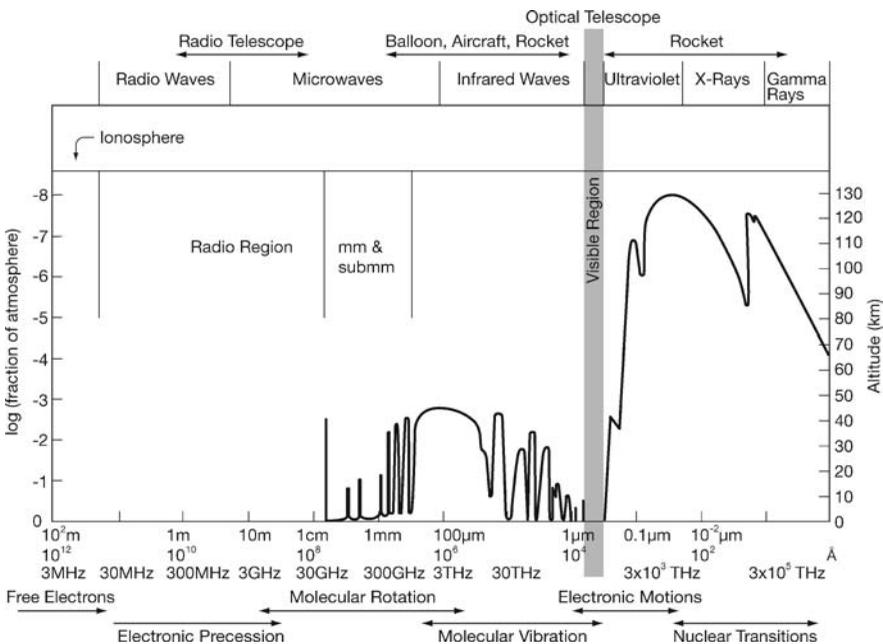
Fig. 1.1 The transmission of the earth's atmosphere for electromagnetic radiation. The diagram gives the height in the atmosphere at which the radiation is attenuated by a factor 1/2

absorption of astronomical signals by other abundant molecules in the atmosphere, N_2 and CO_2, occurs at frequencies above 300 GHz.

There is great interest to extend the upper frequency limits of the measurements to as high a value as possible, since the astronomical sources produce more intense spectral lines in this range. The rotational transitions of carbon monoxide, CO, play an especially important role since this molecule is very widespread and its chemistry is thought to be well understood. The circumstance that water vapor is one of the determining factors for this cut-off makes it possible to extend the accessible frequency range somewhat by carrying out measurements from locations with a low total water vapor content. With respect to the absorption caused by oxygen little can be done from earth's surface. In some parts of the sub-mm wavelength range, measurements must be carried out from satellites such as the *Herschel Space Observatory*, the airborne facility SOFIA (*Stratospheric Observatory for Infrared Astronomy*), or from high flying balloons. Interstellar spectral lines of water vapor and oxygen are best observed from satellites orbiting above the earth's atmosphere. On earth, high-altitude observatories with a dry climate are the best one can do. We will discuss the effects of the atmosphere in the chapter on observational methods.

At the lowest frequencies, the terrestrial atmosphere ceases to be transparent because of free electrons in the ionosphere. Transmission through the atmosphere is not possible if the frequency of the radiation is below the plasma frequency v_p. As we will show later (Eq. 2.77) this frequency is given by:

$$\frac{v_p}{\text{kHz}} = 8.97 \sqrt{\frac{N_e}{\text{cm}^{-3}}},$$

where N_e is the electron density of the plasma in cm^{-3} and v_p is given in kHz. Thus the low-frequency limit of the radio window will be near 4.5 MHz at night when the F_2 layer of the ionosphere has an average maximum density of $N_e \cong 2.5 \times 10^5 \, \text{cm}^{-3}$, and near 11 MHz at daytime, because then $N_e \cong 1.5 \times 10^6 \, \text{cm}^{-3}$. However the electron densities in the ionosphere depend on solar activity, and therefore this low-frequency limit varies with "space weather". Only when the observing frequency is well above this limit do ionospheric properties have no noticeable effect. Radio astronomy in the kHz frequency range must be performed from satellites above the earth's ionosphere.

Radio frequency interference (RFI) has an increasingly detrimental impact on astronomical observations. Man-made sources of radio signals, including intentional emitters (such as cell phones, wireless networks, garage door openers, and satellites) and unintentional radiators (such as computers and automobiles), can swamp very weak cosmic signals being studied. Some forms of RFI can be partially removed, but the presence of RFI always compromises the utility of the data and/or the efficiency of data acquisition. The International Telecommunication Union (ITU), an agency of the United Nations, is responsible for the global management of the radio spectrum, including the protection of radio astronomy. Expert committees, such as the European Science Foundation's Committee on Radio Astronomy Frequencies (CRAF) and the U.S. National Academy of Sciences' Committee on Radio

Frequencies (CORF), study spectrum issues and their impact on radio astronomy. In most radio observatories and at the U.S. National Science Foundation, at least one staff member is dedicated to the protection of radio astronomy observations. More information on the management of interference to radio astronomy can be found in the ITU Handbook on Radio Astronomy (http://www.itu.int/publ/R-HDB-22/en)

1.3 Some Basic Definitions

Electromagnetic radiation in the radio window is a wave phenomenon, but when the scale of the system involved is much larger than the wavelength, we can consider the radiation to travel in straight lines called *rays*. The infinitesimal power dP intercepted by an infinitesimal surface $d\sigma$ (Fig. 1.2) then is

$$dP = I_\nu \cos\theta \, d\Omega \, d\sigma \, d\nu \,, \tag{1.1}$$

where

dP = infinitesimal power, in watts,
$d\sigma$ = infinitesimal area of surface, cm^2,
$d\nu$ = infinitesimal bandwidth, in Hz,
θ = angle between the normal to $d\sigma$ and the direction to $d\Omega$,
I_ν = brightness or specific intensity, in $W\,m^{-2}\,Hz^{-1}\,sr^{-1}$.

Equation (1.1) should be considered to be the definition of the brightness I_ν. Quite often the term *intensity* or *specific intensity* I_ν is used instead of the term *brightness*. We will use all three designations interchangeably.

The total flux of a source is obtained by integrating (1.1) over the total solid angle Ω_s subtended by the source

$$S_\nu = \int_{\Omega_s} I_\nu(\theta, \varphi) \cos\theta \, d\Omega, \tag{1.2}$$
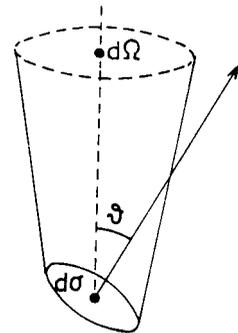
Fig. 1.2 A sketch to illustrate the definition of brightness

and this flux density is measured in units of $\mathrm{W\,m^{-2}\,Hz^{-1}}$. Since the flux density of radio sources is usually very small, a special radio astronomical flux density unit, the Jansky (abbreviated Jy) has been introduced

$$1\,\mathrm{Jy} = 10^{-26}\,\mathrm{Wm^{-2}Hz^{-1}} = 10^{-23}\,\mathrm{erg\,s^{-1}cm^{-2}Hz^{-1}}. \tag{1.3}$$

Very few sources are as bright as 1 Jy, but even such a source would produce a signal of only 10^{-15} W with the 100 m telescope (effective aperture $A \cong 5 \times 10^{3}\mathrm{m^{2}}$, $\Delta\nu = 20MHz$).

The brightness of an extended source is a quantity similar to the surface brightness in optical astronomy: it is independent of the distance to the source, as long as the effects of diffraction and extinction can be neglected. Consider a bundle of rays emitted by a source (Fig. 1.3), which contains the power dW. As long as the surface element $d\sigma$ covers the ray bundle completely, the power remains constant:

$$dP_1 = dP_2. \tag{1.4}$$

For each of these we have

$$dP_1 = I_{\nu_1}d\sigma_1\,d\Omega_1\,d\nu \qquad \text{and}$$
$$dP_2 = I_{\nu_2}d\sigma_2\,d\Omega_2\,d\nu.$$

If the distance between $d\sigma_1$ and $d\sigma_2$ is R, then the solid angles are $d\Omega_2 = d\sigma_1/R^2$, $d\Omega_1 = d\sigma_2/R^2$ and thus

$$dP_1 = I_{\nu_1}d\sigma_1\,\frac{d\sigma_2}{R^2}\,d\nu \qquad \text{and} \qquad dP_2 = I_{\nu_2}d\sigma_2\,\frac{d\sigma_1}{R^2}\,d\nu.$$

Using (1.4) we thus obtain

$$I_{\nu_1} = I_{\nu_2} \tag{1.5}$$

so that the brightness is independent of the distance. As we show next the total flux S_ν density shows the expected dependence of $1/r^2$. Consider a sphere with uniform brightness I_ν with a radius R (Fig. 1.4). The total flux received by an observer at the distance r then is, according to (1.2),
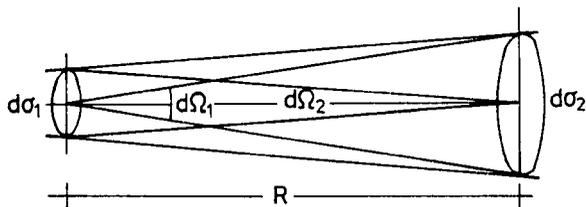
Fig. 1.3 A sketch used to show that the brightness is independent of the distance along a ray

Fig. 1.4 Total flux received at
a point P from an uniformly
bright sphere
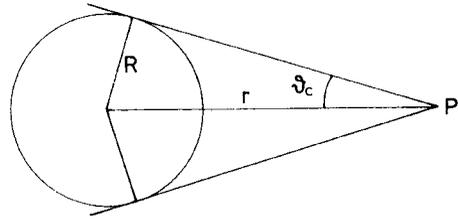
$$S_\nu = \int_{\Omega_s} I_\nu \cos\theta \, d\Omega = I_\nu \int_0^{2\pi} \left(\int_0^{\theta_c} \sin\theta \cos\theta \, d\theta \right) d\varphi ,$$

where

$$\sin\theta_c = \frac{R}{r}$$

defines the angle θ_c that the radius of the sphere subtends at r. We obtain

$$S_\nu = \pi I_\nu \sin^2\theta_c$$

or

$$S_\nu = I_\nu \frac{\pi R^2}{r^2} = I_\nu \Delta\Omega , \tag{1.6}$$

where $\Delta\Omega$ is defined as the area subtended by an object at a distance r.

Another useful quantity related to the brightness is the radiation energy density u_ν in units of erg cm^{-3}. From dimensional analysis, u_ν is intensity divided by speed. Since radiation propagates with the velocity of light c, we have for the *spectral energy density per solid angle*

$$u_\nu(\Omega) = \frac{1}{c} I_\nu . \tag{1.7}$$

If integrated over the whole sphere, 4π steradian, (1.7) results in the *total spectral energy density*

$$u_\nu = \int_{(4\pi)} u_\nu(\Omega) \, d\Omega = \frac{1}{c} \int_{(4\pi)} I_\nu \, d\Omega . \tag{1.8}$$

1.4 Radiative Transfer

Equation (1.5) shows that for radiation in free space the specific intensity I_ν remains independent of the distance along a ray. I_ν will change only if radiation is absorbed or emitted, and this change of I_ν is described by the equation of transfer. The theory to be outlined here is a macroscopic one: for a change in I_ν certain expressions are

adopted which contain free parameters. Only experience will then show whether these expressions are appropriate, or whether different ones should be preferred.

For a change in I_ν along the line of sight, a loss term $dI_{\nu-}$ and a gain term $dI_{\nu+}$ are introduced, and we adopt the form

$$dI_{\nu-} = -\kappa_\nu I_\nu \, ds \,,$$
$$dI_{\nu+} = \varepsilon_\nu \, ds \,,$$

so that the change of intensity in a slab of material of the thickness ds will be

$$\left[I_\nu(s+ds) - I_\nu(s)\right] d\sigma \, d\Omega \, d\nu = \left[-\kappa_\nu I_\nu + \varepsilon_\nu\right] d\sigma \, d\Omega \, d\nu \, ds \,,$$

resulting in the *equation of transfer*

$$\boxed{\frac{dI_\nu}{ds} = -\kappa_\nu I_\nu + \varepsilon_\nu} \quad . \tag{1.9}$$

From general experience, the linear absorption coefficient κ_ν is independent of the intensity I_ν leading to the adoption of the above form for $dI_{\nu-}$; similar arguments hold for the emissivity ε_ν.

There may be situations for which ε_ν depends strongly on I_ν, such as an environment in which radiation is strongly scattered. However, there are many other important situations where ε_ν is independent of I_ν.

There are several limiting cases for which the solution of the differential equation (1.9) is especially simple.

1) Emission only: $\kappa_\nu = 0$

$$\frac{dI_\nu}{ds} = \varepsilon_\nu \,, \qquad I_\nu(s) = I_\nu(s_0) + \int_{s_0}^{s} \varepsilon_\nu(s) \, ds \,. \tag{1.10}$$

2) Absorption only: $\varepsilon_\nu = 0$

$$\frac{dI_\nu}{ds} = -\kappa_\nu I_\nu \,,$$

$$I_\nu(s) = I_\nu(s_0) \exp\left\{ -\int_{s_0}^{s} \kappa_\nu(s) \, ds \right\} \,. \tag{1.11}$$

3) Thermodynamic equilibrium (TE): If there is complete equilibrium of the radiation with its surroundings, the brightness distribution is described by the Planck function, which depends only on the thermodynamic temperature T, of the surroundings

$$\frac{dI_\nu}{ds} = 0 \,, \qquad I_\nu = B_\nu(T) = \varepsilon_\nu / \kappa_\nu \tag{1.12}$$

$$B_v(T) = \frac{2hv^3}{c^2}\frac{1}{e^{hv/kT}-1} \qquad . \qquad (1.13)$$

4) Local thermodynamic equilibrium (LTE): Full thermodynamic equilibrium will be realized only in very special circumstances such as in a black enclosure or, say, in stellar interiors. Often *Kirchhoff's law* is

$$\frac{\varepsilon_v}{\kappa_v} = B_v(T) \qquad (1.14)$$

applicable independent of the material, as is the case with complete thermodynamic equilibrium. In general however, I_v will differ from $B_v(T)$.

If we define the *optical depth* $d\tau_v$ (Fig. 1.5) by

$$d\tau_v = -\kappa_v\,ds \qquad (1.15)$$

or

$$\tau_v(s) = \int_{s_0}^{s} \kappa_v(s)\,ds\,, \qquad (1.16)$$

then the equation of transfer (1.9) can be written as

$$-\frac{1}{\kappa_v}\frac{dI_v}{ds} = \frac{dI_v}{d\tau_v} = I_v - B_v(T) \qquad . \qquad (1.17)$$

The solution of (1.17) is obtained by first multiplying (1.17) by $\exp(-\tau_v)$ and then integrating τ_v by parts:

$$\int_0^{\tau_v(s)} e^{-\tau}\frac{dI_v}{d\tau}d\tau = I_v e^{-\tau}\Big|_0^{\tau_v(s)} + \int_0^{\tau_v(s)} I_v e^{-\tau}d\tau = \int_0^{\tau_v(s)} (I_v - B_v)e^{-\tau}d\tau$$
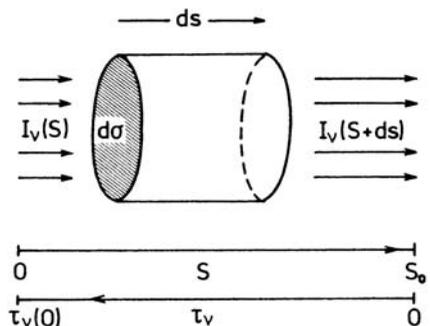
Fig. 1.5 A sketch showing the quantities used in the equation of transfer

$$I_\nu(\tau_\nu(s))\,\mathrm{e}^{-\tau_\nu(s)} - I_\nu(\tau_\nu(s_0))\,\mathrm{e}^0 = -\int\limits_0^{\tau_\nu(s)} B_\nu(T(\tau))\,\mathrm{e}^{-\tau}\mathrm{d}\tau$$

or finally

$$I_\nu(s) = I_\nu(0)\,\mathrm{e}^{-\tau_\nu(s)} + \int\limits_0^{\tau_\nu(s)} B_\nu(T(\tau))\,\mathrm{e}^{-\tau}\mathrm{d}\tau \qquad (1.18)$$

Due to the definition (1.15), s and τ increase in opposite directions as indicated in Fig. 1.5.

If the medium is isothermal, that is, if

$$T(\tau) = T(s) = T = \text{const.}$$

the integral in (1.18) can be computed explicitly resulting in

$$I_\nu(s) = I_\nu(0)\,\mathrm{e}^{-\tau_\nu(s)} + B_\nu(T)\left(1 - \mathrm{e}^{-\tau_\nu(s)}\right) \qquad (1.19)$$

For a large optical depth, that is for $\tau_\nu(0) \to \infty$, (1.19) in LTE approaches the limit

$$I_\nu = B_\nu(T). \qquad (1.20)$$

The observed brightness I_ν for the optically thick case is equal to the Planck black-body brightness distribution independent of the material. If the intensity is to be compared with the result obtained in the absence of an intervening medium, $I_\nu(0)$, we have

$$\Delta I_\nu(s) = I_\nu(s) - I_\nu(0) = (B_\nu(T) - I_\nu(0))(1 - \mathrm{e}^{-\tau}). \qquad (1.21)$$

1.5 Black Body Radiation and the Brightness Temperature

The spectral distribution of the radiation of a black body in thermodynamic equilibrium is given by the Planck law (cf. (1.13))

$$B_\nu(T) = \frac{2h\nu^3}{c^2}\frac{1}{\mathrm{e}^{h\nu/kT} - 1} \qquad .$$

It gives the power per unit frequency interval. Converting this to the wavelength scale, we obtain $B_\lambda(T)$. Because $B_\nu(T)\,\mathrm{d}\nu = -B_\lambda(T)\,\mathrm{d}\lambda$ and $\mathrm{d}\nu = (-c/\lambda^2)\,\mathrm{d}\lambda$ this is

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc/k\lambda T} - 1} \quad . \tag{1.22}$$

Integrating either (1.13) over ν or (1.22) over λ, the total brightness of a black body is obtained

$$B(T) = \frac{2h}{c^2} \int_0^\infty \frac{\nu^3}{e^{h\nu/kT} - 1} \, d\nu \, .$$

Putting

$$x = \frac{h\nu}{kT} \, , \tag{1.23}$$

we get

$$B(T) = \frac{2h}{c^2} \left(\frac{kT}{h}\right)^4 \int_0^\infty \frac{x^3}{e^x - 1} \, dx \, .$$

The integral has the value $\pi^4/15$ [an explicit demonstration of this is given in, e.g., Reif (1965), Sect. A.11]. Thus

$$B(T) = \sigma T^4 \, , \quad \sigma = \frac{2\pi^4 k^4}{15c^2 h^3} = 1.8047 \times 10^{-5} \, \text{erg cm}^{-2} \text{s}^{-1} \text{K}^{-4} \quad . \tag{1.24}$$

In some texts, such as Leighton's *Principles of Modern Physics*, the value of σ is given with an extra factor of π so that in CGS units, the value is $\sigma = 5.67 \times 10^{-5}$.

Equation (1.24) is the Stefan-Boltzmann radiation law which was found experimentally in 1879 by J. Stefan and derived theoretically in 1884 by L. Boltzmann before Planck's radiation law was known. In the literature quite often a different value for σ is given which is obtained, when the total radiation emitted into a solid angle of 2π is computed from (1.24). Both (1.13) and (1.22) have maxima (Fig. 1.6) which are found by solving $\partial B_\nu/\partial \nu = 0$ and $\partial B_\lambda/\partial \lambda = 0$ respectively. Using (1.23), these correspond to solving $3(1 - e^{-x}) - x = 0$ and $5(1 - e^{-x}) - x = 0$ with the solutions

$$x_m = 2.82143937 \quad \text{and} \quad \hat{x}_m = 4.96511423 \, .$$

Thus (1.13) attains its maximum at

$$\frac{\nu_{max}}{\text{GHz}} = 58.789 \left(\frac{T}{K}\right) \quad , \tag{1.25}$$

while from (1.22)
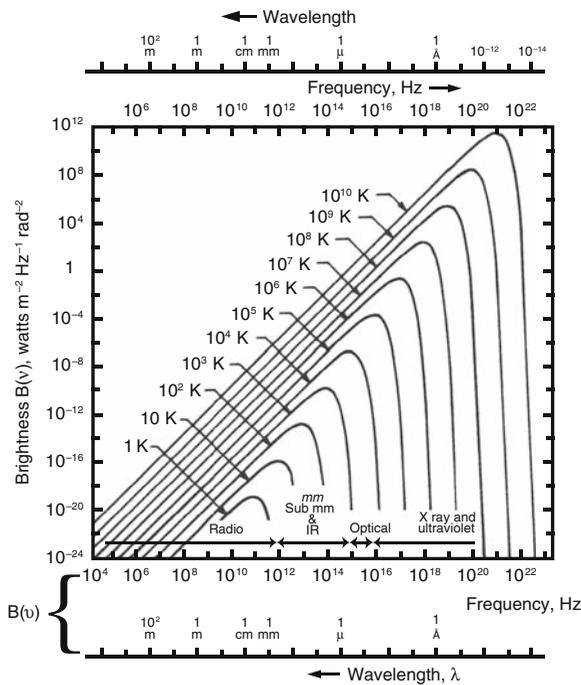
Fig. 1.6 Planck spectra for black bodies of different temperatures

$$\left(\frac{\lambda_{\max}}{\text{cm}}\right)\left(\frac{T}{\text{K}}\right) = 0.28978 \qquad . \tag{1.26}$$

Equations (1.25) and (1.26) are both known as *Wien's displacement law*. If $x = h\nu/kT$ is far from the maximum, (1.13) can be approximated by simpler expressions (Fig. 1.7).

1) $h\nu \ll kT$: *Rayleigh-Jeans Law*. An expansion of the exponential

$$e^{h\nu/kT} \cong 1 + \frac{h\nu}{kT} + \dots \tag{1.27}$$

results in

$$B_{\text{RJ}}(\nu, T) = \frac{2\nu^2}{c^2}kT \qquad . \tag{1.28}$$

This is the classical limit of the Planck law since it does not contain Planck's constant. In the millimeter and submillimeter range, one frequently defines a *radiation temperature*, $J(T)$ as

Fig. 1.7 Normalized Planck
curve and the Rayleigh-Jeans
and Wien approximation
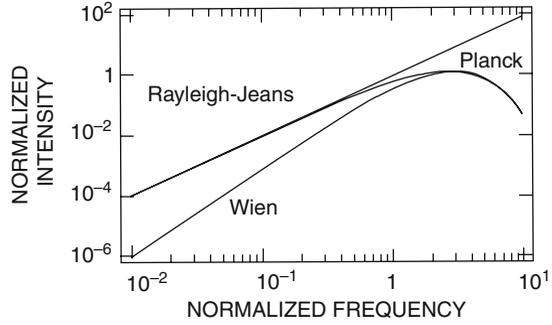
$$J(T) = \frac{c^2}{2kv^2} I = \frac{hv}{k} \frac{1}{e^{hv/kT} - 1} . \tag{1.29}$$

Inserting numerical values for k and h, we see that the Rayleigh-Jeans relation
holds for frequencies

$$\frac{v}{\text{GHz}} \ll 20.84 \left(\frac{T}{\text{K}} \right) . \tag{1.30}$$

It can thus be used for all thermal radio sources except perhaps for low temperatures in the millimeter or sub-mm range.

2) $hv \gg kT$: *Wien's Law.* In this case $e^x \gg 1$, so that

$$\boxed{B_{\text{W}}(v, T) = \frac{2hv^3}{c^2} e^{-hv/kT}} \quad . \tag{1.31}$$

While this limit is quite useful for stellar measurements in the visual and ultraviolet range, it plays no role in radio astronomy.

One of the important features of the Rayleigh-Jeans law is the implication that the brightness and the thermodynamic temperature of the black body that emits this radiation are strictly proportional (1.28). This feature is so useful that it has become the custom in radio astronomy to measure the brightness of an extended source by its *brightness temperature* T_b. This is the temperature which would result in the given brightness if inserted into the Rayleigh-Jeans law

$$T_b = \frac{c^2}{2k} \frac{1}{v^2} I_v = \frac{\lambda^2}{2k} I_v . \tag{1.32}$$

Combining (1.6) with (1.32), we have

$$\boxed{S_v = \frac{2kv^2}{c^2} T_b \Delta\Omega} \quad . \tag{1.33}$$

For a Gaussian source, this relation is

$$\left[\frac{S_v}{\text{Jy}}\right] = 2.65\, T_b \left[\frac{\theta}{\text{arc minutes}}\right]^2 \left[\frac{\lambda}{\text{cm}}\right]^{-2} \tag{1.34}$$

That is, with a measurement of the flux density S_v in Janskys, and the source size, the brightness temperature, T_b, of the source can be determined.

If emitted by a black body and $hv \ll kT$ then T_b gives the thermodynamic temperature of the source, a value that is independent of v. If other processes are responsible for the emission of the radiation, T_b will depend on the frequency; it is, however, still a useful quantity and is commonly used in practical work.

This is the case even if the frequency is so high that condition (1.30) is not valid. Then (1.34) can still be applied, but it should be understood that T_b is different from the thermodynamic temperature of a black body. However, it is rather simple to obtain the appropriate correction factors.

It is also convenient to introduce the concept of brightness temperature into the radiative transfer equation (1.21). Formally one can obtain

$$J(T) = \frac{c^2}{2kv^2}(B_v(T) - I_v(0))(1 - e^{-\tau_v(s)})\,.$$

Usually calibration procedures (see Sect. 8.2) allow one to express $J(T)$ as T. This measured quantity is referred to as T_R^*, the *radiation temperature*, or the *brightness temperature*, T_b. In the centimeter wavelength range, one can apply (1.32) to (1.17) and one obtains

$$\boxed{\frac{dT_b(s)}{d\tau_v} = T_b(s) - T(s)} \quad, \tag{1.35}$$

where $T(s)$ is the thermodynamic temperature of the medium at the position s. The general solution is

$$\boxed{T_b(s) = T_b(0)\,e^{-\tau_v(s)} + \int_0^{\tau_v(s)} T(s)\,e^{-\tau}d\tau} \quad . \tag{1.36}$$

If the medium is isothermal, this becomes

$$\boxed{T_b(s) = T_b(0)\,e^{-\tau_v(s)} + T\,(1 - e^{-\tau_v(s)})} \quad . \tag{1.37}$$

For the sake of simplicity, let us assume that $T_b(0) = 0$. Then two limiting cases that are often applicable are:

1) for optically thin $\tau \ll 1$,

$$T_b = \tau_v T\,, \tag{1.38}$$

and

2) for optically thick $\tau \gg 1$,

$$T_b = T \, . \tag{1.39}$$

These relations are correct only if both the geometry of the source and the radiating medium are not important. One usually expresses this as "the sources are much larger than the telescope beam".

1.6   The Nyquist Theorem and the Noise Temperature

Finally, we relate electrical power and temperature. Suppose a resistor R is connected across the input terminals of a linear amplifier. The thermal motion of the electrons in the resistor will produce a current $i(t)$ which forms a random input to the amplifier. Though the mean value of this current will be zero, its rms value will not be so. Since $\langle i^2 \rangle \neq 0$ represents a power, the resistor provides a power input to the amplifier. In thermal equilibrium, this power is determined by the physical temperature. This is Johnson noise. This situation was investigated in 1929 by H. Nyquist, who showed that this is a problem similar to that of the random walk of a particle in Brownian motion including a friction term. A detailed discussion goes beyond the scope of this book, it can be found in many treatments of stochastic processes [see the appropriate chapters in Reif (1965) or Papoulis and Pillai (2002)].

The average power per unit bandwidth produced by the resistor R in the circuit shown in Fig. 1.8 is

$$P_v = \langle iv \rangle = \frac{\langle v^2 \rangle}{2R} = \frac{1}{4R} \langle v_N^2 \rangle \, , \tag{1.40}$$

where $v(t)$ is the voltage that is produced by i across R, and $\langle \cdots \rangle$ indicates a time average. The first factor $1/2$ arises from the need to transfer maximum power to the element on the right. This condition is met by setting $R_x = R$; then $i = v/2R$. The second factor $1/2$ arises from the time average of v^2. An analysis of the random walk process now shows that

$$\langle v_N^2 \rangle = 4RkT \, . \tag{1.41}$$

Inserting this into (1.40) we obtain

$$P_v = kT \, . \tag{1.42}$$

Fig. 1.8 A sketch of a circuit containing a resistor R, to illustrate the origin of Johnson noise. The resistor R, on the left, at a temperature T, provides a power kT to a matched load R_X, on the right
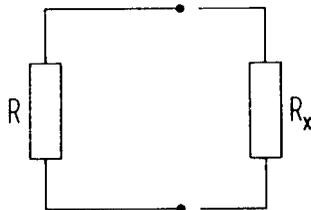
Expression (1.42) can also be obtained by a reformulation of (1.13) for one dimension and the Rayleigh-Jeans limit. Then, the available noise power of a resistor is proportional to its temperature, the *noise temperature* T_N, and independent of the value of R. Throughout the whole radio range, from the longest waves to the far infrared region the noise spectrum is white, that is, its power is independent of frequency. For receivers, since the impedance of a noise source must be matched to that of the amplifier, such a noise source can only be matched over some finite bandwidth.

Not all circuit elements can be characterized by thermal noise. For example a microwave oscillator can deliver the equivalent of more than 10^{16} K, although the physical temperature is only 300 K. Clearly this is a very nonthermal process, and in this case temperature is not a useful concept.

Problems

1. If the average electron density in the interstellar medium (ISM) is $0.03\,\mathrm{cm}^{-3}$, what is the lowest frequency of electromagnetic radiation which one can receive due to the effect of this plasma? Compare this to the ionospheric plasma cutoff frequency if the electron density, N_e, in the ionosphere is $\sim 10^5\,\mathrm{cm}^{-3}$.

2. (a) A researcher measures radio emission at a frequency of 250 kHz and finds that the emission is present over the whole sky with a brightness temperature of 250 K. Could the origin of this radiation be the earth's ionosphere?
(b) Assume that the source fills the entire visible sky, taken to be a half hemisphere. What is the power received by an antenna with $A = 1\,\mathrm{m}^2$ collecting area in a $B = 1\,\mathrm{kHz}$ bandwidth?

3. There is a proposal to orbit a downward-pointing radar in a satellite, *Cloudsat*, moving in a polar orbit. The satellite will orbit at an altitude of 500 km. The operating frequency is 94 GHz. Assume that the power is radiated over a hemisphere. The peak power will be 1500 W, uniformly distributed over a bandwidth of 1 GHz. If no power is absorbed in the earth's atmosphere, what is the peak flux density of this satellite when it is directly overhead? This radar is transmitting 3% of the time (duty cycle). What is the *average* power radiated and the corresponding flux density?

4. A unit commonly used in astronomy is flux density, S_ν, the Jansky (Jy). One Jy is $10^{-26}\,\mathrm{W\,m^{-2}\,Hz^{-1}}$. Calculate the flux density, in Jy, of a microwave oven with an output of 1 kW at a distance of 10 m if the power is radiated over all angles and is uniformly emitted over a bandwidth of 1 MHz.

5. (a) What is the flux density, S_ν, of a source which radiates a power of 1 kW in the microwave frequency band uniformly from 2.6 GHz to 2.9 GHz, when placed at the distance of the Moon (3.84×10^5 km)? Repeat for an identical source if the radiation is in the optical frequency band, from 3×10^{14} to 8×10^{14} Hz.

(b) If we *assume* that the number of photons is uniform over the band, what is the average energy, $E = h\nu$, of a photon? Use this average photon energy and the power to determine N, the number of photons. How many photons pass through a $1\,m^2$ area in one second in the optical and radio frequency bands?

6. In the near future there may be an anti-collision radar installed on automobiles. It will operate at $\sim 70\,GHz$. If the bandwidth is $10\,MHz$, and at a distance of $3\,m$, the power per area is $10^{-9}\,W\,m^{-2}$. Assume the power level is uniform over the entire bandwidth of $10\,MHz$. What is the flux density of this radar at $1\,km$ distance? A typical large radio telescope can measure to the mJy $(=10^{-29}\,W\,m^{-2}\,Hz^{-1})$ level. At what distance will such radars disturb such radio astronomy measurements?

7. If the intensity of the Sun peaks in the optical range, at a frequency of about $3.4 \times 10^{14}\,Hz$, what is the temperature of the Sun? Use the Wien displacement law (1.25). If all of the power is emitted only between 3 and $4 \times 10^{14}\,Hz$, how many photons per cm^2 arrive at the earth when the Sun is directly overhead? What is the power received on earth per cm^2? A value for the solar power is $135\,mW$ per cm^2. How does this compare to your calculation?

8. (a) At what frequency does the intensity of a $2.73\,K$ black body reach a maximum? At what wavelength?
(b) Could the difference between the maximum wavelength and frequency be caused by the different weightings of the Planck relation? Determine B_ν at the maximum frequency.
(c) What is the (integrated) energy density $u = (1/c) \int I d\Omega = (4\pi/c) I$?
(d) Reformulate the derivation of the Stefan–Boltzmann relation to obtain the number density of photons. Make use of the relation

$$\int_0^\infty \frac{x^2}{e^x - 1} dx = 2.404$$

to determine how many photons are present in a volume of $1\,cm^{-3}$.
(e) What is the error in applying the Rayleigh–Jeans approximation, instead of the Planck relation to calculate the intensity of the $2.73\,K$ black body radiation at $4.8\,GHz$, $115\,GHz$ and $180\,GHz$?

9. From Eq. (1.42), the power radiated in one dimension is $P = kT \Delta\nu$. If a microwave oscillator delivers $1\,mW$ of power uniformly over a bandwidth of $1\,Hz$, what is the equivalent temperature T? Since the physical temperature of such an oscillator is $\sim 300\,K$, this is an example of a *non-thermal* process.

10. A cable has an optical depth, τ, of 0.1 and a temperature of $T = 300\,K$. A signal of peak temperature $T_b(0) = 1\,K$ is connected to the input of this cable. Use Eq. (1.37) to analyze this situation. What is $T_b(s)$, the temperature of the output of the cable? Repeat the problem for $T = 100\,K$. What is the signal-to-noise ratio for these two cases, using signal $= 1\,K$, and noise from the cable contribution.?

11. A signal passes through two cables with the same optical depth, τ. These have temperatures T_1 and T_2, with $T_1 < T_2$. Which cable should be connected first to obtain the lowest output power from this arrangement?

12. Show that (1.34) can be obtained from (1.33).

13. If Jupiter has $T_B = 150\,\mathrm{K}$, with $\theta = 40''$, what is S_ν at $1.4\,\mathrm{GHz}$? At $115\,\mathrm{GHz}$? Repeat for the HII region Orion A, with $\theta = 2.5'$, with $T_B = 330\,\mathrm{K}$ at $4.8\,\mathrm{GHz}$, and $T_B = 24\,\mathrm{K}$ at $23\,\mathrm{GHz}$.

Chapter 2
Electromagnetic Wave Propagation Fundamentals

2.1 Maxwell's Equations

Maxwell's theory of electrodynamics describes electromagnetic fields in terms of the space and time variations of electromagnetic field components. In most treatises on electrodynamics, this theory is derived by induction starting with static situations.

Here we give only those features of the theory that are needed to understand the formation, emission and propagation of electromagnetic waves. These will be given in a uniform set of quantities, in the CGS system. These are the electric field intensity E, the electric displacement D, the magnetic field intensity H, the magnetic induction B, and the electric current density J. The electric charge density is designated by ϱ.

The relations of the five vector fields and one scalar field which are required to (properly) describe the electromagnetic phenomena are given by Maxwell's equations. These are conveniently divided into several groups. Some of the field components are related by the properties of the medium in which they exist. These are the so-called material equations

$$J = \sigma E \qquad (2.1)$$
$$D = \varepsilon E \qquad (2.2)$$
$$B = \mu H \qquad . \qquad (2.3)$$

σ, ε and μ are scalar functions that are almost constant in most materials. For the *Gaussian CGS system* the values of ε and μ are unity (=1) in vacuum, while (2.1) is the differential form of Ohm's law, where σ is the specific conductivity.

Maxwell's equations proper can now be further divided into two groups: The first group involves only the spatial structure of the fields

$$\nabla \cdot D = 4\pi\varrho \qquad (2.4)$$
$$\nabla \cdot B = 0 \qquad , \qquad (2.5)$$

while the second group includes time derivatives

$$\nabla \times \boldsymbol{E} = -\frac{1}{c}\dot{\boldsymbol{B}}$$

(2.6)

$$\nabla \times \boldsymbol{H} = \frac{4\pi}{c}\boldsymbol{J} + \frac{1}{c}\dot{\boldsymbol{D}}$$.

(2.7)

Taking the divergence of (2.7) the left side of the resulting equation is found to be equal to zero (see Appendix A). If we use (2.4), we obtain

$$\nabla \cdot \boldsymbol{J} + \dot{\varrho} = 0$$;

(2.8)

that is, charge density and current obey a continuity equation.

2.2   Energy Conservation and the Poynting Vector

By considering the forces that a static electric or magnetic field imposes on a test charge it can be shown that the energy density of an electromagnetic field is given by

$$u = \frac{1}{8\pi}\left(\boldsymbol{E}\cdot\boldsymbol{D} + \boldsymbol{B}\cdot\boldsymbol{H}\right) = \frac{1}{8\pi}\left(\varepsilon E^2 + \mu H^2\right)$$.

(2.9)

If both ε and μ are time-independent, the time derivative of u is given by

$$\dot{u} = \frac{1}{4\pi}\left(\varepsilon\boldsymbol{E}\cdot\dot{\boldsymbol{E}} + \mu\boldsymbol{H}\cdot\dot{\boldsymbol{H}}\right) = \frac{1}{4\pi}\left(\boldsymbol{E}\cdot\dot{\boldsymbol{D}} + \boldsymbol{H}\cdot\dot{\boldsymbol{B}}\right).$$

(2.10)

Substituting both $\dot{\boldsymbol{D}}$ and $\dot{\boldsymbol{B}}$ from Maxwell's equations (2.6) and (2.7), this becomes

$$\dot{u} = \frac{c}{4\pi}\left(\boldsymbol{E}\cdot(\nabla\times\boldsymbol{H}) - \boldsymbol{H}\cdot(\nabla\times\boldsymbol{E})\right) - \boldsymbol{E}\cdot\boldsymbol{J}$$

$$\dot{u} = -\frac{c}{4\pi}\nabla\cdot(\boldsymbol{E}\times\boldsymbol{H}) - \boldsymbol{E}\cdot\boldsymbol{J}$$

(2.11)

if the vector identity (A 9) given in Appendix A is applied. By introducing the Poynting vector $\boldsymbol{S}$ (Poynting 1884)

$$\boldsymbol{S} = \frac{c}{4\pi}\boldsymbol{E}\times\boldsymbol{H}$$,

(2.12)

(2.11) can be written as an equation of continuity for $\boldsymbol{S}$:

$$\frac{\partial u}{\partial t} + \nabla\cdot\boldsymbol{S} = -\boldsymbol{E}\cdot\boldsymbol{J}$$.

(2.13)

Fig. 2.1 A sketch to illustrate energy conservation. We show the Poynting vector for a circular straight wire carrying a steady current density produced by the electric field
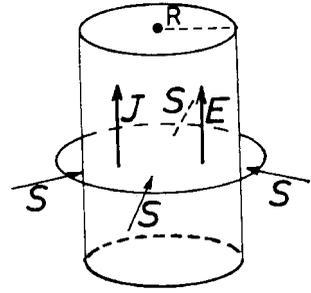
The time variation of the energy density u thus consists of two parts: a spatial change of the Poynting vector or energy flux S and a conversion of electromagnetic energy into thermal energy (Joule's energy theorem).

The significance of (2.13) becomes clearer if we consider a simple example. Let a straight wire of circular cross section carry a steady current I (Fig. 2.1). If all conditions are constant, the total electromagnetic energy density, u, must be constant, so that $\dot{u} = 0$. However if a constant current I is flowing in the wire there is a constant transformation of electric energy into thermal energy. Per unit length l of the wire, this thermal energy is formed at a rate

$$\frac{dW}{dl} = rI^2 , \tag{2.14}$$

where r is the specific resistance of the wire. Obviously

$$r = \frac{1}{\sigma \pi R^2}$$

so that

$$\frac{dW}{dl} = \frac{I^2}{\sigma \pi R^2} .$$

But

$$|\boldsymbol{J}| = \frac{I}{\pi R^2} ,$$

and according to (2.1)

$$\boldsymbol{E} = \frac{1}{\sigma} \boldsymbol{J} ,$$

so that the thermal loss rate is

$$\frac{dW}{dl} = |\boldsymbol{E}| I . \tag{2.15}$$

But according to Ampere's law (see e.g. Jackson Equation (5.6))

$$|\boldsymbol{H}| = \frac{2I}{cR}$$

with a direction perpendicular to I. Then

$$|S| = \frac{c}{4\pi} |E \times H| = \frac{|E| I}{2\pi R},$$

where S is oriented such that E, H and S form a right-handed system. Therefore

$$|E| I = 2\pi R |S|. \tag{2.16}$$

Thus the total flux of S at the surface of the wire and, from the direction of J and H, we see that S flows into it (Fig. 2.1). But according to (2.15) this is just the conversion rate of electrical energy into thermal energy. Therefore the Poynting flux just compensates for this loss, as it must in a steady state.

2.3 Complex Field Vectors

In situations where electromagnetic wave phenomena are considered, the field vectors usually show a harmonic time dependence described by sine or cosine functions. But since these functions are related to the exponential function by the Euler relation

$$\cos x + i \sin x = e^{ix},$$

the inconvenience of having to apply the rather complicated trigonometric addition theorem can be avoided, if complex field vectors are introduced by

$$E = (E_1 + i E_2) e^{-i\omega t}; \quad E_1, E_2 \text{ real vector fields}, \tag{2.17}$$

and

$$H = (H_1 + i H_2) e^{-i\omega t}; \quad H_1, H_2 \text{ real vector fields}. \tag{2.18}$$

In any application the electric or magnetic field considered is then identified with the real part of E and H or the imaginary part, whichever is more convenient. All mathematical operations can then be performed on E or H directly, as long as they are restricted to linear operations. Only if nonlinear operations are involved must one return to real quantities. Even here convenient simplifications exist. Such is the case for the Poynting vector. For S obviously the expression

$$S = \frac{c}{4\pi} \text{Re}\{E\} \times \text{Re}\{H\} \tag{2.19}$$

should be used. But since

$$\text{Re}\{E\} = E_1 \cos \omega t + E_2 \sin \omega t$$

and

$$\text{Re}\{H\} = H_1 \cos \omega t + H_2 \sin \omega t,$$

this is

$$\text{Re}\{\boldsymbol{E}\} \times \text{Re}\{\boldsymbol{H}\} = (\boldsymbol{E}_1 \times \boldsymbol{H}_1) \cos^2 \omega t + (\boldsymbol{E}_2 \times \boldsymbol{H}_2) \sin^2 \omega t$$
$$+ (\boldsymbol{E}_1 \times \boldsymbol{H}_2 + \boldsymbol{E}_2 \times \boldsymbol{H}_1) \cos \omega t \sin \omega t \,.$$

If we now do not consider the instantaneous value of $\boldsymbol{S}$, but the mean value over a full oscillation, and if such mean values are designated by $\langle \rangle$, then since

$$\langle \sin^2 \omega t \rangle = \langle \cos^2 \omega t \rangle = \tfrac{1}{2}$$

and

$$\langle \sin \omega t \cos \omega t \rangle = 0 \,,$$

one obtains

$$\langle \text{Re}\{\boldsymbol{E}\} \times \text{Re}\{\boldsymbol{H}\} \rangle = \tfrac{1}{2} (\boldsymbol{E}_1 \times \boldsymbol{H}_1 + \boldsymbol{E}_2 \times \boldsymbol{H}_2) \,. \tag{2.20}$$

On the other hand

$$\boldsymbol{E} \times \boldsymbol{H}^* = (\boldsymbol{E}_1 + \mathrm{i}\boldsymbol{E}_2) \, \mathrm{e}^{-\mathrm{i}\omega t} \times (\boldsymbol{H}_1 - \mathrm{i}\boldsymbol{H}_2) \, \mathrm{e}^{\mathrm{i}\omega t}$$
$$= (\boldsymbol{E}_1 + \mathrm{i}\boldsymbol{E}_2) \times (\boldsymbol{H}_1 - \mathrm{i}\boldsymbol{H}_2)$$

so that

$$\text{Re}\{\boldsymbol{E} \times \boldsymbol{H}^*\} = \boldsymbol{E}_1 \times \boldsymbol{H}_1 + \boldsymbol{E}_2 \times \boldsymbol{H}_2 \,,$$

where $\boldsymbol{H}^*$ denotes the complex conjugate of $\boldsymbol{H}$. Inserting this in (2.20) the average value of $\boldsymbol{S}$ is

$$\boxed{\langle \boldsymbol{S} \rangle = \frac{c}{4\pi} \text{Re}\{\boldsymbol{E} \times \boldsymbol{H}^*\}} \quad . \tag{2.21}$$

From (2.17) and (2.18), this formula applies only to complex electromagnetic fields that have harmonic time variations.

2.4 The Wave Equation

Maxwell's equations (2.4–2.7) give the connection between the spatial and the time variation of the electromagnetic field. However, the situation is complicated by the fact that the equations relate different fields: e.g. curl $\boldsymbol{E}$ is related to $\dot{\boldsymbol{B}}$ (2.6), and the other equations show a similar behavior.

A better insight into the behavior of the fields can be obtained if the equations are reformulated so that only a single vector field appears in each equation. This is achieved by the use of the wave equations. To simplify the derivation, the conductivity σ, the permittivity ε and the permeability μ will be assumed to be constants both in time and in space. Taking the curl of (2.7)

$$\nabla \times (\nabla \times \boldsymbol{H}) = \frac{4\pi}{c} \nabla \times \boldsymbol{J} + \frac{1}{c} \frac{\partial}{\partial t} \nabla \times \boldsymbol{D}$$

$$= \frac{4\pi}{c} \nabla \times (\sigma \boldsymbol{E}) + \frac{1}{c} \frac{\partial}{\partial t} \nabla \times (\varepsilon \boldsymbol{E})$$

$$= \frac{1}{c} \left(4\pi\sigma + \varepsilon \frac{\partial}{\partial t} \right) \nabla \times \boldsymbol{E},$$

where the order of ∇ and time derivation have been interchanged, and $\boldsymbol{J}$ and $\boldsymbol{D}$ have been replaced by $\sigma \boldsymbol{E}$ and $\varepsilon \boldsymbol{E}$ respectively by application of (2.1) and (2.2). Using (2.6) and (2.3), this can be further modified to

$$\nabla \times (\nabla \times \boldsymbol{H}) = -\frac{\mu}{c^2} \left(4\pi\sigma + \varepsilon \frac{\partial}{\partial t} \right) \frac{\partial}{\partial t} \boldsymbol{H} = -\frac{\mu}{c^2} (4\pi\sigma \dot{\boldsymbol{H}} + \varepsilon \ddot{\boldsymbol{H}}) . \qquad (2.22)$$

By a similar procedure from (2.6)

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\frac{1}{c} \frac{\partial}{\partial t} (\nabla \times \boldsymbol{B}) = -\frac{\mu}{c} \frac{\partial}{\partial t} (\nabla \times \boldsymbol{H}) .$$

Using (2.7) this becomes

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\frac{\mu}{c} \frac{\partial}{\partial t} \left(\frac{4\pi}{c} \boldsymbol{J} + \frac{1}{c} \dot{\boldsymbol{D}} \right) = -\frac{\mu}{c} \frac{\partial}{\partial t} \left(\frac{4\pi}{c} \sigma \boldsymbol{E} + \frac{\varepsilon}{c} \dot{\boldsymbol{E}} \right)$$

$$= -\frac{\mu}{c^2} (4\pi\sigma \dot{\boldsymbol{E}} + \varepsilon \ddot{\boldsymbol{E}}) . \qquad (2.23)$$

The left-hand side of (2.22) and (2.23) can be reduced to a more easily recognisable form by using the vector identity [see Appendix (A.13)]

$$\nabla \times (\nabla \times \boldsymbol{P}) = \nabla(\nabla \cdot \boldsymbol{P}) - \nabla^2 \boldsymbol{P} ;$$

applying this relation to (2.5)

$$\nabla \times (\nabla \times \boldsymbol{H}) = \nabla(\nabla \cdot \boldsymbol{H}) - \nabla^2 \boldsymbol{H} = -\nabla^2 \boldsymbol{H}$$

and, if it can be assumed that there are no free charges in the medium, that is, if

$$\nabla \cdot \boldsymbol{D} = 0 ,$$

similarly

$$\nabla \times (\nabla \times \boldsymbol{E}) = \nabla(\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E} = -\nabla^2 \boldsymbol{E} .$$

we obtain, finally

$$\boxed{\nabla^2 \boldsymbol{H} = \frac{\varepsilon\mu}{c^2} \ddot{\boldsymbol{H}} + \frac{4\pi\sigma\mu}{c^2} \dot{\boldsymbol{H}}} \qquad (2.24)$$

$$\boxed{\nabla^2 \boldsymbol{E} = \frac{\varepsilon\mu}{c^2} \ddot{\boldsymbol{E}} + \frac{4\pi\sigma\mu}{c^2} \dot{\boldsymbol{E}}} \quad . \qquad (2.25)$$

Both $\boldsymbol{E}$ and $\boldsymbol{H}$ obey the same inhomogeneous wave equation, a linear second order partial differential equation. Since these equations are derived from Maxwell's equations, every solution of these will also be a solution of the wave equation. The reverse conclusion is not true under all conditions. For example, in (2.24) and (2.25) the $\boldsymbol{E}$ and the $\boldsymbol{H}$ fields are decoupled, and therefore any arbitrary solution for $\boldsymbol{E}$ can be coupled to any solution for $\boldsymbol{H}$ provided that they obey the initial conditions. In Maxwell's equations this is not true; here $\boldsymbol{E}$ and $\boldsymbol{H}$ are interdependent. For simple cases it is rather easy to specify which $\boldsymbol{H}$ solution belongs to a given $\boldsymbol{E}$ solution of Maxwell's equations; for more complicated situations other methods must be used. Some of these will be outlined in Chap. 6; here a direct solution of the wave equation should suffice to show the principle.

2.5   Plane Waves in Nonconducting Media

Consider a homogeneous, nonconducting medium ($\sigma = 0$) that is free of currents and charges. In rectangular coordinates each vector component u of $\boldsymbol{E}$ and $\boldsymbol{H}$ obeys the homogeneous wave equation

$$\boxed{\nabla^2 u - \frac{1}{v^2}\ddot{u} = 0}\quad, \tag{2.26}$$

where

$$\boxed{v = \frac{c}{\sqrt{\varepsilon\mu}}} \tag{2.27}$$

is a constant with the dimension of velocity. For the vacuum this becomes

$$v = c\,. \tag{2.28}$$

When Kohlrausch and Weber in 1856 obtained this result experimentally, it became one of the basic facts used by Maxwell when he developed his electromagnetic theory predicting the existence of electromagnetic waves. Eventually this prediction was confirmed experimentally by Hertz (1888).

Equation (2.26) is a homogeneous linear partial differential equation of second order. The complete family of solutions forms a wide and sometimes rather complicated group. No attempt will be made here to discuss general solutions, rather we will restrict our presentation to the properties of the harmonic waves.

$$u = u_0\,e^{i(kx \pm \omega t)} \tag{2.29}$$

is a solution of (2.26) if the wave number k obeys the relation

$$\boxed{k^2 = \frac{\varepsilon\mu}{c^2}\,\omega^2} \tag{2.30}$$

This can be confirmed by the substitution of (2.29) into (2.26). If we set

$$\varphi = kx \pm \omega t, \tag{2.31}$$

where φ is the phase of the wave, we see that points of constant phase move with the phase velocity

$$\boxed{v = \frac{\omega}{k} = \frac{c}{\sqrt{\varepsilon\mu}}} \quad, \tag{2.32}$$

This gives a physical meaning to the constant v appearing in (2.26). Introducing the index of refraction n as the ratio of c to v this becomes

$$\boxed{n = \frac{c}{v} = \sqrt{\varepsilon\mu} = \frac{c}{\omega} k} \quad . \tag{2.33}$$

For plane electromagnetic waves, each component of E and H will have solutions (2.29) but with an amplitude, u_0, that generally is complex. The use of (2.29) permits us to introduce some important simplifications. For a traveling plane wave

$$A(\boldsymbol{x},t) = \boldsymbol{A}_0 \, e^{i(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)}, \qquad \boldsymbol{A}_0, \boldsymbol{k}, \omega = \text{const.}, \tag{2.34}$$

$$\dot{\boldsymbol{A}} = -i\,\omega\boldsymbol{A}, \tag{2.35}$$

$$\ddot{\boldsymbol{A}} = -\omega^2\boldsymbol{A}, \tag{2.36}$$

$$\nabla\cdot\boldsymbol{A} = i\boldsymbol{k}\cdot\boldsymbol{A}, \tag{2.37}$$

$$\nabla^2\boldsymbol{A} = -k^2\boldsymbol{A}. \tag{2.38}$$

The E and H fields of an electromagnetic wave are not only solutions of the wave equation (2.26), but these also must obey Maxwell's equations. Because of the decoupling of the two fields in the wave equation, this produces some additional constraints.

In order to investigate the properties of plane waves as simply as possible, we arrange the rectangular coordinate system such that the wave propagates in the positive z direction. A wave is considered to be plane if the surfaces of constant phase form planes $z = \text{const}$. Thus all components of the E and the H field will be independent of x and y for fixed z; that is,

$$\frac{\partial E_x}{\partial x} = 0, \qquad \frac{\partial E_y}{\partial x} = 0, \qquad \frac{\partial E_z}{\partial x} = 0,$$
$$\frac{\partial E_x}{\partial y} = 0, \qquad \frac{\partial E_y}{\partial y} = 0, \qquad \frac{\partial E_z}{\partial y} = 0, \tag{2.39}$$

and a similar set of equations for H. But according to Maxwell's equations (2.4) and (2.5) with $\varrho = 0$ and $\varepsilon = \text{const.}$

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0 \qquad \text{and} \qquad \frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} + \frac{\partial H_z}{\partial z} = 0.$$

Because of (2.39) this results in

$$\frac{\partial E_z}{\partial z} = 0 \quad \text{and} \quad \frac{\partial H_z}{\partial z} = 0 \quad . \tag{2.40}$$

From the remaining Maxwell's equations (2.6) and (2.7) we similarly obtain

$$\frac{\partial E_z}{\partial t} = 0 \quad \text{and} \quad \frac{\partial H_z}{\partial t} = 0 \quad . \tag{2.41}$$

Therefore both the longitudinal components E_z and H_z must be constant both in space and time. Since such a constant field is of no significance here, we require that

$$E_z \equiv 0, \; H_z \equiv 0 \tag{2.42}$$

that is, the plane electromagnetic wave in a nonconducting medium is *transverse* (Fig. 2.2). The remaining components have the form of traveling harmonic waves [as given by (2.29)]. The only components of (2.6) and (2.7) which differ from zero are

$$\frac{\partial E_x}{\partial z} = -\frac{\mu}{c}\frac{\partial H_y}{\partial t}, \qquad \qquad \frac{\partial H_x}{\partial z} = \frac{\varepsilon}{c}\frac{\partial E_y}{\partial t},$$
$$\text{and} \tag{2.43}$$
$$\frac{\partial E_y}{\partial z} = \frac{\mu}{c}\frac{\partial H_x}{\partial t}, \qquad \qquad \frac{\partial H_y}{\partial z} = -\frac{\varepsilon}{c}\frac{\partial E_x}{\partial t}.$$

Applying the relations (2.35) and (2.37) for plane harmonic waves, we find

$$\frac{\partial E_x}{\partial z} = \mathrm{i}kE_x = -\frac{\mu}{c}\dot{H}_y = \frac{\mathrm{i}\omega\mu}{c}H_y,$$
$$\tag{2.44}$$
$$\frac{\partial E_y}{\partial z} = \mathrm{i}kE_y = \frac{\mu}{c}\dot{H}_x = -\frac{\mathrm{i}\omega\mu}{c}H_x,$$
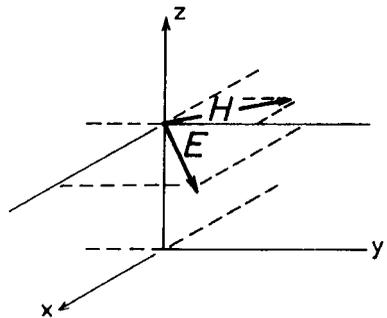
Fig. 2.2 A sketch of the field vectors in a plane electromagnetic wave propagating in the z-direction

resulting in

$$\boldsymbol{E} \cdot \boldsymbol{H} = E_x H_x + E_y H_y = -\frac{ck}{\omega\mu} E_x E_y + \frac{ck}{\omega\mu} E_y E_x = 0\,,$$

$$\boxed{\boldsymbol{E} \cdot \boldsymbol{H} = 0} \quad . \tag{2.45}$$

$\boldsymbol{E}$ and $\boldsymbol{H}$ are thus always perpendicular; together with the wave vector $\boldsymbol{k}$, these form an orthogonal system. For the ratio of their absolute values, (2.44) and (2.30) result in

$$\frac{|\boldsymbol{E}|}{|\boldsymbol{H}|} = \sqrt{\frac{\mu}{\varepsilon}}\,. \tag{2.46}$$

The unit of this *intrinsic impedance* of the medium in which the wave propagates is the Ohm (Ω). In a vacuum it has the value

$$Z_0 = 376.73\,\Omega\,. \tag{2.47}$$

Finally, the energy flux of the Poynting vector of this wave is of interest. As given by (2.12) we find

$$|\boldsymbol{S}| = \frac{c}{4\pi}\sqrt{\frac{\varepsilon}{\mu}}E^2\,, \tag{2.48}$$

and $\boldsymbol{S}$ points in the direction of the propagation vector $\boldsymbol{k}$. The (time averaged) energy density, u, of the wave given by (2.9) is then[1]

$$u = \frac{1}{8\pi}\left(\varepsilon\boldsymbol{E}\cdot\boldsymbol{E}^* + \mu\boldsymbol{H}\cdot\boldsymbol{H}^*\right)\,. \tag{2.49}$$

The argument used in this is quite similar to that used in deriving (2.21). In using (2.46) we find that (2.49) becomes

$$u = \frac{\varepsilon}{4\pi}E^2\,. \tag{2.50}$$

The time averaged Poynting vector is often used as a measure of the intensity of the wave; its direction represents the direction of the wave propagation.

2.6　Wave Packets and the Group Velocity

A monochromatic plane wave

$$u(x,t) = A\,\mathrm{e}^{\mathrm{i}(kx-\omega t)} \tag{2.51}$$

propagates with the phase velocity

---

[1] This energy density should not be confused with the Cartesian component u of $\boldsymbol{E}$ or $\boldsymbol{H}$ in (2.26) and following.

$$v = \frac{\omega}{k}. \tag{2.52}$$

If this velocity is the same for a whole range of frequencies, then a wave packet formed by the superposition of these waves will propagate with the same velocity. In general, however, the propagation velocity, v, will depend on the wave number k. Then such wave packets have some new and interesting properties. A wave with an arbitrary shape can be formed by superposing simple harmonic waves

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} A(k) \, \mathrm{e}^{\,\mathrm{i}\,(kx - \omega t)} \, \mathrm{d}k, \tag{2.53}$$

where $A(k)$ is the amplitude of the wave with the wave number k. The angular frequency of these waves will be different for different k; this distribution is

$$\omega = \omega(k) \tag{2.54}$$

and it will be referred to as the *dispersion equation* of the waves. If $A(k)$ is a fairly sharply peaked function around some k_0, only waves with wave numbers not too different from k_0 will contribute to (2.53), and quite often a linear approximation for (2.54)

$$\omega(k) = \omega_0 + \left.\frac{\mathrm{d}\omega}{\mathrm{d}k}\right|_0 (k - k_0) \tag{2.55}$$

will be sufficient. The symbol after the derivative indicates that it will be evaluated at $k = 0$. Substituting this into (2.53) we can extract all factors that do not depend on k from the integral, obtaining

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \exp\left[\mathrm{i}\left(\left.\frac{\mathrm{d}\omega}{\mathrm{d}k}\right|_0 k_0 - \omega_0\right)t\right] \int\limits_{-\infty}^{\infty} A(k) \exp\left[\mathrm{i}k\left(x - \left.\frac{\mathrm{d}\omega}{\mathrm{d}k}\right|_0 t\right)\right] \mathrm{d}k.$$
$$\tag{2.56}$$

According to (2.53), at the time $t = 0$ the wave packet has the shape

$$u(x,0) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} A(k) \, \mathrm{e}^{\,\mathrm{i}kx} \, \mathrm{d}k.$$

Therefore the integral in (2.56) is $u(x',0)$, where $x' = x - \left.\frac{\mathrm{d}\omega}{\mathrm{d}k}\right|_0 t$. The entire expression is

$$u(x,t) = u\left(x - \left.\frac{\mathrm{d}\omega}{\mathrm{d}k}\right|_0 t, 0\right) \exp\left[\mathrm{i}\left(k_0 \left.\frac{\mathrm{d}\omega}{\mathrm{d}k}\right|_0 - \omega_0\right)t\right]. \tag{2.57}$$

The exponential in (2.57) has a purely imaginary argument and therefore is only a phase factor. Therefore, the wave packet travels undistorted in shape except for an overall phase factor with the group velocity

$$v_g = \frac{d\omega}{dk} \quad . \tag{2.58}$$

This is strictly true if the angular frequency is a linear function of k. If $\omega(k)$ is more general, the group velocity depends on wave number, and the form of the wave packet (made up of waves with a finite range of wave numbers) will be distorted in time. That is, the pulse will disperse.

Whether phase velocity (2.52) or group velocity (2.58), is larger depends on the properties of the medium in which the wave propagates. Writing (2.52) as

$$\omega = kv,$$

one finds

$$\frac{d\omega}{dk} = v_g = v + k\frac{dv}{dk}. \tag{2.59}$$

Recalling the definition of the index of refraction (2.33)

$$n = \frac{c}{v}$$

and that the wavelength is given by

$$\lambda = \frac{2\pi}{k}, \tag{2.60}$$

we see that normal dispersion $dn/d\lambda < 0$ in the medium corresponds to $dv/dk < 0$. In a medium with normal dispersion therefore $v_g < v$. Only for anomalous dispersion will we have $v_g > v$.

Energy and information are usually propagated with the group velocity. The situation is, however, fairly complicated if propagation in dispersive media is considered. These problems have been investigated by Sommerfeld (1914) and Brillouin (1914). Details can be found in Sommerfeld (1959).

2.7 Plane Waves in Conducting Media

In Sect. 2.5 the propagation properties of plane harmonic waves in a *nonconducting* ($\sigma = 0$) medium have been investigated. Now this assumption will be dropped so that $\sigma \neq 0$, but we still restrict the investigation to strictly harmonic waves propagating in the direction of increasing x

$$\boldsymbol{E}(x,t) = \boldsymbol{E}_0 \, e^{i(kx-\omega t)}. \tag{2.61}$$

Both $\boldsymbol{E}_0$ and k are complex constants. Making use of (2.35) to (2.38), the wave equations (2.24) and (2.25) become

$$\boxed{\left[k^2 - \left(\frac{\varepsilon\mu}{c^2}\omega^2 + i\frac{4\pi\sigma\mu\,\omega}{c^2}\right)\right]\begin{Bmatrix} E \\ H \end{Bmatrix} = 0}\quad . \tag{2.62}$$

If these equations are to be valid for arbitrary E or H [of the form (2.61)] the square bracket must be zero, so that the *dispersion equation* becomes

$$\boxed{k^2 = \frac{\mu\varepsilon\,\omega^2}{c^2}\left(1 + i\frac{4\pi\sigma}{\omega\varepsilon}\right)}\quad . \tag{2.63}$$

The wave number k thus is indeed a complex number. Writing

$$k = a + ib, \tag{2.64}$$

we find

$$a = \sqrt{\varepsilon\mu}\,\frac{\omega}{c}\sqrt{\frac{1}{2}\left(\sqrt{1 + \left(\frac{4\pi\sigma}{\varepsilon\,\omega}\right)^2} + 1\right)} \tag{2.65}$$

$$b = \sqrt{\varepsilon\mu}\,\frac{\omega}{c}\sqrt{\frac{1}{2}\left(\sqrt{1 + \left(\frac{4\pi\sigma}{\varepsilon\,\omega}\right)^2} - 1\right)} \tag{2.66}$$

and the field therefore can be written

$$E(x,t) = E_0\,e^{-bx}\,e^{i(ax - \omega t)}\,. \tag{2.67}$$

Thus the real part of the conductivity gives rise to an exponential damping of the wave. If (2.67) is written using the index of refraction n and the absorption coefficient κ,

$$\boxed{E(x,t) = E_0\exp\left(-\frac{\omega}{c}n\kappa x\right)\exp\left[i\omega\left(\frac{n}{c}x - t\right)\right]}\quad , \tag{2.68}$$

we obtain

$$\boxed{n\kappa = \sqrt{\varepsilon\mu}\sqrt{\frac{1}{2}\left(\sqrt{1 + \left(\frac{4\pi\sigma}{\varepsilon\,\omega}\right)^2} - 1\right)}} \tag{2.69}$$

$$\boxed{n = \sqrt{\varepsilon\mu}\sqrt{\frac{1}{2}\left(\sqrt{1 + \left(\frac{4\pi\sigma}{\varepsilon\,\omega}\right)^2} + 1\right)}}\quad . \tag{2.70}$$

2.8 The Dispersion Measure of a Tenuous Plasma

The simplest model for a dissipative medium is that of a tenuous plasma where free electrons and ions are uniformly distributed so that the total space charge density is zero. This model was first given by Drude (1900) to explain the propagation of ultraviolet light in a transparent medium, but this model was later applied to the propagation of transverse electromagnetic radio waves in a tenuous plasma.

The free electrons are accelerated by the electric field intensity; their equation of motion is

$$m_e \dot{\boldsymbol{v}} = m_e \ddot{\boldsymbol{r}} = -e \boldsymbol{E}_0 \, e^{-i\omega t} \tag{2.71}$$

with the solution

$$\boldsymbol{v} = \frac{e}{i m_e \, \omega} \boldsymbol{E}_0 \, e^{-i\omega t} = -i \frac{e}{m_e \, \omega} \boldsymbol{E} \,. \tag{2.72}$$

Equation (2.72) describes the motion of the electrons. Moving electrons, however, carry a current, whose density is

$$\boldsymbol{J} = -\sum_\alpha e \boldsymbol{v}_\alpha = -N e \boldsymbol{v} = i \frac{N e^2}{m_e \, \omega} \boldsymbol{E} = \sigma \boldsymbol{E} \,. \tag{2.73}$$

This expression explains why the ions can be neglected in this investigation. Due to their large mass ($m_i \simeq 2 \times 10^3 \, m_e$), the induced ion velocity (2.72) is smaller than that of the electrons by the same factor, and since the charge of the ions is the same as that of the electrons, the ion current (2.73) will be smaller than the electron current by the same factor.

According to (2.73) the conductivity of the plasma is purely imaginary:

$$\sigma = i \frac{N e^2}{m_e \, \omega} \,. \tag{2.74}$$

Inserting this into (2.63) we obtain, for a thin medium with $\varepsilon \approx 1$ and $\mu \approx 1$

$$\boxed{k^2 = \frac{\omega^2}{c^2} \left(1 - \frac{\omega_p^2}{\omega^2} \right)} \,, \tag{2.75}$$

where

$$\boxed{\omega_p^2 = \frac{4\pi N e^2}{m_e}} \tag{2.76}$$

is the square of the *plasma frequency*. It gives a measure of the mobility of the electron gas. Inserting numerical values we obtain

$$\frac{\nu_p}{\text{kHz}} = 8.97 \sqrt{\frac{N}{\text{cm}^{-3}}} \tag{2.77}$$

if we convert (2.76) to frequencies by $v = \omega/2\pi$. For $\omega > \omega_p$, k is real, and we obtain from (2.52)

$$v = \frac{c}{\sqrt{1 - \dfrac{\omega_p^2}{\omega^2}}} \tag{2.78}$$

for the *phase velocity* v and so $v > c$ for $\omega > \omega_p$. For the *group velocity* it follows from (2.58)

$$v_g = \frac{d\omega}{dk} = \frac{1}{dk/d\omega},$$

so that

$$v_g = c\sqrt{1 - \frac{\omega_p^2}{\omega^2}} \tag{2.79}$$

and $v_g < c$ for $\omega > \omega_p$. Both v and v_g thus depend on the frequency ω. For $\omega = \omega_p$, $v_g = 0$; thus for waves with a frequency lower than ω_p, no wave propagation in the plasma is possible. The frequency dependence of v and v_g are in the opposite sense; taking (2.78) and (2.79) together the relation

$$v\,v_g = c^2 \tag{2.80}$$

is obtained.

For some applications the *index of refraction* is a useful quantity. According to (2.33) and (2.75) it is

$$n = \sqrt{1 - \frac{\omega_p^2}{\omega^2}}\quad. \tag{2.81}$$

Electromagnetic pulses propagate with the group velocity. This varies with frequency so that there is a dispersion in the pulse propagation in a plasma. This fact took on a fundamental importance when the radio pulsars were detected in 1967. The arrival time of pulsar pulses depends on the frequency: The lower the observing frequency, the later the pulse arrives. This behavior can easily be explained in terms of wave propagation in a tenuous plasma, as the following discussion shows.

The plasma frequency of the interstellar medium (ISM) is much lower than the observing frequency. In the ISM, N is typically 10^{-3}–$10^{-1}\,\mathrm{cm}^{-3}$, so v_p is in the range 2.85–0.285 kHz; however, the observing frequency must be $v > 10\,\mathrm{MHz}$ in order to propagate through the ionosphere of the earth. For v_g, we can use a series expansion of (2.79)

$$\frac{1}{v_g} = \frac{1}{c}\left(1 + \frac{1}{2}\frac{v_p^2}{v^2}\right) \tag{2.82}$$

with high precision. A pulse emitted by a pulsar at the distance L therefore will be received after a delay

$$\tau_D = \int_0^L \frac{dl}{v_g} \cong \frac{1}{c} \int_0^L \left(1 + \frac{1}{2}\left(\frac{v_p}{v}\right)^2\right) dl = \frac{1}{c} \int_0^L \left(1 + \frac{e^2}{2\pi m_e} \frac{1}{v^2} N(l)\right) dl,$$

$$\tau_D = \frac{L}{c} + \frac{e^2}{2\pi c m_e} \frac{1}{v^2} \int_0^L N(l)\, dl. \tag{2.83}$$

The difference between the pulse arrival times measured at two frequencies v_1 and v_2 therefore is given by

$$\boxed{\Delta\tau_D = \frac{e^2}{2\pi c m_e}\left[\frac{1}{v_1^2} - \frac{1}{v_2^2}\right]\int_0^L N(l)\, dl} \quad . \tag{2.84}$$

The quantity $\int_0^L N(l)\, dl$ is the column-density of the electrons in the intervening space between pulsar and observer. Since distances in astronomy are measured in parsecs ($1\,\mathrm{pc} = 3.085677 \times 10^{18}$ cm), it has become customary to measure $N(l)$ in cm^{-3} but dl in pc. The integral then is called the *dispersion measure* (Fig. 2.3)

$$\boxed{\mathrm{DM} = \int_0^\infty \left(\frac{N}{\mathrm{cm}^{-3}}\right) d\left(\frac{l}{\mathrm{pc}}\right)} \tag{2.85}$$
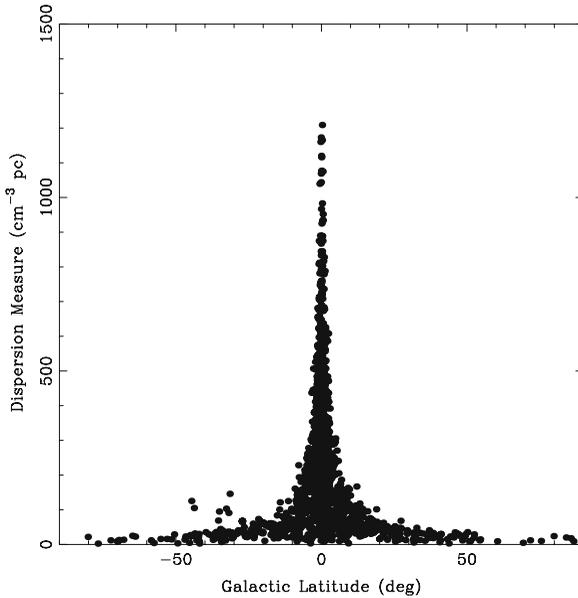
Fig. 2.3 Dispersion measure, DM, for pulsars at different galactic latitudes [adapted from B. Klein (MPIfR) unpublished]

and therefore we find

$$\frac{\Delta \tau_D}{\mu s} = 1.34 \times 10^{-9} \left[\frac{DM}{cm^{-2}} \right] \left[\frac{1}{\left(\frac{\nu_1}{MHz} \right)^2} - \frac{1}{\left(\frac{\nu_2}{MHz} \right)^2} \right] \qquad (2.86)$$

or

$$\frac{\Delta \tau_D}{\mu s} = 4.148 \times 10^9 \left[\frac{DM}{cm^{-3} pc} \right] \left[\frac{1}{\left(\frac{\nu_1}{MHz} \right)^2} - \frac{1}{\left(\frac{\nu_2}{MHz} \right)^2} \right] \qquad . \qquad (2.87)$$

Since both the time delay $\Delta \tau_D$ and the observing frequencies ν_1 and ν_2 can be measured with high precision, a very accurate value of DM for a given pulsar can be determined from

$$\frac{DM}{cm^{-3} pc} = 2.410 \times 10^{-4} \left(\frac{\Delta \tau_D}{s} \right) \left[\frac{1}{\left(\frac{\nu_1}{MHz} \right)^2} - \frac{1}{\left(\frac{\nu_2}{MHz} \right)^2} \right]^{-1} . \qquad (2.88)$$

Provided the distance L to the pulsar is known, this gives a good estimate of the average electron density between observer and pulsar. However since L is usually known only very approximately, only approximate values for N can be obtained in this way. Quite often the opposite procedure is used: From reasonable guesses for N, a measured DM provides information on the unknown distance L to the pulsar.

Dispersion in the ISM, combined with a finite pulse width, sets a limit to the fine structure on can resolve in a pulse. The frequency dependence of the pulse arrival time is τ_D from (2.83). This gives a condition for the bandwidth b needed to resolve a time feature τ

$$\frac{b}{MHz} = 1.205 \times 10^{-4} \frac{1}{\left[\frac{DM}{cm^{-3} pc} \right]} \left[\frac{\nu}{MHz} \right]^3 \frac{\tau}{s} . \qquad (2.89)$$

Since the pulses will have a finite width in both time and frequency, a differential form of (2.89) will give a limit to the maximum bandwidth that can be used at a given frequency and DM if a time resolution τ is wanted. This will be re discussed in the context of pulsar back ends.

Problems

1. There is a proposal to transmit messages to mobile telephones in large U.S. cities from a transmitter hanging below a balloon at an altitude of 40 km. Suppose the city in question has a diameter of 40 km. What is the solid angle to be illuminated?

Suppose mobile telephones require an electric field strength, E, of $200\,\mu V$ per meter. If one uses $S = E^2/R$ with $R = 50\,\Omega$, what is the E field at the transmitter? How much power must be transmitted? At what distance from the transmitter would the microwave radiation reach the danger level, $10\,mW\,cm^{-2}$?

2. Radiation from an astronomical source at a distance of $1.88\,kpc$, $(= 7.1 \times 10^{21}\,cm)$ has a flux density of $10^3\,Jy$ over a frequency band of $600\,Hz$. If it is isotropic, what is the power radiated? Suppose the source size is 1 milli arc second (see (1.34)). What is the value of T_b? Compare to the surface temperature of an O star $\approx 40,000\,K$.

3. A plane electromagnetic wave perpendicularly approaches a surface with conductivity σ. The wave penetrates to a depth of δ. Apply (2.25), taking $\sigma \gg \varepsilon/4\pi$, so $\nabla^2 E = (4\pi\sigma\mu/c^2)\dot{E}$. The solution to this equation is an exponentially decaying wave. Use this to estimate the $1/e$ penetration depth, δ.

4. Estimate the value of $\delta = c/\sqrt{4\pi\sigma\mu\omega}$ for copper, which has (in CGS units) $\sigma = 10^{17}\,s^{-1}$, and $\mu \approx 1$, for $\nu = 10^{10}\,Hz$.

5. Suppose that $\nu_{phase} = \dfrac{c}{\sqrt{1-(\lambda_0/\lambda_c)^2}}$. What is ν_{group}? Evaluate both of these quantities for $\lambda_0 = \frac{1}{2}\lambda_c$.

6. There is a 1 D wave packet. At time $t=0$, the amplitudes are distributed as $a(k) = a_0 \exp(-k^2/(\Delta k)^2)$, where a_0 and Δk are constant. From the use of Fourier transform relations in Appendix B, determine the product of the width of the wave packet, Δk, and the width in time, Δt.

7. Repeat problem 7 with $a(k) = a_0 \exp(-(k-k_0)^2/(\Delta k)^2)$.

8. Repeat problem 7 for $a(k) = a_0$ for $k_1 < k < k_2$, otherwise $a(k) = 0$.

9. Assume that pulsars emit narrow periodic pulses at all frequencies simultaneously. Use (2.83) to show that a narrow pulse (width of order $\sim 10^{-6}\,s$) will traverse the radio spectrum at a rate, in $MHz\,s^{-1}$, of $\dot{\nu} = 1.2 \times 10^{-4}\,(DM)^{-1}\,\nu\,[MHz]^3$.

10. (a) Show that using a receiver bandwidth B will lead to the smearing of a very narrow pulse, which passes through the ISM with dispersion measure DM, to a width $\Delta t = 8.3 \times 10^3\,DM\,[\nu\,[(MHz)]^{-3}\,B\,s$.
(b) Show that the ionosphere (electron density $10^5\,cm^{-3}$, height $20\,km$) has little influence on the pulse shape at $100\,MHz$.

11. (a) Show that the smearing Δt, in milli seconds, of a short pulse is $(202/\nu_{MHz})^3$ DM ms per MHz of receiver bandwidth.
(b) If a pulsar is at a distance of 5 kpc, and the average electron density is $0.05\,cm^{-3}$, find the smearing at $400\,MHz$. Repeat for $800\,MHz$.

12. Suppose you would like to detect a pulsar located at the center of our Galaxy. The pulsar may be behind a cloud of ionized gas of size $10\,pc$, and electron density $10^3\,cm^{-3}$. Calculate the dispersion measure, DM. What is the bandwidth limit if the observing frequency is $1\,GHz$, and the pulsar frequency is $30\,Hz$?

13. A typical value for DM is $30 \, \text{cm}^{-3}$ pc, which is equivalent to an electron column density of $10^{20} \, \text{cm}^{-2}$. For a frequency of 400 MHz, use (2.87) to predict how much a pulse will be delayed relative to a pulse at an infinitely high frequency. Repeat for a frequency of 1000 MHz.

14. To resolve a pulse feature with a width of 0.1 μs at a received frequency of 1000 MHz and DM $= 30 \, \text{cm}^{-3}$ pc, what is the maximum receiver bandwidth?

Chapter 3
Wave Polarization

3.1 Vector Waves

In the preceding Chapter we have shown that plane electromagnetic waves in a dielectric medium are transverse and that the x and the y component of both E and H for a wave propagating in the z direction obey the same wave equation. For the sake of simplicity, we have investigated the propagation of only one component of these fields. In this Chapter, we present the theory of polarization. This can be caused by a number of mechanisms that will be presented in Chaps. 10 and 11. In the references for this Chapter are a few papers that present the analysis and interpretation of polarization data.

In general both the x and the y component have to be specified but, in a strictly monochromatic wave, they are not independent, since both share the same harmonic dependence, although with a different phase:

$$
\begin{aligned}
E_x &= E_1 \cos\left(kz - \omega t + \delta_1\right), \\
E_y &= E_2 \cos\left(kz - \omega t + \delta_2\right), \\
E_z &= 0.
\end{aligned}
$$
(3.1)

Here $k = 2\pi/\lambda$, where λ is the wavelength in cm, and $\omega = 2\pi\nu$, where ν is frequency in Hz. Regarding (E_x, E_y, z) as the coordinates of a point in a rectangular coordinate system we find that (3.1) describes a helical path on the surface of a cylinder. The cross section of this cylinder can be determined by eliminating the phase of this wave, abbreviated by

$$
\tau = kz - \omega t.
$$
(3.2)

Rewriting the first two equations of (3.1) as

$$
\begin{aligned}
\frac{E_x}{E_1} &= \cos\tau\cos\delta_1 - \sin\tau\sin\delta_1, \\
\frac{E_y}{E_2} &= \cos\tau\cos\delta_2 - \sin\tau\sin\delta_2
\end{aligned}
$$
(3.3)

gives

$$\frac{E_x}{E_1}\sin\delta_2 - \frac{E_y}{E_2}\sin\delta_1 = \cos\tau\sin(\delta_2 - \delta_1),$$

$$\frac{E_x}{E_1}\cos\delta_2 - \frac{E_y}{E_2}\cos\delta_1 = \sin\tau\sin(\delta_2 - \delta_1).$$

Squaring and adding we obtain

$$\left(\frac{E_x}{E_1}\right)^2 + \left(\frac{E_y}{E_2}\right)^2 - 2\frac{E_x}{E_1}\frac{E_y}{E_2}\cos\delta = \sin^2\delta \qquad (3.4)$$

$$\delta = \delta_1 - \delta_2 \qquad (3.5)$$

This is the equation of an ellipse, since the discriminant is not negative

$$\begin{vmatrix} \dfrac{1}{E_1^2} & -\dfrac{\cos\delta}{E_1 E_2} \\[2mm] -\dfrac{\cos\delta}{E_1 E_2} & \dfrac{1}{E_2^2} \end{vmatrix} = \frac{1-\cos^2\delta}{E_1^2 E_2^2} = \frac{\sin^2\delta}{E_1^2 E_2^2} \geq 0. \qquad (3.6)$$

The wave is said to be elliptically polarized, and this applies to both the electric and the magnetic field of the wave; $\sin\delta$ determines the sense in which the electric vector rotates.

The ellipse (3.4) usually is arbitrarily oriented with respect to the coordinate system. Its geometric properties are seen best by selecting a coordinate system oriented along the major and minor axes (Fig. 3.1). In this system the ellipse equation is

$$\begin{aligned} E_\xi &= E_a \cos(\tau + \delta), \\ E_\eta &= E_b \sin(\tau + \delta), \end{aligned} \qquad (3.7)$$

and the relation between the coordinate systems (x,y) and (ξ,η) is given by the linear transformation

$$\begin{aligned} E_\xi &= E_x \cos\psi + E_y \sin\psi, \\ E_\eta &= -E_x \sin\psi + E_y \cos\psi. \end{aligned} \qquad (3.8)$$
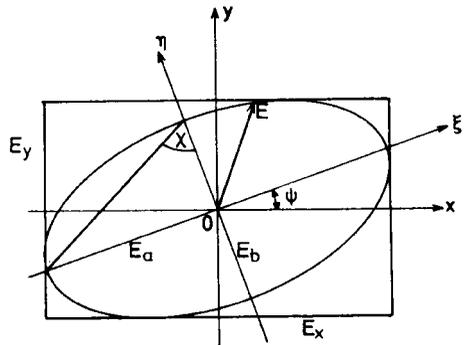
Fig. 3.1 The polarization ellipse for the electric vector, **E**, of an elliptically polarized wave

The intrinsic parameters of the polarization ellipse E_a and E_b, as well as the angle ψ by which the major axis is tilted with respect to the x axis, can then be determined by requiring that (3.4) transformed by (3.8) should lead to (3.7). Substituting (3.3) and (3.7) into (3.8) while simultaneously expanding the $\cos(\tau + \delta)$ term leads to

$$E_a (\cos \tau \cos \delta - \sin \tau \sin \delta) = E_1 (\cos \tau \cos \delta_1 - \sin \tau \sin \delta_1) \cos \psi$$
$$+ E_2 (\cos \tau \cos \delta_2 - \sin \tau \sin \delta_2) \sin \psi \quad (3.9)$$

and

$$E_b (\sin \tau \cos \delta + \cos \tau \sin \delta) = -E_1 (\cos \tau \cos \delta_1 - \sin \tau \sin \delta_1) \sin \psi$$
$$+ E_2 (\cos \tau \cos \delta_2 - \sin \tau \sin \delta_2) \cos \psi . \quad (3.10)$$

These equations are valid for all τ, i.e. also for $\tau = 0$ and $\tau = \frac{\pi}{2}$, resulting in

$$E_a \cos \delta = E_1 \cos \delta_1 \cos \psi + E_2 \cos \delta_2 \sin \psi , \quad (3.11)$$
$$-E_a \sin \delta = -E_1 \sin \delta_1 \cos \psi - E_2 \sin \delta_2 \sin \psi , \quad (3.12)$$
$$E_b \cos \delta = E_1 \sin \delta_1 \sin \psi - E_2 \sin \delta_2 \cos \psi , \quad (3.13)$$
$$E_b \sin \delta = -E_1 \cos \delta_1 \sin \psi + E_2 \cos \delta_2 \cos \psi . \quad (3.14)$$

Squaring these equations and adding we obtain

$$\boxed{S_0 \equiv E_a^2 + E_b^2 = E_1^2 + E_2^2} \quad . \quad (3.15)$$

Recalling (2.50), we find that this can be interpreted that the total Poynting flux of the polarized wave is equal to the sum of the fluxes of two orthogonal, but otherwise arbitrary directions.

Multiplying (3.11) by (3.13) and (3.12) by (3.14) and subtracting the results, we obtain

$$E_a E_b = E_1 E_2 \sin \delta , \quad (3.16)$$

while division and addition of the same pairs of equations result in

$$- (E_1^2 - E_2^2) \sin \psi \cos \psi = E_1 E_2 \cos \delta (\sin^2 \psi - \cos^2 \psi) ,$$
$$(E_1^2 - E_2^2) \sin 2\psi = 2E_1 E_2 \cos \delta \cos 2\psi . \quad (3.17)$$

If we now define α by

$$\boxed{\frac{E_1}{E_2} = \tan \alpha} \quad , \quad (3.18)$$

(3.17) can be rewritten as

$$\tan 2\psi = \frac{2E_1 E_2}{E_1^2 - E_2^2} \cos \delta = -\frac{2 \tan \alpha}{1 - \tan^2 \alpha} \cos \delta$$

or

$$\boxed{\tan 2\psi = -\tan 2\alpha \cos \delta} \quad . \tag{3.19}$$

Dividing (3.16) by (3.15) results in

$$\frac{2E_a E_b}{E_a^2 + E_b^2} = \frac{2E_1 E_2}{E_1^2 + E_2^2} \sin \delta .$$

Defining

$$\boxed{\frac{E_a}{E_b} = \tan \chi} \quad , \tag{3.20}$$

(3.19) is equivalent to

$$\boxed{\sin 2\chi = \sin 2\alpha \sin \delta} \quad . \tag{3.21}$$

Equations (3.15, 3.18, 3.19, 3.20 and 3.21) now permit the computation of all intrinsic polarization properties of the elliptically polarized wave from the intensities specified in an arbitrary coordinate system. Values for E_1, E_2 and δ (3.15) give S_0, the total intensity, while (3.19) combined with (3.18) allows the determination of the angle ψ, while the angle χ is determined from (3.21). E_a and E_b can be computed from (3.20) and (3.15).

The phase difference δ is important in several respects. Its sign determines the sense in which the wave vector is rotating. If $\sin \delta > 0$ or equivalently $\tan \chi > 0$, the polarization is called *right-handed*; conversely $\sin \delta < 0$ or $\tan \chi < 0$ describes *left-handed* elliptical polarization. For right-handed polarization, the rotation of the E vector and the direction of propagation form a right-handed screw. This convention is the one generally adopted in microwave physics and modern physical optics. According to this definition, right-handed helical beam antennas radiate or receive right-circular polarization, a result which is easy to remember. Traditional optics used a different definition resulting in just the opposite sense of rotation based on the apparent behavior of E when "viewed" face-on by the observer. Here we will follow the modern definition, but care should be taken when comparing some of our results with those in older texts.

If the phase difference is

$$\delta = \delta_1 - \delta_2 = m\pi, \quad m = 0, \pm 1, \pm 2 \ldots \tag{3.22}$$

the polarization ellipse degenerates into a straight line and E is *linearly polarized*. As we have seen, an elliptically polarized wave can be regarded as the superposition of two orthogonal linearly polarized waves.

Another important special case is that of a *circularly polarized* wave. For this

$$E_1 = E_2 = E \tag{3.23}$$

and

$$\delta = \frac{\pi}{2}(1+m), \quad m = 0, 1, \pm 2 \pm 3, \cdots, \tag{3.24}$$

so that (3.4) reduces to the equation of a circle

$$E_x^2 + E_y^2 = E \tag{3.25}$$

with the orthogonal linear components

$$E_x = E \cos \tau,$$
$$E_y = \pm E \cos\left(\tau - \frac{\pi}{2}\right). \tag{3.26}$$

From this we see that an arbitrary elliptically polarized wave can be decomposed into the sum of two circularly polarized waves, because (3.7) can be written as

$$E_\xi = E_a \cos(\tau + \delta) = (E_r + E_l) \cos(\tau + \delta),$$
$$E_\eta = E_b \sin(\tau + \delta) = (E_r - E_l) \cos\left(\tau + \delta - \frac{\pi}{2}\right).$$

Solving for E_r and E_l, we find that
$$E_r = \tfrac{1}{2}(E_a + E_b),$$
$$E_l = \tfrac{1}{2}(E_a - E_b),$$
and, for the total Poynting flux of the wave, we obtain
$$\boxed{S_0 = E_a^2 + E_b^2 = E_r^2 + E_l^2} \quad . \tag{3.28}$$

3.2   The Poincaré Sphere and the Stokes Parameters

The results of the preceding section show that three independent parameters are needed to describe the state of the polarization of a monochromatic vector wave. For this we have introduced several sets of parameters:

1) the amplitudes E_1, E_2 and the relative phase δ of two orthogonal, linearly polarized waves;
2) the amplitudes E_r and E_l, and the relative phase δ of a right- and a left-hand circularly polarized wave;
3) the major and minor axis E_a, E_b and the position angle ψ of the polarization ellipse.

Poincaré (1892) introduced another representation that permits an easy visualization of all the different states of polarization of a vector wave. If we interpret
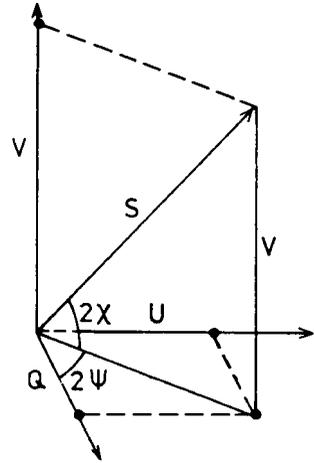
the angles 2ψ of (3.19) and 2χ of (3.21) as longitude and latitude on a sphere with
the radius S_0 of (3.15) there is a one-to-one relation between polarization states and
points on the sphere (Fig. 3.2). The equator represents linear polarization; the north
pole corresponds to right-circular and the south pole to left-circular polarization
(Fig. 3.3).

There is a natural relation between the Poincaré sphere and the Stokes parame-
ters (1852). These are the Cartesian coordinates of the points on the sphere with the
definitions:
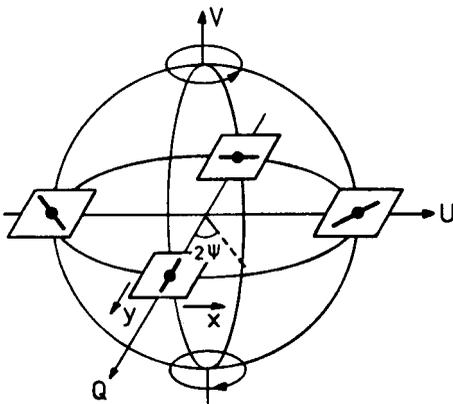
Fig. 3.3 Polarization and the Poincaré sphere. Considering the angles 2ψ and 2χ as angles in a
polar coordinate system, each point on the surface of the resulting sphere corresponds to a unique
state of polarization. The positions on the equator $(2\chi = 0)$ correspond to *linear polarization*,
those at the northern latitudes $(2\chi > 0)$ contain *right-handed* circular polarization, while those on
the southern hemisphere contain *left-handed*. If we orient the (x, y) coordinate system parallel to
Q and U, the linear polarization of the waves are oriented as indicated

$$\begin{array}{|l|}\hline S_0 = I = E_a^2 + E_b^2 \\ S_1 = Q = S_0 \cos 2\chi \cos 2\psi \\ S_2 = U = S_0 \cos 2\chi \sin 2\psi \\ S_3 = V = S_0 \sin 2\chi \\ \hline \end{array} \qquad . \tag{3.29}$$

Only three of these parameters are independent, since according to the construction of the Poincaré sphere

$$\begin{array}{|l|}\hline S_0^2 = S_1^2 + S_2^2 + S_3^2 \\ I^2 = Q^2 + U^2 + V^2 \\ \hline \end{array} \qquad . \tag{3.30}$$

The Stokes parameters can also be directly expressed by the parameters of the polarization ellipse (3.4). To do this we derive from (3.18)

$$\tan 2\alpha = \frac{2\tan\alpha}{1 - \tan^2\alpha} = -\frac{2E_1 E_2}{E_1^2 - E_2^2}, \tag{3.31}$$

$$\cos 2\alpha = \frac{1}{\sqrt{1 + \tan^2 2\alpha}} = -\frac{E_1^2 - E_2^2}{E_1^2 + E_2^2}, \tag{3.32}$$

$$\sin 2\alpha = \frac{2E_1 E_2}{E_1^2 + E_2^2}. \tag{3.33}$$

Then from (3.21), using (3.33) and (3.15),

$$\sin 2\chi = \frac{2E_1 E_2}{E_1^2 + E_2^2} \sin\delta = \frac{2E_1 E_2}{I} \sin\delta, \tag{3.34}$$

$$\cos 2\chi = \frac{1}{I} \sqrt{I^2 - (2E_1 E_2)^2 \sin^2\delta}. \tag{3.35}$$

And from (3.19) with (3.31),

$$\tan 2\psi = \frac{2E_1 E_2}{E_1^2 - E_2^2} \cos\delta \tag{3.36}$$

and

$$\cos 2\psi = \frac{1}{\sqrt{1 + \tan^2 2\psi}} = \frac{E_1^2 - E_2^2}{\sqrt{I^2 - (2E_1 E_2)^2 \sin^2\delta}}, \tag{3.37}$$

$$\sin 2\psi = \frac{2E_1 E_2 \cos \delta}{\sqrt{I^2 - (2E_1 E_2)^2 \sin^2 \delta}} \ . \tag{3.38}$$

Substituting (3.34), (3.35) and (3.37), (3.38) into (3.29) we then obtain the desired result

$$\begin{array}{l} S_0 = I = E_1^2 + E_2^2 \\ S_1 = Q = E_1^2 - E_2^2 \\ S_2 = U = 2E_1 E_2 \cos \delta \\ S_3 = V = 2E_1 E_2 \sin \delta \end{array} \tag{3.39}$$

These equations permit us to express the Stokes parameters directly in terms of observable quantities. A few special cases will illustrate the principle.

1) For a right-handed circularly polarized wave we have $E_1 = E_2$ and $\delta = \frac{\pi}{2}$, so that

$$S_0 = I = S,$$
$$S_1 = Q = 0,$$
$$S_2 = U = 0,$$
$$S_3 = V = S.$$

2) For a left-handed circularly polarized wave we have

$$S_0 = I = S,$$
$$S_1 = Q = 0,$$
$$S_2 = U = 0,$$
$$S_3 = V = -S.$$

3) For a linearly polarized wave we have $E_b = E$ and $E_a = 0$, so that $\chi = 0$ and

$$S_0 = I = E^2 = S,$$
$$S_1 = Q = I \cos 2\psi,$$
$$S_2 = U = I \sin 2\psi,$$
$$S_3 = V = 0.$$

Finally, one should note that so far we have implied (but not explicitly stated) that a strictly monochromatic wave is always polarized; there is no such thing as an unpolarized monochromatic wave. This becomes evident if we remember that for a monochromatic plane harmonic wave, E_1, E_2, δ_1 and δ_2 are always constants. This situation will be different when we consider quasi-monochromatic radiation, in which ω is restricted to some small but finite bandwidth. Radiation of this kind can be unpolarized or partially polarized. To analyze this, one must have a convenient way to describe such radiation. This will be done in the next section.

3.3 Quasi-Monochromatic Plane Waves

To this point, the description of the polarization properties of electromagnetic waves applies only to strictly monochromatic waves. The problem is how to modify the results to allow for a finite frequency interval.

Both the electric and the magnetic field intensity of the wave at a given fixed position can then be expressed by an integral of the form

$$V^{(r)}(t) = \int_0^\infty a(v)\cos[\phi(v) - 2\pi vt]\, dv. \tag{3.40}$$

Equation (3.40) has precisely the form of a Fourier integral. Therefore it is convenient to associate $V^{(r)}$ with the complex function

$$V(t) = \int_0^\infty a(v)\, e^{i[\phi(v) - 2\pi vt]}\, dv \tag{3.41}$$

where

$$V(t) = V^{(r)}(t) + i V^{(i)}(t), \tag{3.42}$$

$$V^{(i)}(t) = \int_0^\infty a(v)\sin[\phi(v) - 2\pi vt]\, dv. \tag{3.43}$$

$V^{(i)}$ does not contain information not already contained in $V^{(r)}$. V is referred to as the *analytic signal* associated with $V^{(r)}$. The integral in (3.41) formally extends over an infinite range in frequency. This allows phase to be determined. Frequently $a(v)$ has a form that effectively limits this range to an interval Δv which is small compared with the mean frequency $\bar{v}$; i.e.,

$$\Delta v / \bar{v} \ll 1. \tag{3.44}$$

If this condition is fulfilled, the signal is said to be *quasi-monochromatic*. If we express V in the form

$$V(t) = A(t)\, e^{i[\Phi(t) - 2\pi \bar{v}t]}, \tag{3.45}$$

$A(t)$ will only vary slowly with t, if the bandwidth Δv of the signal is small. However, even this variation is often too rapid to be directly measured; all that is really needed is some kind of time average. Such an average will be denoted by $\langle \cdots \rangle$:

$$\langle F(t) \rangle = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} F(t)\, dt \tag{3.46}$$

so that

$$\langle A^2(t)\rangle = \langle VV^*\rangle = \lim_{T\to\infty}\frac{1}{2T}\int\limits_{-T}^{T} V(t)V^*(t)\,\mathrm{d}t\,. \qquad (3.47)$$

If we require that $\langle A^2\rangle$ has a finite value, then $\int_{-\infty}^{\infty}VV^*\,\mathrm{d}t$ diverges. However, according to Wiener (1949), the techniques of Fourier analysis can be extended to such a generalized harmonic analysis; therefore we will assume that time averaged values for A can be computed from (3.47) This will be rediscussed in Sect. 4.1. We give an example of a quasi-monochromatic wave in problem $8ff$.

3.4 The Stokes Parameters for Quasi-Monochromatic Waves

The observable intensity of a wave is given by its time averaged Poynting flux which is, apart from a constant that is of no importance in this connection given by

$$I(P) = \langle V(P,t)V^*(P,t)\rangle\,. \qquad (3.48)$$

Let us now consider a quasi-monochromatic wave of frequency ν propagating in the z direction:

$$E_x(t) = a_1(t)\,\mathrm{e}^{\mathrm{i}(\phi_1(t)-2\pi\bar{\nu}t)}\,, \qquad E_y(t) = a_2(t)\,\mathrm{e}^{\mathrm{i}(\phi_2(t)-2\pi\bar{\nu}t)} \qquad (3.49)$$

where E_x and E_y are the *analytic signals* associated with the components $E_x^{(r)}(t) = a_1(t)\cos[\phi_1(t) - 2\pi\bar{\nu}t]$ and $E_y^{(r)}(t) = a_2(t)\cos[\phi_2(t) - 2\pi\bar{\nu}t]$. If the y component is retarded in phase by ε relative to the x component, then the electric vector in the θ direction is

$$E(t;\theta,\varepsilon) = E_x\cos\theta + E_y\,\mathrm{e}^{\mathrm{i}\varepsilon}\sin\theta \qquad (3.50)$$

and the intensity in this polarization angle is

$$I(\theta,\varepsilon) = \langle E(t;\theta,\varepsilon)E^*(t;\theta,\varepsilon)\rangle\,, \qquad (3.51)$$

The Stokes parameters of a quasi-monochromatic wave are straightforward generalizations of the expressions in (3.39). For the wave field (3.49), they are

$$\boxed{\begin{aligned} S_0 &= I = \langle a_1^2\rangle + \langle a_2^2\rangle \\ S_1 &= Q = \langle a_1^2\rangle - \langle a_2^2\rangle \\ S_2 &= U = 2\langle a_1 a_2\cos\delta\rangle \\ S_3 &= V = 2\langle a_1 a_2\sin\delta\rangle \end{aligned}} \qquad (3.52)$$

and these can be calculated from 6 intensity measurements. Using (3.51) we find

$$
\boxed{
\begin{aligned}
S_0 &= I = I(0°,0) + I(90°,0) \\
S_1 &= Q = I(0°,0) - I(90°,0) \\
S_2 &= U = I(45°,0) - I(135°,0) \\
S_3 &= V = I(45°,\tfrac{\pi}{2}) - I(135°,\tfrac{\pi}{2})
\end{aligned}
}
\tag{3.53}
$$

These are the relationships used to analyze the outputs of radio polarimeters. We will return to this later. For partially polarized light we find from (3.52)

$$
\boxed{
\begin{aligned}
S_0^2 &\overset{\geq}{=} S_1^2 + S_2^2 + S_3^3 \\
I^2 &\overset{\geq}{=} Q^2 + U^2 + V^2
\end{aligned}
}
\tag{3.54}
$$

instead of (3.30), which is valid for strictly monochromatic waves. It is then easy to express the *degree of polarization*

$$
\boxed{ p = \frac{\sqrt{S_1^2 + S_2^2 + S_3^3}}{S_0} }
\tag{3.55}
$$

The Stokes parameters of the superposition of several independent vector waves will be the sum of the Stokes parameters of the individual waves.

3.5 Faraday Rotation

In 1845, Faraday detected that the polarization angle of dielectric materials will rotate if a magnetic field is applied to the material in the direction of the light propagation. This indicated to him that light must be an electromagnetic phenomenon. In radio astronomy this Faraday rotation has become an important tool to investigate the interstellar magnetic field (see, e.g., Fig. 3.4). As shown in Sect. 2.8 interstellar gas must be treated as a tenuous plasma. Wave propagation in such a medium in the presence of an external magnetic field is a rather complicated subject with many different wave modes, cut-offs, etc. It is treated rather extensively in most textbooks on plasma physics and we refer to a few of these in the reference list for this chapter.

Here we will disregard all these complications and treat only the one remaining mode in the high-frequency limit where the frequency of the electromagnetic wave is well above all the resonances, though still low enough that the interaction of the free electrons in the plasma with the external magnetic field cannot be neglected altogether. Since the effects of wave propagation in the direction of the magnetic field are so much larger than those of propagation perpendicular to the field, only this case will be considered.

In Sect. 2.7, we have obtained the dispersion equation linking wave number $k = 2\pi/\lambda$, and circular frequency $\omega = 2\pi\nu$ for wave propagation in a dispersive
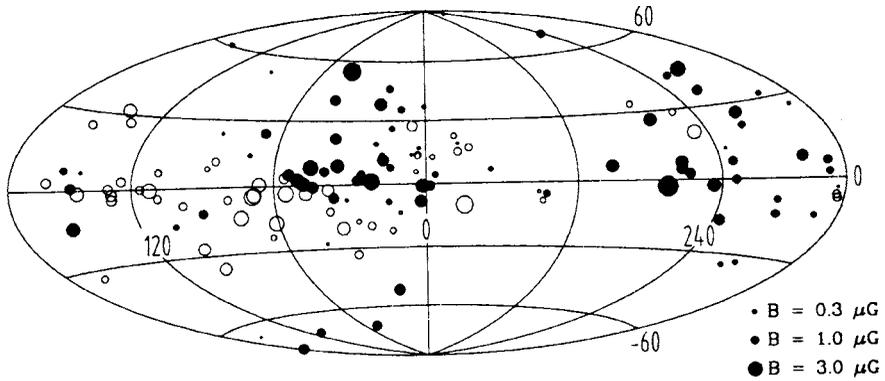
Fig. 3.4 A plot of the line-of-sight magnetic field strength determined from Faraday rotation. From the rotation measure and dispersion measure one can obtain the column density of electrons. This data is for pulsars with distances < 3 kpc. Positive fields are shown by filled circles, negative fields by open circles. The size of the symbols are proportional to field strength [adapted from Backer, in Verschuur and Kellermann (1988)]

medium. In Sect. 2.8, we studied wave propagation in a tenuous plasma by examining the effects of the conductivity σ on an electromagnetic wave in a medium with free electrons. Here we will repeat this process but will include an external magnetic field.

To be exact, the material constants ε, μ and σ should be treated as tensors with 9 components each. However, by choosing a small angle between the direction of the magnetic field and the propagation direction and a high enough frequency, we can use scalar values for ε, μ and σ.

We assume that the interstellar gas is a tenuous plasma with free electrons and ions. As in Sect. 2.8, only electrons need to be considered, since the motion of the ions is at least three orders of magnitude less than that of the electrons. The equation of motion for an electron in the presence of a magnetic field $\boldsymbol{B}$ is

$$m\dot{\boldsymbol{v}} = m\ddot{\boldsymbol{r}} = -e\left(\boldsymbol{E} + \frac{1}{c}\dot{\boldsymbol{r}} \times \boldsymbol{B}\right). \tag{3.56}$$

If the magnetic field $\boldsymbol{B}$ is oriented in the z direction (3.56) becomes

$$\begin{aligned}
\ddot{r}_x + \frac{e}{mc}B\dot{r}_y &= -\frac{e}{m}E_x \\
\ddot{r}_y - \frac{e}{mc}B\dot{r}_x &= -\frac{e}{m}E_y.
\end{aligned} \tag{3.57}$$

Multiplying (3.57) by the factors 1 and $\pm\mathrm{i}$ and adding, this becomes

$$\boxed{\begin{aligned}
\ddot{r}_\pm \mp \mathrm{i}\frac{e}{mc}B\dot{r}_\pm &= -\frac{e}{m}E_\pm \\
r_\pm &= r_x \pm \mathrm{i}r_y \\
E_\pm &= E_x \pm \mathrm{i}E_y
\end{aligned}} \tag{3.58}$$
.

Equation (3.58) is a differential equation for the complex quantities r and E. Depending on the sign of $\mathrm{i}\,(e/mc)\,B\dot{r}$, we distinguish between the solutions E_+ and E_-. These can be regarded as circularly polarized waves because the rectangular coordinates are given by

$$E_x = \frac{1}{2}\,(E_+ + E_-), \quad E_y = \frac{1}{2\mathrm{i}}\,(E_+ - E_-). \tag{3.59}$$

To obtain solutions of (3.58) in the form of a harmonic wave we put

$$E_\pm = A\,\mathrm{e}^{\mathrm{i}(k_\pm z - \omega t)} \tag{3.60}$$

where A is assumed to be real. Inserting this into (3.58), we see a solution for r of the form

$$r_\pm = r_0\,\mathrm{e}^{\mathrm{i}(k_\pm z - \omega t)} \tag{3.61}$$

with r_0 being in general a complex quantity. This is possible provided that

$$r_\pm \left(-\omega^2 \mp \frac{e}{mc}\,B\omega\right) = -\frac{e}{m}E_\pm$$

or

$$r_\pm = \frac{-\frac{e}{m}}{-\omega^2 \mp \frac{e}{mc}\,B\omega}\,E_\pm \tag{3.62}$$

and

$$\dot{r}_\pm = \frac{\mathrm{i}\,\frac{e}{m}}{-\omega^2 \mp \frac{e}{mc}\,B\omega}\,\omega E_\pm\,.$$

Thus, we find a current density

$$|\,J\,| = -Ne\,\dot{r}_\pm = \mathrm{i}\,\frac{Ne^2}{m\left(\omega \pm \frac{e}{mc}\,B\right)}\,E_\pm = \sigma_\pm E_\pm$$

with

$$\sigma_\pm = \mathrm{i}\,\frac{Ne^2}{m\left(\omega \pm \frac{e}{mc}\,B\right)} \quad . \tag{3.63}$$

The conductivity therefore is purely imaginary. For $\omega = \omega_c$, where

$$\boxed{\begin{aligned} \omega_c &= \frac{e}{mc}\,B \\ \nu_c &= \frac{e}{2\pi mc}\,B \end{aligned}} \tag{3.64}$$

is the *cyclotron frequency*, the frequency of the wave is in resonance with the gyration frequency of the electrons in the magnetic field. Then $|\sigma_-| \to \infty$, and the E_- wave cannot propagate. This is seen most easily when (3.63) is substituted in the dispersion equation (2.63), again assuming $\varepsilon \cong 1$ and $\mu \cong 1$. Then

$$k_{\pm}^2 = \frac{\omega^2}{c^2}\left(1 - \frac{\omega_p^2}{\omega(\omega \pm \omega_c)}\right) \quad, \qquad (3.65)$$

where we have introduced the plasma frequency (2.76). The index of refraction thus becomes according to (2.33)

$$n_{\pm}^2 = 1 - \frac{\omega_p^2}{\omega(\omega \pm \omega_c)} \quad, \qquad (3.66)$$

and consequently, the two modes E_+ and E_- have slightly different phase propagation velocities $v_{\pm} = c/n_{\pm}$. Then the two circularly polarized waves E_+ and E_- will have a relative phase difference $2\Delta\psi$ after a propagation length Δz given by

$$2\Delta\psi = (k_+ - k_-)\Delta z. \qquad (3.67)$$

The two circularly polarized waves can be superposed to form an elliptically polarized wave. If one does this first for the original wave, and then after the wave has left the slab Δz, we find that the polarization angle has changed by $\Delta\psi$.

Truncating the series expansions of (3.65) after the second term, which is permissible for $\omega \gg \omega_p$ and $\omega \gg \omega_c$, we obtain

$$\Delta\psi = \frac{\omega_p^2 \omega_c}{2c\,\omega^2}\Delta z = \frac{2\pi N e^3 B}{m^2 c\,\omega^2}\Delta z. \qquad (3.68)$$

For a finite slab with variable density $N(z)$ and magnetic flux density $B(z)$, we thus obtain the total rotation of the polarization direction

$$\Delta\psi = \frac{e^3}{2\pi m^2 c}\frac{1}{v^2}\int_0^L B_{\parallel}(z)N(z)\,\mathrm{d}z \quad . \qquad (3.69)$$

In astronomy a system of mixed units is usually employed. Using this system, we have

$$\frac{\Delta\psi}{\mathrm{rad}} = 8.1 \times 10^5 \left(\frac{\lambda}{\mathrm{m}}\right)^2 \int_0^{L/\mathrm{pc}} \left(\frac{B_{\parallel}}{\mathrm{Gauss}}\right)\left(\frac{N_e}{\mathrm{cm}^{-3}}\right)\mathrm{d}\left(\frac{z}{\mathrm{pc}}\right) \quad . \qquad (3.70)$$

The dependence of $\Delta\psi$ on ν^{-2} can be used to determine the value of $\int BN\,dz$ from the measurement of the polarization direction at two frequencies:

$$\frac{RM}{rad\,m^{-2}} = 8.1 \times 10^5 \int\limits_{0}^{L/pc} \left(\frac{B_{\parallel}}{Gauss}\right)\left(\frac{N_e}{cm^{-3}}\right) d\left(\frac{z}{pc}\right)$$
$$= \frac{\left(\frac{\Delta\psi_1}{rad}\right) - \left(\frac{\Delta\psi_2}{rad}\right)}{\left(\frac{\lambda_1}{m}\right)^2 - \left(\frac{\lambda_2}{m}\right)^2} \quad . \tag{3.71}$$

In this expression the unknown intrinsic polarization angle of the source cancels. The units of RM are radians per m^2, and positive RM indicates that $B_{\parallel}$ points toward us. Equation (3.71) can, conversely, be used to determine the intrinsic polarization angle from (3.70) and thus be used to correct the measured polarization. For pulsars, one can combine the values of RM from the Faraday rotation of pulsars and DM, from the pulse dispersion from (2.85). The resulting ratio gives the average magnetic field parallel to the line-of-sight

$$\frac{\bar{B}_{\parallel}}{Gauss} = 1.23 \times 10^{-6} \frac{RM}{DM} \quad . \tag{3.72}$$

If there are line-of-sight reversals, $B_{\parallel}$ is a lower limit to the actual value. Results for pulsars at distances less than 3 kpc show a scatter, but in the galactic longitude range $0°$ to $180°$, the direction of $\bar{B}_{\parallel}$ is away from the Sun, and at longitudes $180°$ to $360°$, towards the Sun. This is in the sense of galactic rotation. The $\bar{B}_{\parallel}$ fields obtained from pulsar studies are in the range of 0.3 µGauss to 3 µGauss. Faraday rotation measurements in our galaxy can be affected by field reversals. This is especially the case for the inner parts of our galaxy, where reversals in B field direction are thought to be present.

Problems

1. A source is 100% linearly polarized in the north–south direction. Express this in terms of Stokes parameters.

2. If the degree of polarization is 10% in Eq. (3.55) with $S_3=0$, $S_1=S_2$ in Eq. (3.53), what is the state of polarization?

3. Intense spectral line emission at 18 cm wavelength is caused by maser action of the OH molecule. At certain frequencies, such emission shows nearly 100% circular polarization, but little or no linear polarization. Express this in terms of Stokes parameters.

4. Determine the *upper limit* of the angle through which a linearly polarized elec-
tromagnetic wave is rotated when it traverses the ionosphere. **(a)** Find RM using
(Eq. (3.73)) with the following parameters: an ionospheric depth of 20 km, an av-
erage electron density of 10^5 cm^{-3} and a magnetic field strength (assumed to be
parallel to the direction of wave propagation) of 1 G.
(b) Carry out the calculation for the Faraday rotation, $\Delta\psi$, for frequencies of
100 MHz, 1 GHz and 10 GHz, if the rotation is $\Delta\psi/\mathrm{rad} = (\lambda/\mathrm{m})^2$ RM.
(c) What is the effect if the magnetic field direction is perpendicular to the direction
of propagation? What is the effect on circularly polarized electromagnetic waves?
(d) Repeat for the conditions which hold in the solar system: the average charged
particle density in the solar system is 5 cm^{-3}, the magnetic field 5 μG and the av-
erage path 10 AU (=1.46×10^{14} cm). What is the maximum amount of Faraday
rotation of an electromagnetic wave of frequency 100 MHz, 1 GHz? Must radio as-
tronomical results be corrected for this?

5. A 100% linearly polarized interstellar source is 3 kpc away. The average electron
density in the direction of this source is 0.03 cm^{-3}. The magnetic field along the
line-of-sight direction, $B_\parallel$, is 3 μG. What is the change in the angle of polarization
at 100 MHz, at 1 GHz?

6. A right hand circularly polarized electromagnetic wave is sent perpendicular to a
perfectly conducting metallic flat surface. The electromagnetic energy must be zero
inside this conductor.
(a) Use a qualitative argument to show that the sense of the polarization of the
reflected wave is opposite to that of the incoming wave.
(b) What is the effect of reflection on a linearly polarized signal?

7. If the DM for a given pulsar is 50, and the value of RM is 1.2×10^2, what is the
value of the *average* line-of-sight magnetic field? If the magnetic field perpendicular
to the line of sight has the same strength, what is the total magnetic field

8. Consider a quasi-monochromatic wave with $\Delta v/\bar{v} = 0.1$ and $v = v_0$, a constant.
Use (3.42) with a(v)=a_0, a constant, and $\phi(\bar{v} + \mu) = \phi_0$ likewise a constant. With
these values, calculate A(t). This is an idealization, however is a commonly used
approximation to describe wide band signals limited by narrow filters.

9. Repeat problem 8 for the function

$$a(v) = a_0 \, e\left(\frac{(v - v_0)^2}{\Delta v^2}\right).$$

Show that $\Delta v \, \Delta t$=1.

Chapter 4
Signal Processing and Receivers: Theory

In this chapter, we cover some general topics concerned with signal processing and noise analysis (Sects. 4.1 and 4.2). These are needed to understand the general properties of radiometers. It is not expected that these topics will change greatly with time. Specifics of actual receivers will be presented in the next chapter. It is essential to have a working knowledge of Fourier transforms in order to make use of the concepts presented in Chaps. 4–8. We give a summary of the relevant concepts of Fourier transforms (FT) in Appendix B, including convolutions and related topics.

4.1 Signal Processing and Stationary Stochastic Processes

The concept of spectral power density was introduced in Chap. 1 in a purely phenomenological way. Radio receivers are devices that measure spectral power density. A detailed understanding of the principles governing the operation of certain receivers, such as autocorrelation spectrometers, as well as the discussion of the limiting receiver sensitivity is possible only if this concept is discussed more thoroughly.

In the preceding chapters, the signals considered were periodic functions of the time which could be conveniently expressed as the superposition of simple harmonic functions of time. It is now necessary to consider a more general class of time variable functions; that is, those allowing representations of signals as *stationary random processes*, $x(t)$. The signal $x(t)$ is a function of time t, but it is not fully determined. One can only specify certain statistical properties of the signal.

4.1.1 Probability Density, Expectation Values and Ergodicity

Perhaps the most important of these statistical quantities is the probability density function, $p(x)$, which gives the probability that at any arbitrary moment of time the value of the process $x(t)$ falls within an interval $(x - 1/2\,dx,\ x + 1/2\,dx)$. For a stationary random process, $p(x)$ will be independent of the time t.

The *expected value* $E\{x\}$ or *mean value* of the random variable x is given by the integral

$$E\{x\} = \int_{-\infty}^{\infty} x\, p(x)\, \mathrm{d}x$$

(4.1)

and, by analogy, the expectation value $E\{f(x)\}$ of a function $f(x)$ is given by

$$E\{f(x)\} = \int_{-\infty}^{\infty} f(x)\, p(x)\, \mathrm{d}x \qquad .$$

(4.2)

This is different from the expected value of the transformation $y = f(x)$

$$E\{y\} = \int_{-\infty}^{\infty} y\, p_y(y)\, \mathrm{d}y = \int_{-\infty}^{\infty} f(x)\, p_x(x)\, \frac{\mathrm{d}x}{|f'(x)|} \, .$$

(4.3)

Frequently encountered expected values are the *mean value*

$$\mu = E\{x\}$$

(4.4)

and the *variance* or dispersion

$$\sigma^2 = E\{x^2\} - E^2\{x\} \, .$$

(4.5)

Another average that can be formed for a stationary random process is the *time average* of the values of the function f. This average will be designated (as in earlier chapters) by acute brackets:

$$\langle f(x) \rangle = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f(x(t))\, \mathrm{d}t \qquad .$$

(4.6)

There are stochastic signals for which this limit does not exist. However, conditions can be formulated [the ergodic theorem of Birkoff, see Khinchin (1949)] so that the results of the definitions (4.2) and (4.6) agree. We will assume this to be the case in the following.

4.1.2 Autocorrelation and Power Spectrum

The concept of Fourier Transforms plays a fundamental role in many branches of physics and engineering, and it is convenient to use this in the discussion of noise

signals. However, there are difficulties in doing this because a stationary time series does not decrease to zero for $t \to \pm\infty$. Therefore the simple definition for the FT

$$X(\nu) = \lim_{T \to \infty} \int_{-1/2T}^{1/2T} x(t) \, e^{-2\pi i \nu t} \, dt \tag{4.7}$$

does not exist; the integral varies irregularly as T increases. As first shown by N. Wiener, the concept of the Cesaro sum of an improper integral can be used to advantage in this situation. The Cesaro sum is defined as

$$\int_{-\infty}^{\infty} A(x) \, dx = \lim_{N \to \infty} \frac{1}{N} \int_0^N \left[\int_{-r}^r A(x) \, dx \right] dr, \tag{4.8}$$

that is, as the limit of the average over the finite integrals. This limit will exist for a wide class of functions where the ordinary improper integral does not exist. For those cases where the ordinary limit exists, this will equal the Cesaro sum, as can be seen if the sequence of the integrations in (4.8) is interchanged using Dirichlet's theorem on repeated integrations (see Whittaker and Watson, Sect. 4.3):

$$\frac{1}{N} \int_0^N \left[\int_{-r}^r A(x) \, dx \right] dr = \int_{-N}^N \left(1 - \frac{|x|}{N} \right) A(x) \, dx. \tag{4.9}$$

For any finite section of a stochastic time series we can define the Fourier transform

$$X_T(\nu) = \int_{-1/2T}^{1/2T} x(t) \, e^{-2\pi i \nu t} \, dt \,.$$

The mean-squared expected value is

$$E_T \left\{ |X(\nu)|^2 \right\} = E \left\{ \int_{-1/2T}^{1/2T} \int_{-1/2T}^{1/2T} x(s) x(t) \, e^{-2\pi i \nu (t-s)} \, ds \, dt \right\}. \tag{4.10}$$

Because $x(t)$ is assumed to be stationary, we must have

$$R_T(\tau) = E_T \left\{ x(s) x(s+\tau) \right\} = E_T \left\{ x(t-\tau) x(t) \right\} \tag{4.11}$$

where $R_T(\tau)$ is the autocorrelation function (ACF). Introducing the ACF into the above expression and performing the integration with respect to s, we find

$$E_T\left\{|X(v)|^2\right\} = T\int_{-T}^{T}\left(1 - \frac{|\tau|}{T}\right)R_T(\tau)\,e^{-2\pi i v \tau}\,d\tau. \tag{4.12}$$

But the right-hand side is a Cesaro sum, and therefore by defining the power spectral density (PSD), $S(v)$, as

$$S(v) = \lim_{T\to\infty}\frac{1}{T}E_T\left\{|X(v)|^2\right\}, \tag{4.13}$$

we obtain from (4.12)

$$\boxed{S(v) = \int_{-\infty}^{\infty} R(\tau)\,e^{-2\pi i v \tau}\,d\tau} \quad . \tag{4.14}$$

This is the *Wiener-Khinchin theorem* stating that the ACF, $R(\tau)$, and the PSD, $S(v)$, of an ergodic random process are FT pairs (see a graphical representation in Fig. 4.1). Taking the inverse FT of (4.14) we obtain

$$\boxed{R(\tau) = \int_{-\infty}^{\infty} S(v)\,e^{2\pi i v \tau}\,dv} \quad . \tag{4.15}$$
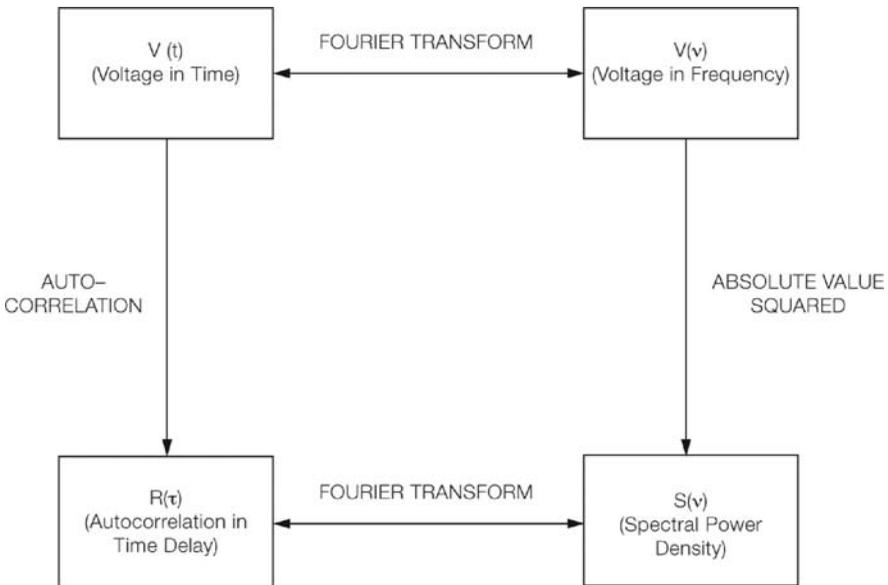
Fig. 4.1 A sketch of the relation between the voltage input as a function of time, $V(t)$, and frequency, $V(v)$, with the autocorrelation function, ACF, $R(\tau)$, and corresponding power spectral density, PSD, $S(v)$. The two-headed arrows represent reversible processes

Thus the total power transmitted by the process is given by

$$R(0) = \int_{-\infty}^{\infty} S(v)\,dv = E\left\{x^2(t)\right\}. \tag{4.16}$$

The limit $T \to \infty$ of the autocorrelation function (ACF) $R_T(\tau)$ can be found using the Cesaro sum resulting in

$$R(\tau) = E\left\{x(s)x(s+\tau)\right\} = \lim_{T\to\infty} \int_{-T}^{T} \left(1 - \frac{|s|}{2T}\right) x(s)x(s+\tau)\,ds. \tag{4.17}$$

Using the concept of ensemble average, this can also be written as

$$R(\tau) = \int\!\!\int_{-\infty}^{\infty} x_1(s)x_2(s+\tau)\,p(x_1,x_2;\tau)\,dx_1\,dx_2 \tag{4.18}$$

where $p(x_1,x_2;\tau)$ is the joint probability density function for the appearance of values x_1 and x_2 which are separated by the time τ. For ergodic stationary processes, (4.17) and (4.18) lead to identical results, but sometimes one or the other is easier to apply.

Applications of these concepts will be illustrated in the following two sections in which we illustrate the influence that linear systems and square-law detectors have on a random process. The results for square law detectors will be used later in the discussion of the limiting sensitivity of radio receivers. A schematic representation of these concepts is given in Fig. 4.2.

4.1.3 Linear Systems

Let the signal $x(t)$ be passed through a fixed linear filter whose time response to a unit impulse $\delta(t)$ is $h(t)$, see (Fig. 4.2). The output of this system is the convolution of $x(t)$ with $h(t)$ that is,

$$y(t) = \int_{-\infty}^{\infty} x(t-\tau)h(\tau)\,d\tau = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)\,d\tau. \tag{4.19}$$

Fig. 4.2 A schematic diagram to illustrate the analysis of noise in a linear system. The symbols above represent the time behavior, those below the frequency behavior
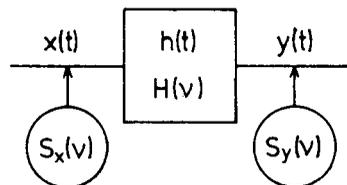
In physical systems the impulse response $h(t) = 0$ for $t < 0$. This permits a corresponding change of the integration limits in (4.19). However, in the following it will not be necessary to make this assumption.

The FT of the filter response is

$$H(v) = \int_{-\infty}^{\infty} h(t)\,\mathrm{e}^{-2\pi\mathrm{i}vt}\,\mathrm{d}t\,. \qquad (4.20)$$

Taking the expectation value of (4.19) and exchanging the order of expectation value and integration we find

$$E\left\{y(t)\right\} = \int_{-\infty}^{\infty} E\left\{x(t-\tau)\right\}h(\tau)\,\mathrm{d}\tau \qquad (4.21)$$

or using (4.4)

$$E\left\{y(t)\right\} = \mu_y = E\left\{x(t)\right\}\int_{-\infty}^{\infty} h(\tau)\,\mathrm{d}\tau \qquad \quad . \qquad (4.22)$$

With (4.20) this can be written as

$$\mu_y = H(0)\,\mu_x \qquad (4.23)$$

showing how the mean value of a stochastic process will be affected if passed through a linear system. If the mean value of the input signal is zero, this will also be true for the output signal.

The autocorrelation $R_{yy}(\tau)$ of the output $y(t)$ is most easily determined by first considering the cross-correlation $R_{xy}(\tau)$ between $x(t)$ and $y(t)$. Multiplying both sides of (4.19) by $x(t-\vartheta)$ we have

$$y(t)x(t-\vartheta) = \int_{-\infty}^{\infty} x(t-\tau)x(t-\vartheta)h(\tau)\,\mathrm{d}\tau\,. \qquad (4.24)$$

But

$$E\{x(t-\tau)x(t-\vartheta)\} = R_{xx}((t-\tau)-(t-\vartheta)) = R_{xx}(\vartheta-\tau)\,.$$

Taking the expectation value of both sides of (4.24) and again exchanging integration and expectation value, we get

$$E\{y(t)x(t-\vartheta)\} = \int_{-\infty}^{\infty} R_{xx}(\vartheta-\tau)h(\tau)\,\mathrm{d}\tau\,.$$

This integral is obviously time independent and equal to the convolution of $R_{xx}(\tau)$ with $h(\tau)$; the left side is the cross-correlation of $y(t)$ and $x(t)$, so that

$$R_{yx}(\tau) = R_{xx}(\tau) \otimes h(\tau),\tag{4.25}$$

where $\otimes$ indicates convolution. Multiplying (4.19) by $y(t + \vartheta)$ we have

$$y(t + \vartheta)y(t) = \int_{-\infty}^{\infty} y(t + \vartheta)x(t - \tau)h(\tau)\,d\tau$$

and

$$R_{yy}(\vartheta) = \int_{-\infty}^{\infty} R_{yx}(\vartheta + \tau)h(\tau)\,d\tau = R_{yx}(\vartheta) \otimes h(-\vartheta).\tag{4.26}$$

From the definition of ACF,

$$R_{xy}(\tau) = R_{yx}(\tau),$$

if we combine (4.25) and (4.26), we obtain

$$\boxed{R_{yy}(\tau) = R_{xx}(\tau) \otimes h(\tau) \otimes h(-\tau)}\ .\tag{4.27}$$

written in full, this is

$$R_{yy}(\tau) = \int_{-\infty}^{\infty} R_{xx}(\tau - t)\left[\int_{-\infty}^{\infty} h(\vartheta + t)h(\vartheta)\,d\vartheta\right]\,dt\ .$$

Therefore, in order to compute a single value of the output autocorrelation function (ACF) of a linear filter, the entire input ACF must be known.

If we take the FT of (4.27) we obtain the following relation for the input and output power spectral densities

$$\boxed{S_y(\nu) = S_x(\nu)\,|H(\nu)|^2}\ .\tag{4.28}$$

4.1.4 Filters

Filters are devices that limit the frequencies passed through a system or change the phase of an input. Filters can be grouped in a number of categories. The most commonly encountered are the following:

1) A *band pass filter* allows a range of frequencies, $\nu_{\min} < \nu < \nu_{\max}$ to pass further in the system

2) A *low pass filter* allows a range of frequencies up to but not beyond a specified frequency, $v < v_{max}$ to pass further in the system.
3) A *high pass filter* allows a range of frequencies, $v > v_{max}$ to pass further in the system.
4) A *band stop filter* eliminates a range of frequencies, $v_{min} < v < v_{max}$ from the system.
5) An *all pass filter* allows all of the input frequencies to pass further, but changes the phase of the input signal.

The properties of such filters are more easily appreciated in plots of their frequency behavior rather than time behavior. Such filters may be either analog or digital. One example relevant for the following is a low pass filter. Low pass filters allow frequencies $0 < v < v_{max}$ to pass unchanged to the digitization and sampling stage. This is usually referred to as a *Video* band. In practical systems, the D. C. term is not passed in order to avoid large offsets.

4.1.5 Digitization and Sampling

The essential part of any digital system is the device that produces a digital output from the analog input. Functionally, the operation of such devices can be divided into two parts:

1) *Analog-to-Digital converters* (A/D converters) and
2) *Samplers*.

In both cases, the input is usually in the video band, that is, from very close to zero frequency to a maximum frequency, which we call *B*.

First, the signal is digitized in an A/D converter. The quality of an A/D converter depends on the speed at which it operates (in either MHz or GHz) and the accuracy used to determine the amplitude of the result (the *quantization* usually expressed in bits). Commercial A/D converters typically have quantizations of 8–12 bits but can accomodate only relatively narrow input bandwidths. In Fig 4.3, we show a one bit (2 level) and two bit (4 level) quantization of an analog function. The one bit quantization of the input results in a positive or negative output level. This is referred to as "hard clipping"; this will result in a lower signal-to-noise ratio since only part of the information contained in the input is retained. Remarkably the properties of the input can be recovered, albeit with a lower S/N ratio. The mathematical details (first derived by Van Vleck) of the recovery from the input from hard clipped data are given in Appendix C. Clearly multi-level quantization of an input will preserve more information, and will thus result in an improved signal-to-noise ratio. An improved but still simple scheme uses a 3 level (sometimes called *1.5 bit*) digitization. This scheme allows a differentiation between amplitudes that are very positive, very negative, positive or negative but close to zero. The limits chosen for 4 level (*2 bit*) digitization are: (1) larger than $+1\sigma$, (2) between $+1\sigma$ and 0, (3) between 0 and -1σ, and (4) lower than -1σ. For a multi-level output, the reconstruction of the

Fig. 4.3 A sketch to illustrate the digitization in an analog-to-digital converter (A/D converter). In (**a**) we show the analog input. In (**b**) is shown the one bit digitization of the input. In (**c**) is shown the 2 bit digitization
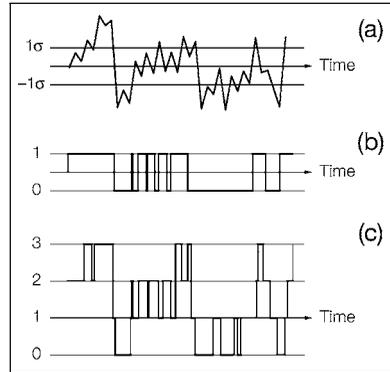
input is usually based on tables generated from computer simulations. In all practical versions of such devices, there is a long term (i.e. a few seconds) average of the input that is used to compensate the input so that the average does not drift far from a given value, usually zero. A recent application in radio astronomy is the A/D converter designed by Recoquillon et al. (2005). For a video input, the low frequency side of the input band is usually not well determined and DC offsets become important. The ALMA (Atacama Large Millimeter Array) design operates between 2 and 4 GHz for this reason. The ALMA the A/D converter operates at a rate of 4 Gigasamples per second.

Second, the digitized function must be *sampled* at regular intervals. The sampling of a sinusoid is shown in Fig. 4.4. In this example, the maximum frequency of the input is v_0. Samples of this function are shown for rates of $2v_0$ and $4v_0$. Given an input from 0 Hz to B Hz, the sampling rate, v_0, must be $v_0 = 2B$ to characterize the sinusoid, that is, at twice the highest frequency to be analyzed. This is referred to as the *Nyquist Sampling Rate*. Clearly a higher sampling rate can only improve the characterization of the input. The sampling functions must occupy an extremely small time interval. A higher sampling rate will allow the input to be better characterized, thus giving a better S/N ratio.

In both the digitization and sampling we have assumed that the reaction of the devices and that the sampling interval is shorter than any changes of the input.

An example of the sampling process in time and frequency is shown in Fig. 4.5. The time variation of an analog function in panel (a) determines the maximum range of frequencies in panel (b). Note that negative frequencies are also plotted to allow a determination of phase of the input. The process of sampling in the time domain is a multiplication of the function in panel (a) with the sampling function in (c). In the frequency domain this is a convolution (see Fig. 4.1). In the frequency domain, it is clear that a minimum sampling rate is needed to prevent an overlap of the sampled function in frequency. If an overlap, there will be a mixture of frequency compontents. This effect, *aliasing*, usually causes a degration of the sampled signal.

If only a portion of the input function is retained in the quantization and sampling process, information is lost. This results in a lowering of the signal-to-noise (S/N)

Fig. 4.4 An illustration of
the *Nyquist Sampling Rate*. In
panel (**a**) the sine wave input.
In panel (**b**) two samples per
period, at the best possible
position. In panel (**c**), four
samples per period. If the
sampling rate is $2\nu_0$, the
properties of the sinusoid
can be characterized. If the
sampling rate is higher the
characterization will be even
better
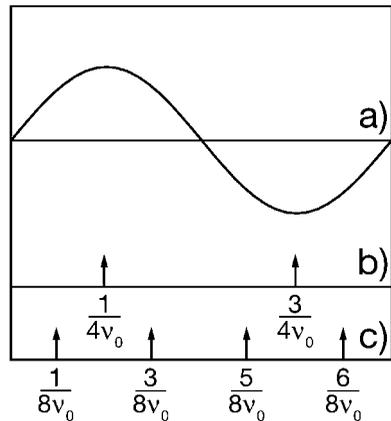
ratio. The effects of sampling rate and quantization on the S/N ratio are quantified
in Table 4.1. The effect on the signal-to-noise ratios for 1, 2 and 3 bit quantizations
at the Nyquist sampling rate (where $1/2\Delta\nu$) are listed in Col. 2. In addition, the im-
provement of the S/N ratios with sampling at *twice* the Nyquist sampling. Sampling
at a rate much higher than the Nyquist rate is referred to as "oversampling". rate are
listed in Col. 3.
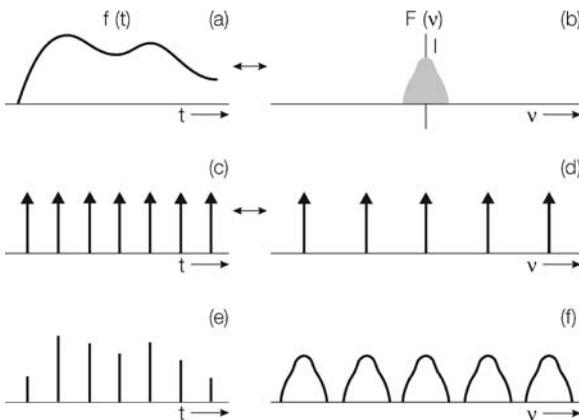
Fig. 4.5 The time and frequency distribution of a sampled function: (**a**) The time variation, (**b**), the
frequency behavior, (**c**) the time behavior of a regularly spaced sampling function (referred to as
a "picket fence" function), (**d**) the frequency behavior of the "picket fence" function, (**e**) the time
behavior of the sampled function, and (**f**) the frequency behavior of the function sampled with a
"picket fence". The result in (**f**) is low pass filtered. The maximum frequency extent in (**b**), ν_{m}, is
smaller than the sampling rate, ν_0 as shown in (**d**). See Appendix B for the "picket fence" function

Table 4.1 Signal-to-noise ratio as function of quantization and sampling rate

(1) No. of bits	(2) Sampling rate	(3)
	$\frac{1}{2\Delta\nu}$	$\frac{1}{4\Delta\nu}$
1	0.64	0.74
2	0.81	0.89
3	0.88	0.94
∞	1.00	1.00

[from: D'Addario (1989)]

4.1.6 Gaussian Random Variables

For the practical analysis of complicated systems the class of stationary random processes is often too large, so one restricts the analysis to functions with less general properties to simplify the investigations. Here stationary *normally distributed random variables* or *Gaussian noise*, for which the probability density distribution function is a Gaussian function with the mean $\mu = 0$, are frequently used. For example, a function with a Gaussian distribution of its values can be used to represent *white noise* which is passed through a band limiting filter. At each instant of time, the probability distribution of a variable x is given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-x^2/2\sigma^2} \,. \tag{4.29}$$

For this random variable we have

$$E\{x\} = \mu = 0 \quad \text{and} \quad E\{x^2\} = \sigma^2 \,.$$

It is particularly important to note that the FT of a Gaussian is also a Gaussian, and that the widths of these FT pairs are inversely related.

$$\Delta\nu \cdot \Delta t = 1 \,.$$

Similar situations encountered in Quantum Mechanics, under the description *Heisenberg Uncertainty Principle*. This represents the fact that certain variables are FT pairs. In Table 4.2 we give a set of values for the area within the positive half of a normalized Gaussian curve in terms of the RMS standard deviation, σ. These values give the probability that a Gaussian distributed quantity lies *above* the average.

4.1.7 Square Law Detectors

In radio receivers, the noise is passed through a device that produces an output signal $y(t)$ which is proportional to the power in a given input $v(t)$:

Table 4.2 Gaussian noise statistics

σ	Value outside the curve	Value inside
1	0.3174	0.6826
2	0.0456	0.9544
3	0.0026	0.9974
4	0.0020	0.9980

$$y(t) = a v^2(t).$$ (4.30)

This involves an evaluation of the integral

$$E\{y(t)\} = E\{a v^2(t)\} = \frac{a}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} v^2 e^{-v^2/2\sigma^2} dv$$

The standard approach used to evaluate this expression is the following. First, take the square of this integral. Then in each of the factors of the square, use the variables x and y. Transform from rectangular (x, y) to two-dimensional polar coordinates (ρ, θ). The result is

$$E\{y(t)\} = E\{v^2(t)\} = a\sigma_v^2$$ (4.31)

For the evaluation of $E\{y^2(t)\}$, one must evaluate

$$E\{y(t)^2\} = E\{v^4(t)\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} v^4 e^{-v^4/2\sigma^2} dv$$

This is best done using an integration by parts, with $u = x^3$, and $dv = xe^{-x^2/2\sigma^2}$. The result of this integration is

$$E\{y^2(t)\} = 3 a^2 \sigma_v^4$$ (4.32)

and hence

$$\sigma_y^2 = E\{y^2(t)\} - E^2\{y(t)\} = 2 E^2\{y(t)\}.$$ (4.33)

Thus, in contrast to linear systems, the mean value of the output signal of a square-law detector does not equal zero, even if the input signal has a zero expected mean value.

4.2 Limiting Receiver Sensitivity

Radio receivers are devices used to measure the PSD. A receiver should contain the following basic units:

1) A reception (usually band pass) filter that defines the spectral range of the receiver.
2) A square-law detector used to produce an output signal that is proportional to the average power in the reception band.
3) A smoothing filter or averager, which determines the time response of the output.

In some cases, processes in (2) and (3) are carried out after digitizing the signal, so the operations could be carried out in a computer. In some cases, a receiver might record the sampled voltages on a storage device for later analysis (see, e.g., Problem 1(c) for example). In other cases, for a fast receiver response, item (3) might be dispensed with.

A receiver must be *sensitive*, that is, be able to detect faint signals in the presence of noise. Just as with any other measuring device there are limits for this sensitivity, since the receiver input and the receiver itself are affected by noise. We will derive the expression for the limiting sensitivity as a function of receiver parameters. Even when no input source is connected to a receiver, there is an output signal, since any receiver generates thermal noise. This noise is amplified together with the signal. Since signal and noise have the same statistical properties, these cannot be distinguished. To analyze the performance of a receiver we will use the model of an ideal receiver producing no internal noise, but connected simultaneously to two noise sources, one for the external source noise and a second for the receiver noise. These form a 2 port network which is characterized by noise power, bandwidth, and gain. The system gain is the *available gain*, $G(v)$. This is the ratio of the output power to the input power. To be useful, receivers must increase the input power level. The power per unit bandwidth entering a receiver can be characterized by a temperature, as given by Eq. $P_v = kT$ (1.42). Furthermore, it is *always* the case that the noise contributions from source, atmosphere, ground and receiver, T_i, are additive,

$$T_{\text{sys}} = \sum T_i$$

We apply these concepts to a 2 port system, as shown in Fig. 4.6. The signal input, S_1, and output S_2 are related by the system gain, G. The noise output, N_2 is the noise input, N_1, multiplied by the gain, plus the noise added by the system, N_{int}. An often-used figure of merit is the *Noise Factor*, F. This is defined as

$$F = \frac{S_1/N_1}{S_2/N_2} = \frac{N_2}{GN_1} = 1 + \frac{T_{\text{R}}}{T_1} \tag{4.34}$$

that is, any additional noise generated in the receiver contributes to N_2. For a direct detection system, $G = 1$. If T_1 is set equal to $T_0 = 290$K, we have

$$T_{\text{R}} = F - 1$$

Given a value of F, one can determine the receiver noise temperature. The section relates receiver properties to the minimum uncertainty in a measurement.
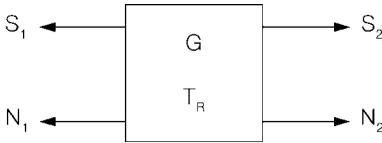
Fig. 4.6 A schematic of a two port system. The receiver is represented as a box, with the signal T_1, and noise T_R, shown on the left. On the right are these quantities after an amplification G. For a direct detection device, $G=1$

4.2.1 Noise Uncertainties due to Random Processes

The following is an exact derivation that makes use of Nyquist sampling of the input.

We assume that the signal is a Gaussian random variable with mean zero which is sampled at a rate equal to twice the bandwidth.

Refer to Fig. 4.7. By assumption $E(v_1) = 0$. The input, v_1 has a much larger bandwidth, B, than the bandwidth of the receiver, that is, $\Delta v \ll B$. The output of the receiver is v_2, with a bandwidth Δv. The power corresponding to the voltage v_2 is $\langle v_2^2 \rangle$.

$$P_2 = v_2^2 = \sigma^2 = kT_{\text{sys}}\, G\Delta v \;, \tag{4.35}$$

where Δv is the receiver bandwidth, G is the gain, and T_{sys} is the total noise from the input T_A and the receiver T_R. The contributions to T_A are the external inputs from the source, ground and atmosphere. Given that the output of the square law detector is v_3

$$\langle v_3 \rangle = \langle v_2^2 \rangle \tag{4.36}$$

then after square-law detection we have

$$\langle v_3 \rangle = \langle v_2^2 \rangle = \sigma^2 = kT_{\text{sys}}G\Delta v \;. \tag{4.37}$$

Crucial to a determination of the noise is the mean value and variance of $\langle v_3 \rangle$ From (4.32) the result is

$$\langle v_3^2 \rangle = \langle v_2^4 \rangle = 3 \langle v_2^2 \rangle \tag{4.38}$$

this is needed to determine $\langle \sigma_3^2 \rangle$. Then,

$$\sigma_3^2 = \langle v_3^2 \rangle - \langle v_3 \rangle^2 \tag{4.39}$$
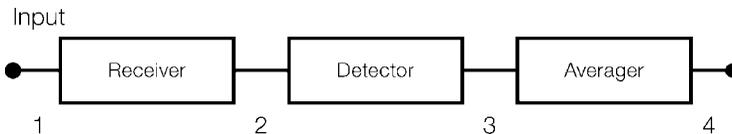
Input

Fig. 4.7 The principal parts of a receiver

$\langle v_3^2 \rangle$ is the total noise power (= receiver plus input signal). Using the Nyquist sampling rate, the averaged output, v_4, is $(1/N)\Sigma v_3$ where $N = 2\Delta v\,\tau$.

From v_4 and $\sigma_4^2 = \sigma_3^2/N$, we obtain the result

$$\sigma_4 = k\Delta v G(T_A + T_R)/\sqrt{\Delta v\,\tau} \tag{4.40}$$

We have explicitly separated T_{sys} into the sum $T_A + T_R$. Finally, we use the calibration procedure to be described in Sect. 4.2.3, to eliminate the term $kG\Delta v$.

$$\boxed{\frac{\Delta T}{T_{sys}} = \frac{1}{\sqrt{\Delta v\,\tau}}} \quad . \tag{4.41}$$

This result is so important that we review the process. We have assumed that the receiver input is a stationary broadband signal. This is referred to as a *white noise* spectrum. The voltage is assumed to follow a Gaussian distribution with zero mean value. Then $E(v_1) = 0$. After passing through the receiver, v_2, is limited to a bandwidth Δv. After passing through the square law detector, one has the result in Eq. (4.37). One must determine the variance of v_3 to find the RMS uncertainty in the receiver output. Using the Nyquist theorem we can describe such a white noise PSD by an equivalent noise temperature [cf. (1.42)] that would produce such a thermal PSD. The calibration process allows us to specify the PSD of the receiver output in degrees Kelvin instead of in Watts per Hz. We therefore characterize the receiver quality by the system noise temperature $T_{sys} = T_A + T_R$. The analysis of another type of detector is presented in Problem 4.2.3.

This result was first obtained by Dicke (1946), using a more complex derivation (given in "Tools" 4th edition). Equation (4.41) is *the* fundamental relation between system noise, bandwidth, integration time and rms fluctuations: For a given system, the improvement in the RMS noise cannot be better than as given in Eq. (4.41). Systematic errors will only increase ΔT, although the time behavior may follow relation (4.41). We repeat for emphasis: T_{sys} is the noise from the *entire* system. That is, it includes the noise from the receiver, atmosphere, ground, and *the source*. Therefore ΔT is larger for an intense source. However this is an ideal situation since the receiver noise is dominated by the signal.

4.2.2   Receiver Stability

Sensitive receivers are designed to achieve a low value for T_{sys}. Since the signals received are of exceedingly low power, receivers must also provide sufficient output power. This requires a large receiver gain. Then even very small gain instabilities can dominate the thermal receiver noise. Therefore receiver stability considerations are also of prime importance. Because the power measured at the receiver output is that generated in the receiver plus the input, $T_{sys} = T_A + T_R$,

$$P = k(T_A + T_R)G\Delta v, \tag{4.42}$$

variations of the total system gain ΔG leading to

$$P + \Delta P = k(T_A + T_R)(G + \Delta G)\Delta v \tag{4.43}$$

are indistinguishable from variations of T_A or T_R

$$P + \Delta P = k(T_A + \Delta T + T_R)G\Delta v. \tag{4.44}$$

Comparing (4.44) and (4.43) using (4.42) we obtain

$$\boxed{\frac{\Delta T_{RMS}}{T_R} = \frac{\Delta G}{G}} \quad , \tag{4.45}$$

This shows that variations of the output power caused by gain variations enter directly into the determination of limiting sensitivity. If a total power receiver is to measure an input of $10^{-4}\,T_R$, the total gain must be kept constant to less than this value. This is exceedingly difficult to achieve with an absolute measurement, so therefore one must employ a receiver system based on a differential or comparison measurement.

This was first applied to radio astronomical receivers by Dicke (1946). This is a straightforward application of the compensation principle such as the Wheatstone bridge. We show a schematic of such a system in Fig. 4.8. In this scheme a receiver is switched periodically between an input T_A and a resistive load at the thermodynamic temperature T_{ref}. If both input, T_A, and reference, T_{ref}, are matched to the receiver input, the antenna gives the output

$$P_A = k(T_A + T_R)G\Delta v$$
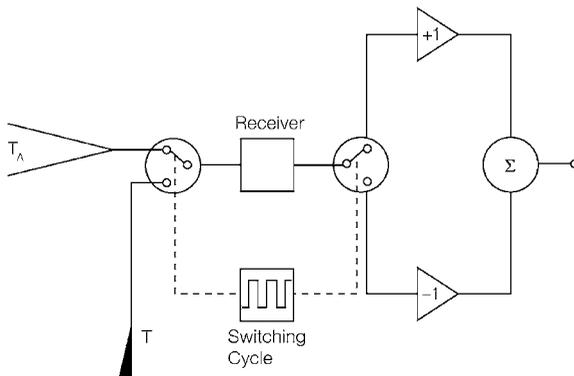
Fig. 4.8 A balanced receiver employing Dicke switching between a load at the temperature T_{ref}, and the sky at T_A. The box joined by *dotted lines* connecting the two switches indicates a square wave switching cycle. This is alternately multiplied by $+1$ or -1, so that the response from the reference is subtracted from sky synchronously in the adding section on the far right

while the reference load produces

$$P_{\text{ref}} = k\left(T_{\text{ref}} + T_{\text{R}}\right)G\Delta\nu.$$

At the output of the receiver, the difference of these two signals as measured by a phase sensitive detector or lock-in amplifier is then

$$P_{\text{A}} - P_{\text{R}} = k\left(T_{\text{A}} - T_{\text{ref}}\right)G\Delta\nu,$$

provided T_{sys} does not change its value between a measurement of signal and reference. If a gain variation ΔG is wrongly interpreted as a variation ΔT of the input we have, eliminating k and $\Delta\nu$,

$$\left(T_{\text{A}} - T_{\text{ref}}\right)\left(G + \Delta G\right) = \left(T_{\text{A}} + \Delta T - T_{\text{ref}}\right)G$$

or

$$\boxed{\frac{\Delta T_{\text{RMS}}}{T_{\text{R}}} = \frac{\Delta G}{G}\frac{T_{\text{A}} - T_{\text{ref}}}{T_{\text{R}}}} \qquad . \tag{4.46}$$

The influence of gain fluctuations depends on the difference $T_{\text{A}} - T_{\text{ref}}$. For a *balanced receiver* with $T_{\text{A}} = T_{\text{ref}}$, ΔT is completely independent of any gain variations. Then the receiver is functioning as a zero point indicator. This is true only when $T_{\text{A}} \cong T_{\text{ref}}$. If $T_{\text{A}} \neq T_{\text{R}}$ the receiver is no longer balanced, then gain variations will influence the signal.

The rate at which the receiver is switched depends on the time behavior of the stability of the receiver. While in the 1950s and early 1960s fast switching rates were needed, present-day receiver systems are so stable that switching rates of a Hertz or slower can be used.

There are different means of producing the comparison T_{ref}. A straightforward implementation is a resistive load at the temperature T_{ref}. For low-noise systems, the reference temperature should not be too different from T_{A}. This might be provided by an absorber immersed in a liquid nitrogen or liquid helium bath.

If the input power levels vary over a wide range, it is not always possible to maintain a well-balanced system with $T_{\text{A}} \cong T_{\text{ref}}$. One could add noise to the load, increasing T_{ref}. Alternatively the system can be stabilized by periodically injecting a constant noise step for part of the measuring cycle. If this calibration cycle is faster than the rate of gain changes, one can compare the output at appropriate phases of the cycle, and with this information correct both the zero point and the gain of the system. In the millimeter and sub-mm wavelength range, the sky temperature can vary. This will have a large effect on T_{A}. The compensation involves a determination of sky conditions at the receiver frequency. This involves a "chopper wheel" calibration to be described in Chap. 8.

At all wavelengths, fluctuations in the atmosphere will affect high resolution images. At millimeter and sub-mm wavelengths, these fluctuations are mostly due to water vapor, but for wavelengths in the range of a meter, the ionospheric fluctuations can distort images. Corrections for such effects are complex, and will be described in Chap. 9.

4.2.2.1 Effect of Switching on Receiver Noise

The time spent measuring references or performing calibrations will *not* contribute to an improvement in the S/N ratio. Thus this amount of time must be subtracted from the total integration time in (4.41). So a system in which one half the integration time is used to measure T_{ref} will achieve a temperature resolution

$$\frac{\Delta T_{\text{RMS}}}{T_{\text{sys}}} = \frac{\sqrt{2}}{\sqrt{\Delta \nu \, \tau}} . \tag{4.47}$$

Often the ΔT quoted for a switched receiver has an additonal factor of $\sqrt{2}$ compared to (4.47). This is caused by the fact that ΔT is computed as the difference $\Delta z = T_{\text{A}} - T_{\text{ref}}$, where both T_{A} and T_{ref} have equal errors due to noise. There is a factor $\sqrt{2}$ from spending only $\sqrt{2}$ of the total time on the source, and an additional $1/2$ caused by subtracting two equally noisy signals. The time τ is the *total* time taken for the measurement (i.e. on-source and off-source).

Even for the output of a total power receiver there will be additional noise in excess of that given by (4.41) since the signals to be differenced are $\Delta T + T_{\text{sys}}$ and T_{sys}. This is needed since $\Delta T \ll T_{\text{sys}}$. The error of this difference signal is given by (4.47).

If the time variation of G is included in the expression for the sensitivity limit, the generalization of (4.41) for stochastic time variations of $\Delta G/G$ will be Eq. (4.48). In Table 4.3 we list the noise performance for different types of receivers. The case of total power and switched receivers have been discussed previously. Correlation receivers are treated in Sect. 5.4.1; these use 2 identical receivers to reduce gain variations, but require combining two noisy inputs, hence an additional factor of $\sqrt{2}$. The additional noise contributions introduced by the use of one and two bit quantization are listed in Table 4.1.

$$\frac{\Delta T}{T_{\text{sys}}} = K \sqrt{\frac{1}{\Delta \nu \, \tau} + \left(\frac{\Delta G}{G}\right)^2} . \tag{4.48}$$

One can model the time behavior of ΔG. For a time dependence

$$\left(\frac{\Delta G}{G}\right)^2 = \gamma_0 + \gamma_1 \tau ,$$

we obtain the smallest value for the resolution $\Delta T / T_{\text{sys}}$ at the integration time

Table 4.3 Noise performance K of different receiver configurations

Receiver type	K	
Total power receiver (4.41)	1	
Switched receiver	$\sqrt{2}$	$\dfrac{\Delta T_{\text{RMS}}}{T_{\text{sys}}} = \dfrac{K}{\sqrt{\Delta \nu \, \tau}}$
Correlation receiver	$\sqrt{2}$	
1-bit digital receiver	2.21	
2-bit digital receiver	1.58	

$$\tau_m = \frac{1}{\sqrt{\Delta\nu\,\gamma_1}}\,.$$ (4.49)

For actual receivers this time can be determined by sampling the normalized output x_i at equal time intervals τ. If

$$S_N = \sum_{i=1}^{N} x_i\,, \quad Q_N = \sum_{i=1}^{N} x_i^2\,,$$

then the mean error of the mean value

$$\bar{x}_N = \frac{1}{N} S_N$$

and the RMS uncertainty is

$$\sigma(N) = \frac{1}{N}\sqrt{Q_N - S_N^2}\,,$$ (4.50)

and this can be computed using running averages of N, S_N, and Q_N without needing individual values of x_i.

A plot of σ^2 versus $T = N\tau$ is sometimes referred to as an Allan plot, after D.W. Allan (1966) who applied this type of analysis to determine the stability of frequency standards. The value of τ_M depends on the stability of the receiver, and the stability of the power entering the receiver. In the millimeter and sub-millimeter wavelength range, the stability of the atmosphere plays a dominant role. The Allan plot is the ultimate way to measure stability, but requires a great amount of measurement time. Therefore it is often used to test receivers in a laboratory, but only rarely on telescopes.

Before reaching a time τ_m, one must take a comparison measurement to prevent an increase in ΔT_{RMS}. This may involve directing the receiver to another part of the sky, or connecting the receiver to an internal source, or changing frequency. A plot of the behaviour of τ_m is shown in Fig. 4.9

4.2.3 Receiver Calibration

In the calibration process, a noise power scale must be established at the receiver input. While the detailed procedures depend on the actual instruments in use, the basic principles are following. In radio astronomy the noise power of coherent receivers (those which preserve the phase of the input) is usually measured in terms of the noise temperature. To calibrate a receiver, one relates the noise temperature increment ΔT at the receiver input to a given measured receiver output increment Δz (this applies to heterodyne receivers. For the calibration of bolometer receivers, see Sect. 8.2.6). In principle, the receiver noise temperature, T_R, could be computed from the output signal z provided the detector characteristics are known. In practice

Fig. 4.9 The time depen-
dence of the relative receiver
uncertainty, $\Delta T_{\mathrm{RMS}}/T_{\mathrm{sys}}$ nor-
malized to unit bandwidth
$\Delta \nu$. K is a factor accounting
for data taking procedures
(Table 4.3). For a total power
system, $(K = 1)$. The differ-
ent one-parameter curves are
characterized by the value τ_m
in (4.49). The turn off in each
curve gives the integration
time leading to the smallest
value for $\Delta T_{\mathrm{RMS}}/T_{\mathrm{sys}}$
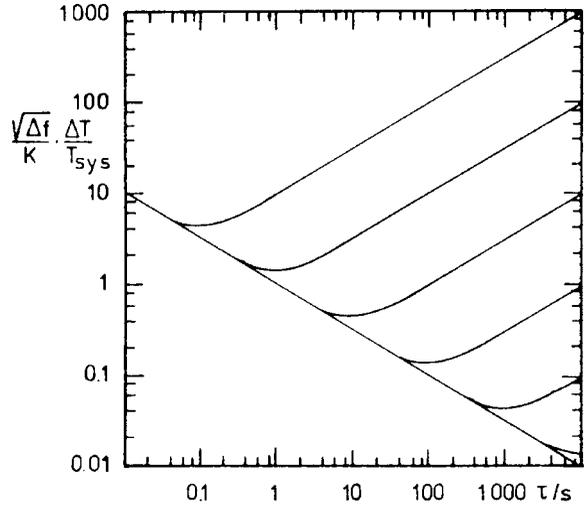
the receiver is calibrated by connecting two or more known power sources to the
input. Usually matched resistive loads at the known (thermodynamic) temperatures
T_{L} and T_{H} are used. The receiver outputs are then

$$z_{\mathrm{L}} = (T_{\mathrm{L}} + T_{\mathrm{R}})\,G,$$
$$z_{\mathrm{H}} = (T_{\mathrm{H}} + T_{\mathrm{R}})\,G,$$

from which

$$T_{\mathrm{rx}} = \frac{T_{\mathrm{H}} - T_{\mathrm{L}}\,y}{y - 1} \qquad , \qquad (4.51)$$

where

$$y = z_{\mathrm{H}}/z_{\mathrm{L}}. \qquad (4.52)$$

The noise temperatures T_{H} and T_{L} are usually produced by matched resistive
loads (absorbers in the millimeter/sub-millimeter wavelength ranges) at the ambient
temperature ($T_{\mathrm{H}} \cong 293\,\mathrm{K}$ or 20°C) and at the temperature of liquid nitrogen ($T_{\mathrm{L}} \cong$
78 K or -195°C) or sometimes liquid helium, which has a boiling point $T_{\mathrm{L}} \cong 4.2\,\mathrm{K}$.
In this process, the receiver is assumed to be a linear power measuring device (i.e. we
assume that the non-linearity of the receiver is a small quantity). Usually such a
fundamental calibration of the receiver need be done infrequently. At centimeter
wavelengths, secondary standards are used. In the millimeter/sub-mm wavelength
range, measurements of the emission from the atmosphere and then from an ambient
resistive load are combined with models to provide an estimate of the atmospheric
transmission. For a determination of the receiver noise, an additional measurement,
usually with a cooled resistive load is needed. Note that the **y** factor as presented here
is determined in the Rayleigh-Jeans limit, and thus using the concepts of classical
physics.

Problems

1. The Gaussian probability distribution function with mean m is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} .$$

(a) Show that $\int_{-\infty}^{+\infty} p(x)dx = 1$. If the first moment, or mean value m, is

$$m = \langle x \rangle = \int_{-\infty}^{+\infty} xp(x)dx$$

and the second moment is

$$\langle x^2 \rangle = \int_{-\infty}^{+\infty} x^2 p(x)dx ,$$

(b) find m and σ, the RMS standard deviation, where $\sigma = \langle x^2 \rangle - \langle x \rangle^2$. The third and fourth moments are defined in analogy with the definitions above. Determine the third and fourth moments of the Gaussian distribution.

(c) The relation between $\langle x^2 \rangle$ and $\langle x^4 \rangle$ has been used to study the noise statistics for very intense narrow band emission from an astronomical source at 18 cm (see Evans et al. 1972 Phys. Rev. **A6**, 1643). If the noise input has zero mean, and if the voltages $\langle v^2 \rangle$ and $\langle v^4 \rangle$ are compared, what would you expect the relation to be for a Gaussian distribution of noise?

2. For an input

$$v(t) = A\sin 2\pi vt$$

calculate the FT, autocorrelation function and power spectrum. Note that this function extends to negative times. Repeat the calculation for

$$v(t) = A\cos 2\pi vt .$$

3. Calculate the power spectrum, S_v, for the sampling function $v(t) = A$ for $-\tau/2 < t < \tau/2$, otherwise $v(t) = 0$, by taking the Fourier transform to obtain $V(v)$ and then squaring this. Next, form the autocorrelation of this function, and then FT to determine the power spectrum. Show that these two methods are equivalent.

4. Repeat the analysis in Problem 5, but shifting this function by a time $+\tau/2$: that is, $v(t) = A$ for $0 < t < \tau$, otherwise $v(t) = 0$. The FT shift theorem is given by equation (B5) in Appendix B

$$f(x-a) \leftrightarrow e^{-i2\pi as} F(s) .$$

Show that the result of this problem can be obtained from the result of Problem 5 by applying the shift theorem. What is the value of the shift constant, a?

5. Repeat the above for the function $v(t) = A$ for $\tau < t < 2\tau$, and $-2\tau < t < -\tau$, otherwise $v(t) = 0$. The result can be interpreted as the frequency distribution calculated in Problem 5, modulated by $\cos 2\pi v \tau$. This is an example of the modulation property of Fourier transforms, as in equation (B6) in Appendix B, namely,

$$f(x)\cos x = \frac{1}{2}F(s-v) + \frac{1}{2}F(s+v).$$

6. Consider another aspect of the situation described in the last problem. We have a function $\cos(2\pi v_c t)\cos(2\pi v_s t)$, where $v_s = v_c + \Delta$, where $\Delta \ll v_c$. Apply the identity $\cos(x+y) = (1/2)[\cos(x+y) + \cos(x-y)]$. Check whether the modulation property of the Fourier transform applies.

7. A table of Gaussian integrals to determine the area within the boundary of the curve at the σ, 2σ, 3σ and 4σ levels is given in Table 4.2.
(a) If you want to determine whether a feature is *not* noise at the 1% level, how many standard deviations from the mean must this signal be?
(b) Suppose you want to detect a continuum signal of peak temperature 10^{-2} K with a total power receiver with a system noise of 200 K, and a bandwidth of 500 MHz. Assume that this system is perfectly stable, that is random noise is the only source of error. How long must you integrate to obtain a 3σ detection?
(c) For an emission line with a total width of 10 kHz, use the same system, but using a spectrometer which has a bandwidth equal to the linewidth. How long must one integrate so that a detection is 99% certain if random noise is the only effect?
(d) If the spectrometer has 1000 channels, how many "false" emission lines, i.e. noise peaks, will be found at the 1σ, 2σ, 3σ levels?
(e) Now suppose the signal could appear as either a positive or negative deflection. How does this change the probabilities?

8. (a) On two days, labelled as 1 and 2, you have taken data which are represented by Gaussian statistics. The mean values are x_1 and x_2, with σ_1 and σ_2. Assume that the average is given by $\bar{x} = fx_1 + (1-f)x_2$ and the corresponding $\sigma^2 = f^2\sigma_1^2 + (1-f)^2\sigma_2^2$. Determine the value of f which gives the smallest $\bar{\sigma}$ by differentiating the relation for σ and setting the result equal to zero. Show that

$$\bar{x} = \left(\frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)x_1 + \left(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)x_2$$

and

$$\bar{\sigma}^2 = \left(\frac{\sigma_2^4}{(\sigma_1^2 + \sigma_2^2)^2}\right)\sigma_1^2 + \left(\frac{\sigma_1^4}{(\sigma_1^2 + \sigma_2^2)^2}\right)\sigma_2^2.$$

(b) Use the relation $\sigma^2 \sim 1/(\text{time})$ to show that the expression for $\bar{x}$ reduces to the result, $\bar{x} = (1/(t_1 + t_2))(t_1 x_1 + t_2 x_2)$.

9. Obtain (4.34) from the quantities in Fig. 4.6. Justify the definition of the noise factor F in Eq. (4.34) based on the case of a *noiseless* receiver, i.e. one with

$F = 1$. Show that this definition is consistent with the definition of receiver noise temperature

$$T_R = (F - 1) \cdot 290\,\text{K}$$

if a room-temperature load is connected to the receiver input. Suppose $F = 2$, what is T_R? Repeat for $F = 1.2$ and 1.5.

10. Use the analysis in Sect. 4.2.1, step for step, for a *linear* detector. In this device, the output is taken to be the absolute value of the voltage input. Assume that the signal is small compared to the receiver noise. Complete each calculation as in the previous problem. The output of the linear detector is

$$v_3 = \int |v_2| \exp(-v_2^2/2\sigma_2^2)\,dx,$$

while the noise depends on $\langle v_3 \rangle^2 = \langle v_2 \rangle^2 = \sigma^2$.
To obtain the final result, one must make use of the relation

$$\Delta T_{\text{RMS}} = \frac{\sigma_4}{(\Delta \langle v_4 \rangle / \Delta T_s)}.$$

11. The y factor is used to determine receiver noise. Given that T_L is 77 K and $T_H = 290\,\text{K}$, show that the plot in Fig. 4.10 correctly expresses the relation between T_{rx} and the y factor.

Fig. 4.10 A plot of receiver noise versus **y** factor
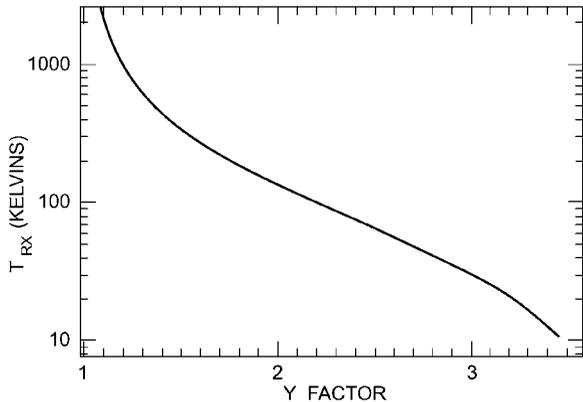
12. Suppose a receiver accepts inputs from 2 frequencies, v_u and v_l. The response of the receiver is the same at these frequencies.
(a). If all factors are equal, and the signal is present in both v_u and v_l, how does the value of T_R change in Eq. 4.41? **(b).** Suppose the signal is present in v_u only. Repeat part (a). **(c).** Repeat (b) for the situation in which the response at v_u is twice that at v_l. What is the value of T_R?

13. Derive the result in Eq. (4.49).

14. To detect a source one samples a large region of the sky. The receiver is perfectly stable. If one has 10 samples at the position of the source, and 10^3 samples away from the source. One can fit a curve to the off-source data and subtract this from the on-source data. Justify the assertion the if the RMS noise of the on-source data is N, the noise in the difference of on-source and off-source is $N\sqrt{1+0.01}$.

Chapter 5
Practical Receiver Systems

5.1 Historical Introduction

The first receivers used by Jansky and Reber were *coherent radiometers*. These radiometers preserve the phase of the received wave field and are sensitive to a single polarization. Usually coherent radiometers are *superheterodyne* systems. In such systems the frequency of the input is translated to another (usually lower) frequency before further processing. Usually this processing consists of amplification in the Intermediate Frequency (IF) section and detection. Such receivers allow greater flexibility in the analysis of the radiation, but involve a number of individual components. One can divide coherent receivers into *front ends* and *back ends*. The dividing point is somewhat arbitrary. Usually the front end operates at the sky frequency, while the back end operates at lower frequencies.

Front ends consist of amplifiers that operate at the sky frequency and/or mixers, which are frequency converters. The trend has been to improve the sensitivity of front ends while extending operation to higher frequency. Initially front ends consisted of room temperature mixers. Later these were replace by exotic devices such as uncooled and then cooled parametric amplifiers, maser amplifiers, cooled transistor amplifiers, and, at millimeter and sub-mm wavelengths, superconducting mixers.

The back ends are devices that analyze the polarization, time structure or spectral properties of the broadband radiation. The trend has been toward digital components for all types of backends. Frequently these components are developed for use in commercial electronics, but have been successfully adapted for radio astronomical applications. Because both phase and amplitude are preserved, only coherent radiometers are used in radio interferometers.

Incoherent radiometers do not preserve phase; these operate as direct detection systems. The most common type of incoherent radiometer at millimeter wavelengths is a bolometer. Bolometers are basically very sensitive thermometers. These have wide bandwidths and high sensitivities. Bolometers are sensitive to both polarizations. For single telescope continuum measurements in the millimeter and sub-mm ranges, semiconductor bolometers have dominated the field. These all follow the practical design pioneered by F. J. Low.

5.1.1 Bolometer Radiometers

The operation of a bolometer makes use of the effect that the resistance, R, of a material varies with the temperature. When radiation is absorbed by the bolometer material, the temperature varies; this temperature change is a measure of the intensity of the incident radiation. Because this thermal effect is rather independent of the frequency of the radiation absorbed, bolometers are intrinsically broadband devices. The frequency discrimination needed is be provided by external filters. A bias voltage must be applied to a bolometer for optimum performance. Although of great practical importance, especially for superconducting bolometers, we neglect the bias voltage in the following. This treatment follow the analysis of Mather (1982) and Jones (1953).

Let the receiving element of the thermal detector (Fig. 5.1) be a piece of radiation absorbing material coupled to a heat sink at a constant temperature T_0. The temperature response of this element to power absorption will be influenced both by the thermal capacity and the thermal conductance between receiving element and heat sink. A relation for the temperature response can be deduced from an analogy with an $R-C$ circuit. If capacity and conductance are denoted by $\mathscr{C}$ and $\mathscr{G} = 1/R$, respectively, the energy balance equation is

$$\mathscr{C}\frac{\mathrm{d}\Delta T}{\mathrm{d}t} + \mathscr{G}\cdot\Delta T = P, \tag{5.1}$$

where ΔT is the temperature increase of the receiving element above its (thermodynamic) equilibrium value T_0. P denotes the power absorbed. For a steady power flow, eventually a constant temperature is reached; when $\mathrm{d}\Delta T/\mathrm{d}t = 0$ we find

$$\Delta T = \frac{P}{\mathscr{G}}. \tag{5.2}$$
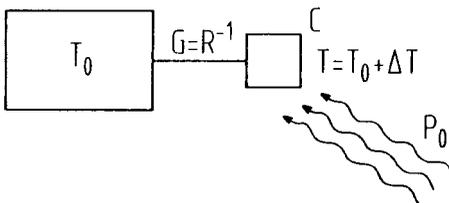
Fig. 5.1 A sketch of a bolometer represented by the smaller square to the *right*. The power from an astronomical source, P_0, raises the temperature of the bolometer element by ΔT, which is much smaller than the temperature T_0 of the heat sink. Heat capacity, $\mathscr{C}$, is analogous to capacitance. The conductance, $\mathscr{G}$ is analogous to electrical conductance, G, which is $1/R$. The noise performance of bolometers depends critically on the thermodynamic temperature, T_0, and on the conductance $\mathscr{G}$. The temperature change causes a change in the voltage drop across the bolometer (the electric circuit is not shown)

If the power flow is suddenly stopped, the temperature at a time t later is

$$\Delta T = \frac{P}{\mathscr{G}} \, e^{-t/\tau},\tag{5.3}$$

where

$$\tau = \mathscr{C}/\mathscr{G}\tag{5.4}$$

is the thermal time constant of the element. Usually the radiation incident on the bolometer is modulated at a chopper frequency v. Then we can write

$$P = P_0 \, e^{2\pi i v t}.\tag{5.5}$$

Ignoring any phase shifts in the system response, the solution of (5.1) is

$$\Delta T = \frac{P_0 \, e^{2\pi i v t}}{\mathscr{G}(1 + 2\pi i v \tau)}.\tag{5.6}$$

The amplitude of the resulting temperature variation is

$$|\Delta T| = \frac{P_0}{\mathscr{G}\sqrt{1 + (2\pi v \tau)^2}}.\tag{5.7}$$

For modulating frequencies fast compared to $1/\tau$ the temperature response falls off as $1/v$, while for $v\tau \ll \frac{1}{2}\pi$ the result reduces to the steady-state response. In practical use of bolometers τ falls within the range of milliseconds to seconds.

For a bolometer to be a useful detector in astronomy it must fulfill several requirements. It should

- respond with a maximum temperature step ΔT to a given power input,
- have a short thermal time constant τ so that chopping frequencies faster than instrumental and weather changes can be used,
- produce a detector noise which is as close to the theoretical minimum as possible.

The first two items require a detector for which both the thermal heat capacity $\mathscr{C}$ and the thermal coupling to the heat sink $\mathscr{G}$ are optimal for a given situation. In the ideal case, one wants to maximize the absorption and minimize the capacity.

5.1.2   The Noise Equivalent Power of a Bolometer

We next consider the minimum noise obtainable with a bolometer. The noise sources are:

- Johnson noise in the bolometer,
- thermal fluctuations, or *phonon noise*,
- background photon noise, and
- noise from the amplifier and load resistor.

Cooling will reduce all of these noise contributions. For ground based bolometers the background photon noise will determine the noise of the system. We will give a simplified semiclassical derivation of this noise, a full quantum statistical derivation is possible [see e.g. Mather (1982) or Griffin and Holland (1988)].

A frequently used measure for the quality of a detector is its *noise equivalent power*, NEP, defined as the power which must fall on the detector to raise output by an amount equal to the RMS noise. This is defined as the response to a sinusoidally modulated input which is switched between two temperatures.

For a Black Body radiation field, the square of the RMS fluctuations in the number of photons is

$$(\Delta n_{\text{RMS}})^2 = n(n+1) = n^2 + n, \tag{5.8}$$

where n is the photon occupation number

$$n = \frac{1}{e^{h\nu/kT} - 1}. \tag{5.9}$$

The first term in (5.8) dominates in the Rayleigh-Jeans limit when $n \gg 1$. Thus we retain only this term. To relate occupation numbers to macroscopic quantities such as power, we must account for density of states factor, collecting area, A, and the solid angle, Ω, of the background as seen by the bolometer. The density of states factor is $(2h\nu^3/c^2)$; to establish the proper units, we need a factor for photon energy, $h\nu$. The total RMS value of fluctuations is twice the simpler expression, because of arguments similar to those used for the extra factor of 2 in connection with Johnson noise (Chap. 1). Then we have

$$(\Delta P_{\text{RMS}})^2 = 2\Omega A \int_0^\infty \left(\frac{2h\nu^3}{c^2}\right) \left(\frac{1}{e^{h\nu/kT} - 1}\right) h\nu \frac{d\nu}{e^{h\nu/kT} - 1}. \tag{5.10}$$

For a narrow band of frequencies ν_0 to $\nu_0 + \Delta \nu$, this is

$$(\Delta P_{\text{RMS}})^2 = 4\Omega A \frac{h^2}{c^2} \int_{\nu_0}^{\nu_0 + \Delta \nu} \frac{\nu^4}{(e^{h\nu/kT} - 1)^2} d\nu. \tag{5.11}$$

Using $h\nu \ll kT$, we obtain

$$(\Delta P_{\text{RMS}})^2 = \frac{4\Omega A}{\lambda^2} (kT)^2 \Delta \nu. \tag{5.12}$$

The bolometer area, A, can be considered in many respects to be an antenna receiving energy in the field of the background radiation. Then, for a simple antenna (see Eq. 7.11), $\Omega A = \lambda^2$, so this expression can be simplified. We have neglected the factor ε, for the emissivity of the background. If we include this, we have,

$$\boxed{\text{NEP}_{\text{ph}} = 2\varepsilon k T_{\text{BG}} \sqrt{\Delta \nu}} \quad . \tag{5.13}$$

If $\varepsilon = 0.5, T_{BG} = 300\,K$ and $\Delta\nu = 50\,GHz$ then $NEP_{ph} = 9.3 \times 10^{-16}$ Watts $Hz^{-1/2}$.

With the collecting area of the 30 m IRAM telescope and a 100 GHz bandwidth one can easily detect mJy sources. This analysis is based on the assumption that the Johnson noise and thermal fluctuations in the bolometers are negligible, which is usually the case. There are other drawbacks: large bolometer bandwidths may lead to a contamination of the continuum response by intense spectral lines in the passband.

5.1.3   Currently Used Bolometer Systems

Bolometers mounted on ground based radio telescopes are background noise limited, so the only way to substantially increase mapping speed for extended sources is to construct large arrays consisting of many pixels. In present systems, the pixels are separated by 2 beamwidths, because of the size of individual bolometer feeds. The systems which best cancel atmospheric fluctuations are composed of rings of close-packed detectors surrounding a single detector placed in the center of the array. Two large bolometer arrays have produced many significant published results. The first is MAMBO2 (MAx-Planck-Millimeter Bolometer). This is a 117 element array used at the IRAM 30-m telescope. This system operates at 1.3 mm, and provides an angular resolution of $11''$. The portion of the sky that is measured at one instant is the *field of view*, (FOV). The FOV of MAMBO2 is $240''$. The second system is SCUBA (Submillimeter Common User Bolometer Array). This is used on the James-Clerk-Maxwell (JCMT) 15-m sub-mm telescope at Mauna Kea, Hawaii. SCUBA consists of a 37 element array operating at 0.87 mm, with an angular resolution of $14''$ and a 91 element array operating at 0.45 mm with an angular resolution of $7.5''$; both have a FOV of about $2.3'$. The LABOCA (LArge Bolometer CAmera) array operates on the APEX 12 m telescope. APEX is on the 5.1 km high Chaijnantor plateau, the ALMA site in north Chile. The LABOCA camera operates at 0.87 mm waqvelength, with 295 bolometer elements. These are arranged in 9 concentric hexagons around a center element. The angular resolution of each element is $18.6''$, the FOV is $11.4'$.

5.1.3.1   Superconducting Bolometers

A promising new development in bolometer receivers is *Transition Edge Sensors* referred to as TES bolometers. These superconducting devices may allow more than an order of magnitude increase in sensitivity, if the bolometer is not background limited. For broadband bolometers used on earth-bound telescopes, the warm background limits the performance. With no background, the noise improvement with TES systems is limited by photon noise; in a background noise limited situation, TES's should be $\sim$2–3 times more sensitive than semiconductor bolometers.

For ground based telescopes, TES's greatest advantage is multiplexing many detectors with a superconducting readout device, so one can construct even larger arrays of bolometers. SCUBA will be replaced with SCUBA-2 now being constructed at the Astronomy Technology Center, Edinburgh. SCUBA-2 is an array of 2 TES bolometers, each consisting of 6400 elements operating at 0.87 mm and 0.45 mm. The FOV of SCUBA-2 will be $8'$. The SCUBA-2 design is based on photo-deposition technology similar to that used for integrated circuits. This type of construction allows for a closer packing of the individual bolometer pixels. In SCUBA-2 these will be separated by 1/2 of a beam, instead of the usual 2 beam spacing.

5.1.3.2   Polarization Measurements

In addition to measuring the continuum total power, one can mount a polarization-sensitive device in front of the bolometer and thereby measure the direction and degree of linear polarization. The polarimeter used with SCUBA consists of a rotatable quartz half-wave plate and a fixed etched grid mounted in front of the SCUBA cryostat. The waveplate introduces a $\lambda/2$ phase lag between the planes of polarization. The signal is switched between sky positions by use of a nutating subreflector. Then the direction of the $\lambda/2$ plate is changed, and the procedure is repeated. Another instrument is PolKA. With PolKA one rotates the $\lambda/2$ plate continuously, without nutating the subreflector. This rotation of the $\lambda/2$ plate gives rise to a modulated signal which is proportional to the polarized signal. Polarized thermal emission from dust grains has been measured in a number of sources with this device (see Chap. 10 for the details of dust emission).

5.1.3.3   Spectral Line Measurements

Thus far, the presentation of bolometers has concentrated on broadband continuum emission. It is possible to also carry out spectroscopy, if frequency sensitive elements, either Michelson or Fabry-Perot interferometers, are placed before the bolometer element. Since these spectrometers operate at the sky frequency, the frequency resolution ($\nu/\Delta\nu$) is limited. One such instrument is the South Pole Imaging Fabry-Perot Interferometer, SPIFI (Stacey et al. 2002). SPIFI is a multi-beam Fabry-Perot system working at 0.3 mm with a velocity resolution of about 300 km s^{-1}. SPIFI is designed to measure $J = 7 - 6$ carbon monoxide rotational spectra and the $^3P_2 - ^3P_1$ fine structure line of carbon simultaneously (see Chap. 13, Table 13.1 and Chap. 15, Sect. 15.8ff).

5.2   Coherent Receivers

We first provide a simplified derivation of the minimum noise of a coherent system. We then give descriptions of the major components of a receiver, and describe specific types of *front ends*. Then we give a description of *back ends* which are used to extract continuum, polarization, spectral line or pulsar data.

5.2.1   The Minimum Noise in a Coherent System

The ultimate limit for a coherent receiver or an amplifier is obtained by application of the *Heisenberg uncertainty principle*. We start with the familiar relation:

$$\Delta E \Delta t \geq h/4\pi \,, \tag{5.14}$$

This must be cast in a slightly different form. It can be rewritten in terms of the uncertainty in the number of photons and the uncertainty in phase:

$$\Delta E = h\nu \Delta n \tag{5.15}$$

and

$$2\pi\nu\Delta t = \Delta\phi\,. \tag{5.16}$$

Inserting relations (5.15) and (5.16) in relation (5.14), we obtain

$$\Delta\phi\,\Delta n \geq 1/2\,. \tag{5.17}$$

The equality in relation (5.17) is reached when both the photon number and phase are Gaussian distributed.

We now apply relation (5.17) to obtain the desired result. A noiseless amplifier with gain $G > 1$ has the property that n_1 photons at the input produce $n_2 = Gn_1$ photons at the output. In addition, the output phase ϕ_2 equals the input phase ϕ_1 plus some constant phase shift. Then an ideal detector at the output of the amplifier must obey relation (5.17):

$$\Delta\phi_2\,\Delta n_2 = 1/2\,. \tag{5.18}$$

But then the uncertainty in input photon number is $\Delta n_1 = \Delta n_2/G$, and the uncertainty in the phase remains the same. Then at the amplifier input, the uncertainty relation would be

$$\Delta\phi_1\,\Delta n_1 = 1/2\,G\,. \tag{5.19}$$

But this is in contradiction to relation (5.17). The only way to avoid this contradiction is to assume that the amplifier adds some noise. The minimum amount, per unit bandwidth, needed to satisfy relation (5.18), at the output of the amplifier is $(G-1)h\nu$. When referred to the input of the amplifier, this is $(1-1/G)h\nu$. To minimize the noise contribution from following stages of amplification, we let G

increase to a large value. Then the *minimum* noise of the amplifier is $h\nu$, which results in a receiver noise temperature of

$$T_{\mathrm{rx}}(\mathrm{minimum}) = \frac{h\nu}{k} \qquad .$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (5.20)

For incoherent detectors, such as bolometers, phase is not preserved, so this limit does not exist. In the centimeter and even millimeter wavelength regions, this noise temperature limit is quite small. For example, at 2.6 mm, it is 5.5 K. However, at a wavelength of 0.3 mm, the limit is 47.8 K. Presently, the best receiver noise is $\cong 5$ times these values. This derivation is valid for receiver noise temperatures that are rather far above the quantum limits. As pointed by Kerr, Feldman and Pan, (1996), for receiver noise temperatures below 40 K, the effect of the zero point energy expressed in temperature units, $\frac{h\nu}{2k}$ may be important. There are a number of subtle effects, but in practice, for a given value of the **y** factor (see Fig. 4.10), this effect raises the receiver noise estimate by 10%.

5.2.2 Basic Components: Passive Devices

5.2.2.1 Thermal Noise of an Attenuator

Attenuators appear at many positions in the circuit of a radiometer, either deliberately in order to reduce the amplitude of a too large input or simply present as a "lossy" piece of connecting cable, connector, switch etc. The equation of radiative transfer together with Kirchhoff's law can be used to determine the noise power emitted by such a device if in Local Thermodynamic Equilibrium, LTE. The PSD at the output of the attenuator is obtained by integrating the transfer equation (1.9) along the signal path.

5.2.2.2 Isolators

Isolators are non-reciprocal devices, that is, these circuit elements allow power to pass in one direction only. Isolators are used to prevent power reflections that arise in one part of the receiver system from affecting other parts of the system. Isolators consist of circuit elements containing magnetic materials that are in strong magnetic fields. These elements are arranged so that a linearly polarized wave entering from one direction is Faraday rotated so that it can propagate further. A wave entering from the other direction cannot propagate. Thus, for a given direction of propagation and magnetic field, this device will favor one direction over the other.

5.2.2.3   Directional Couplers

These elements allow a certain amount of power to be diverted into another part
of the system. In waveguides, in the simplest case, these elements consist of two
openings separated by a quarter of a wavelength. In one direction the waves emitted
from these openings reinforces, while in the opposite direction, the waves cancel.
More complex versions consist of multi-hole couplers.

5.2.2.4   Phase Lock Systems

The purpose of a phase lock loop system (PLL) is to provide a stable frequency,
in both phase and frequency. This is needed for the coversion of frequencies in
heterodyne receivers. The essential features of a PLL are: (1) a voltage controlled
oscillator (VCO), i.e., one that changes frequency when the input voltage changes,
(2) a phase comparitor that produces a signal proportional to the difference of phases
of two inputs, and (3) a low pass filter. For item (2), the two inputs are from a
reference source and from the output of the VCO. We show a schematic of a PLL in
Fig. 5.2.
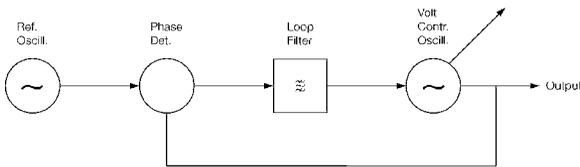
Fig. 5.2 A sketch of the Phase Lock Loop (PLL) which is used to control the LO frequencies in
the microwave range

5.2.3   Basic Components: Active Devices

5.2.3.1   Cascading of Amplifiers

The power amplification needed for a practical receiver is of the order of 80–100 dB.
Such a large amplification can only be obtained by cascading (Fig. 5.3) several am-
plification stages each with the gain G_i resulting in the total gain

$$G = \prod_{i=1}^{n} G_i .$$

The question is now: what is the total noise temperature of the cascaded system
if each individual stage contributes the noise temperature T_{Si}?
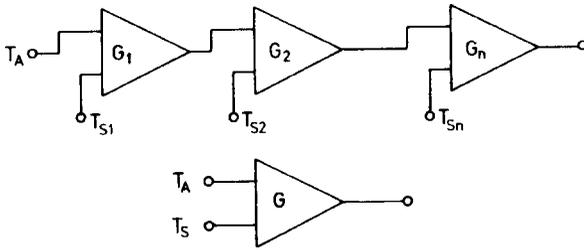
Fig. 5.3 Cascading of amplifiers. In the upper part is a sketch showing the cascading of amplifiers. The inputs to each amplifier are the signal (*upper*) and internal noise (*lower*). In the lower part of this figure is the combined amplifier chain, with signal and noise inputs

If the input PSD of stage 1 is

$$P_0 = kT_A \tag{5.21}$$

then stage i produces an output PSD

$$P_i(\nu) = [P_{i-1}(\nu) + kT_{Si}] G_i(\nu). \tag{5.22}$$

The appropriate definition of the total system noise temperature T_S of a system with the total gain $\prod G_i$ is

$$P_n(\nu) = k(T_A + T_S) \prod_{i=1}^{n} G_i(\nu). \tag{5.23}$$

Substituting (5.21) and (5.22) into (5.23), we obtain the *Friis formula* which takes into account the effect of having cascaded amplifiers :

$$T_S = T_{S1} + \frac{1}{G_1} T_{S2} + \frac{1}{G_1 G_2} T_{S3} + \cdots + \frac{1}{G_1 G_2 \ldots G_{n-1}} T_{Sn} \tag{5.24}$$

If several amplification stages are necessary, these should be arranged so that the amplifier with the lowest noise temperature is used first; for the second and following stages, the noise temperature can be higher. Another important point is that after amplification the output power can be divided into several branches without introducing much additional noise into the system. Thus the output of a single receiver can be used to provide a signal to many devices without worsening the signal-to-noise ratio.

Purely lossy devices such as filters or mixers have $G < 1$. This is usually written as $L = 1/G$, and is referred to as *conversion loss*. Classical mixers operated in the DSB mode with equal response in the signal and image sidebands typically have 3 db loss.

In the case of interferometry (Chap. 9), amplified signals from an individual antenna can be correlated with a large number of other antennas without a significant loss in the signal-to-noise ratio.

5.2.3.2 Mixers

Shifting the signal frequency is useful for two reasons:

- one avoids a feedback of amplified signals into the frontend. High-gain cascaded amplifier chains are often affected by instabilities. If the total gain is of the order of 10^8 to 10^{10} (80–100 dB) an exceedingly small amount of power leaking from the output port back to the input port is sufficient to cause the system to oscillate violently.
- one can choose a frequency at which the signal is more easily amplified. One shifts the frequency of output signal from that of the input by mixing the signal with a monochromatic signal from a *local oscillator*.

The local oscillator is usually referred to as the local oscillator, or *LO*. The process of mixing may shift the phase of the signal by a constant value. Except for additional noise from the mixing process, the information contents of the sifted signal should not be changed by the mixing process.

A *mixer* is the device that is performing the actual frequency shift. In principle any circuit element with a nonlinear relation between input voltage and output current can be used as a mixer. However, derivations of mixer properties are simplest for a device with a purely quadratic characteristic. Mixers are an essential part of *heterodyne* receiver. A semiconductor metal junction can be used as a mixer. Applying both a signal and a local oscillator frequency at the input of a Schottky junction, one can produce a microwave mixer device. That is, the sum and difference of the frequencies at the input will appear at the output. The quality of such a mixer is dependent on the change in the current-voltage characteristic near the voltage at which it is operated, that is, the operating point.

$$I = \alpha U^2. \tag{5.25}$$

Where U is the sum of the signal $E\sin(2\pi v_S t + \delta_S)$ and the local oscillator $V\sin(2\pi v_{LO} t + \delta_{LO})$. Then the output is

$$
\begin{aligned}
I &= \alpha\left[E\sin(2\pi v_S t + \delta_S) + V\sin(2\pi v_{LO} t + \delta_{LO})\right]^2 \\
&= \alpha E^2 \sin^2(2\pi v_S t + \delta_S) + \alpha V^2 \sin^2(2\pi v_{LO} t + \delta_{LO}) \\
&\quad + 2\alpha EV \sin(2\pi v_S t + \delta_S)\sin(2\pi v_{LO} t + \delta_{LO}) \tag{5.26}
\end{aligned}
$$

Using trigonometric addition formulae, one obtains

$$
\begin{aligned}
I = {}&\tfrac{1}{2}\alpha\left(E^2 + V^2\right) &&\text{(DC component)} \\
&-\tfrac{1}{2}\alpha E^2 \sin(4\pi v_S t + 2\delta_S + \tfrac{\pi}{2}) &&\text{(2nd harmonic of signal)} \\
&-\tfrac{1}{2}\alpha V^2 \sin(4\pi v_{LO} t + 2\delta_{LO} + \tfrac{\pi}{2}) &&\text{(2nd harmonic of LO)} \\
&+\alpha VE \sin[2\pi(v_S - v_{LO})t + (\delta_S - \delta_{LO} + \tfrac{\pi}{2})] &&\text{(difference frequency)} \\
&-\alpha VE \sin[2\pi(v_S + v_{LO})t + (\delta_S + \delta_{LO} + \tfrac{\pi}{2})] &&\text{(sum frequency).} \tag{5.27}
\end{aligned}
$$
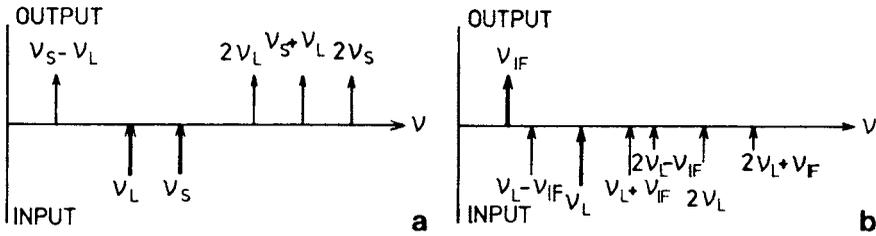
Fig. 5.4 Input and output frequencies of a mixer. The thick arrows are the given values. There are two methods to specify the system: (**a**) when two input frequencies, ν_{LO} and ν_S are given. In case (**b**) when ν_{LO} and ν_{IF} are specified. In this case, signals from both the upper sideband $(\nu_{LO} + \nu_{IF})$ and lower sideband $(\nu_{LO} - \nu_{IF})$ contribute to the IF signal

The output consists of the superposition of several components at different frequencies (Fig. 5.4): a DC signal, signals at twice the signal and the local oscillator frequencies, and two components at the difference and the sum of signal and oscillator frequencies. While the amplitudes of all other components depend on the second power of signal or local oscillator amplitude, the sum and difference frequency signals depend on the first power. Thus their amplitudes are accurate reproductions of the amplitude of the input signal.

By use of an appropriate bandpass filter, all but the desired signal can be suppressed. In this way the mixer can be considered to be a *linear* device producing an output at the frequency $\nu_{IF} = \nu_S - \nu_{LO}$. This is also the case for devices with characteristic curves different from (5.25). Filters give rise to a loss of signal, so for some applications a filter will not be placed before the mixer. Then the mixer is used as a double sideband (DSB) device. The output of a DSB mixer is shown in Fig. 5.5. For a given local oscillator frequency, two frequency bands, above and below the LO frequency, separated by the intermediate frequency (IF) frequency, are shifted into the IF band. Thus, a mixer will shift two frequency bands into the same band. Usually one sideband is wanted, but the other not. These are referred to as the *signal* and *image* bands. Mixers that consist of a single non-linear circuit element accept both sidebands. In the millimeter or sub-millimeter wavelength
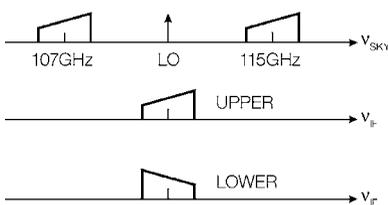
Fig. 5.5 A sketch of the frequencies shifted from the sky frequency (*top*) to the output (*lower*) of a double sideband mixer. In this example, the input is at the sky frequencies for the Upper Side Band (USB) of 115 GHz, and Lower Side Band (LSB) of 107 GHz while the output frequency is 4 GHz. The slanted boxes represent the passbands; the direction of the slant in the boxes indicate the upper (*higher*) and lower (*lower*) edge of the bandpass in frequency

ranges, such mixers are still commonly used as the first stage of a receiver. For single dish continuum measurements, both sidebands contain the signal, so in this case, DSB operation does not decrease the signal-to-noise (S/N) ratio. However, for single dish spectral line measurements, the spectral line of interest is in one sideband only. The other sideband is then a source of extra noise and perhaps confusing lines. Therefore single sideband (SSB) operation is desired. If the image sideband is eliminated, the mixer is said to operate in SSB mode. This can be accomplished by inserting a filter before the mixer. However, filters are lossy elements. Thus this procedure will increase the system noise temperature. If the mixer is first element of a receiver, the degradation of the system will be significant, so the filter-mixer combination should be used after the signal is amplified. If a mixer is used as the first circuit element in a receiver, it is better to make use of a single sideband mixer. See Fig. 5.6 for asketch of suc a device. SSB mixers require two matched mixers fed by a single local oscillator as well as additional circuit elements. Noise in mixers has 3 causes. The first is the mixer itself. Since one half of the input signal is shifted to a frequency $v_{LO} + v_{IF}$, the signal input power is a factor-of-two (3 db) loss of signal. This is referred to as *conversion loss*. The simplest form of classical mixers typically have 3 db loss. In addition there will be an additional noise contribution from the mixer itself. Second, the LO may have "phase noise", that is a rapid change of phase, which will add to the uncertainty. Third, the amplitude of the LO may vary; however this last effect can be minimized. For low levels of LO power, the output power and thus the response of a mixer will increase linearly with LO power. However, variations in the local oscillator power will appear as variations of the total gain of the system. Usually the mixer LO power is adjusted so that the mixer output is saturated. Then no variation of the output signal power is seen if LO power varies. This insures that the output remains within the operational range.
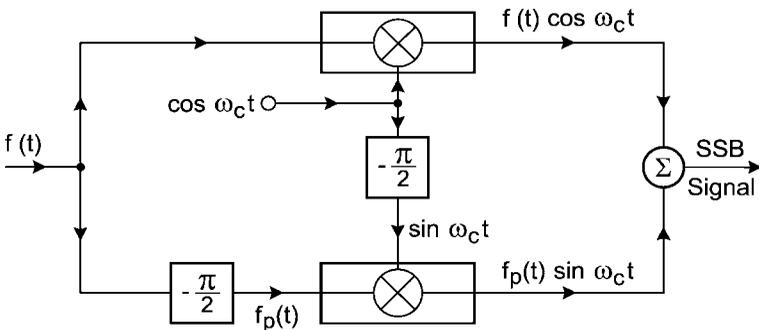
Fig. 5.6 A sketch of the single sideband mixer (SSB). The input signal, $f(t)$, is divided into two equal parts. There are two identical mixers located in an *upper* and *lower* branch of the sketch. The LO frequency from a central source, ω_c, is shifted in phase by $\pi/2$ from the input to the output of the mixer in the *lower part* of the sketch. In the *lower* branch, the phase of the input signal is also shifted by $\pi/2$. After mixing the signals are added to produce the single sideband output

5.2.3.3   Local Oscillator Sources

There are many possible LO sources. In the meter and centimeter wavelength range, one can make direct use of the output from commercially available frequency synthesizers. These devices are rather stable, but their output should be compared to known signals. Ultimately the frequency might be derived from a frequency standard (see the upper left of Fig. 5.2). The method used in comparisons with this standard will be discussed in the next section. In the few GHz to perhaps 100 GHz range, YIG (Yttrium-Iron-Garnet) oscillators are used in receivers and test instruments. YIG oscillators have wide tuning range, and produce a signal that has a small frequency width. YIG oscillators are tuned by varying an external magnetic field. At frequencies higher than 40 GHz, the output from YIG oscillators or commercial frequency synthesizers becomes impractical. Thus the output of a synthesizer is passed through a non-linear element, a multiplier, to produce a higher frequency. Frequently the output of an LO source is doubled, tripled or quadrupled. This process is similar to that described for mixers, but with only a local oscillator input. In Fig. 5.4 the output at $2\nu_{LO}$ would represent the output of a doubler; usually the non-linear element is optimized to enhance the desired harmonic. In some cases, a series of multipliers is needed to reach the desired frequency. It is possible, for example, that the output of a frequency source is doubled, then amplified, then tripled, then amplified again. The amplifiers would be tuned to the desired frequency to avoid spurious output frequencies. For frequencies of about 100 GHz and higher, Gunn oscillators, perhaps with multipliers, are used to produce the LO signal at a micro-Watts levels. Gunn oscillators are Gallium Arsenide crystals which oscillate when a voltage is applied.

For the ALMA project, a completely different approach is used. Here a phase and frequency stable microwave signal for the range 30–900 GHz must be distributed over more than 10 km to each receiver. This is done by first producing two modulated laser signals which are brought to the receivers by optic fibers. At the receivers, the laser signals are mixed to produce a microwave signal. This signal is then multiplied and amplified to produce the needed LO frequency.

5.2.4   Semiconductor Junctions

Semiconductor amplifiers are the first stages of the best centimeter systems. First, a review of a few essential concepts of the quantum theory of crystalline materials. In this outline, the relevant concepts are presented. For an isolated atom, a bound electron can only possess certain allowed energies, but when identical atoms are combined in a solid, such as a metal, we know from experience that electrons can move freely. Within a highly ordered crystal, a free electron can easily move only if it has certain energies. That is, the electron can move only in certain *energy bands*. By varying the material, both the width of the band, the *band gap*, and the minimum energy to reach a conduction band can be varied.

A widely used material for low-noise microwave applications is gallium arsenide, GaAs. In order to increase the current, a small number of impurity atoms is introduced. Usually silicon is adopted for a GaAs crystal. This addition of impurities is referred to as *doping*. These impurities might have one or two excess electrons compared to the basic material. In some cases, the doped material might have fewer electrons. Most importantly, the doping atoms are chosen to have about the same size, so that the crystal structure remains the same. The obvious choices are neighboring elements in the periodic table. There are some extra conditions dependent on purely chemical considerations.

The crucial part of any semiconductor device is the junction. On the one side there is an excess of material with negative carriers, forming n-type material and the other side material with a deficit of electrons, that is p-type material. The p-type material has an excess of positive carriers, or *holes*. At the junction of a p- and n-type material, the electrons in the n-type material can diffuse into the p-type material (and vice-versa), so there will be a net potential difference. The diffusion of charges, p to n and n to p, cannot continue indefinitely, but a difference in the charges near the boundary of the n and p material will remain, because of the low conductivity of the semiconductor material.

From the potential difference at the junction, a flow of electrons in the positive direction is easy, but a flow in the negative direction will be hindered.

The current caused by the positive carriers is the same, and the relation remains valid (see Fig. 5.7). Such p-n junctions have a large capacitance, so there can be no fast response. Thus these are suitable only as square-law detectors. Schottky junctions have a lower capacitance, so are better suited to applications such as microwave mixers. The *I-V* curves of Schottky mixers are similar to the curves for conventional diodes. A sharper curve provides a more efficient conversion.

A simple extension of the p-n junction is the combination of three layers, p-n-p, in a so-called "sandwich". In Field Effect Transistors, FET's, the electric field of the gate, G, controls the carrier flow from source, S, to drain, D. Small variations in the gate potential have large effects on the current flow from source to drain, so this is an amplifier. The direct extension of such a concept is the *bipolar transistor*, which operates by the motion of both *holes* and *electrons*. Such devices have
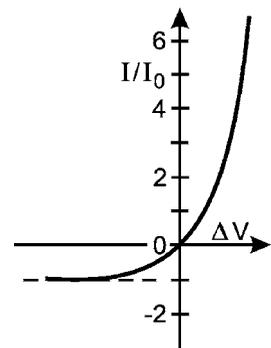
Fig. 5.7 A sketch of the current flow in a diode as a function of applied voltage, this relation, the *I-V* curve, is typical for classical mixers

slow response times, so their use is restricted to less than a few GHz. For example, uncooled Heterojunction Bipolar Transistors are useful up to 6 GHz.

At higher frequencies, unipolar devices, which have only one type of carrier, are used as microwave amplifier front ends. These are Field Effect Transistors, FETs. High Electron Mobility Transistors, HEMTs, are an evolution of FETs. The design goals of HEMT's are: (1) to obtain lower intrinsic amplifier noise and (2) operation at higher frequency. In HEMTs, the charge carriers are present in a channel of small size. This confinement of carriers is arranged by having the channel at the interface of two materials. In the first HEMTs, one used GaAs and AlGaAs as the two materials. In Fig. 5.8 we show a sketch of a High Electron Mobility Transistor or HEMT. The AlGaAs contributes electrons. These diffuse only a small distance into the GaAs because of the positive space charge in AlGaAs. Thus the electrons are confined to a narrow layer which is a potential well. This confinement gives rise to a two dimensional electron gas, or "2 DEG". We have denoted this region as "channel" in Fig. 5.8. Flows in regions containing doped ions give rise to larger scattering of electrons. Since the carriers are located in the 2 DEG region where there are no doped ions, there is less scattering and hence lower noise. When cooled, there is a significant improvement in the noise performance, since the main contribution is from the oscillations of nuclei in the lattice, which are strongly temperature dependent.

To extend the operation of HEMT to higher frequencies, one must increase electron mobility, μ, and saturation velocity V_s. A reduction in the scattering by doped ions leads to a larger electron mobility, μ, and hence faster transit times, in addition
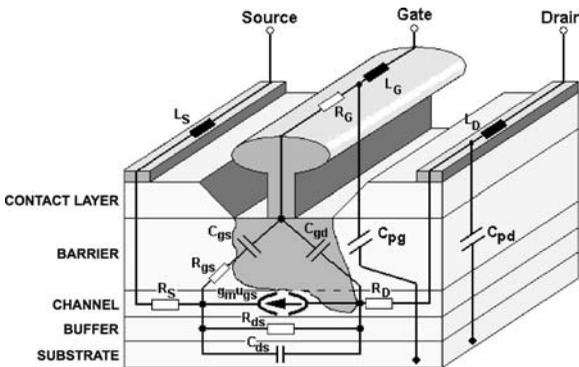
Fig. 5.8 This figure shows a HEMT amplifier. As with FETs, the current flow from Source to Drain, around the Gate. The electric field from the Gate is shown as a darker, irregular region, has a large effect on the current flow from Source to Drain. Thus, this amplifies the signal placed on the Gate. Because of the potentials in the interface layers, the electrons can flow from Source to Drain only in a very thin layer. This is shown enclosed in semicircles; this part of the HEMT provides the gain. The quantity g_m is the transconductance, and u_{gs} is the velocity from gate to source. The product of these is the gain of the amplifier. The quantities labelled "L_s, L_D, etc." represent inductances internal to the HEMT; the "R"s are internal resistances, and "C"s are internal capacitances

to lower amplifier noise. The maximum saturated velocity is the limit to the value of carrier velocity as the source-drain voltage is increased. Reducing the scattering and increasing V_{sat} leads to higher maximum operating frequencies, that is, higher cut-off frequencies. An exact analysis of HEMT behavior shows that the cut-off frequency is directly proportional to the saturation velocity, and inversely proportional to the sum of two terms: first, the width of the gate region and second, a correction for (effectively) the fringing of the electric field from the gate.

5.2.5 Practical HEMT Devices

For use up to $v = 115$ GHz with good noise performance, one has turned to modifications of HEMTs based on advances in material-growth technology. This technology has led to the fabrication of junctions between dissimilar semiconductors. These are referred to as heterojunctions. There has been a significant improvement of carrier-transport properties for two reasons. The first is the quantum confinement of the electron carriers created by the heterostructure. The second is the use of modulation doping, which reduces ionized impurity scattering in the channel where conduction takes place. The performance improvements such as higher gain and lower noise, are directly related to the electron mobility, μ, saturated electron velocity, v_{sat}, and the channel sheet carrier concentration, ns_0. From the use of these different structures comes the name "pseudomorphic" HEMT, or PHEMTs. The heterostructure devices have evolved from GaAs HEMT, to Pseudomorphic HEMT (so-called PHEMT) grown on GaAs, to a composition lattice matched HEMT grown on Indium Phosphide, a so-called InP HEMT, to a GaAs metamorphic HEMT (MHEMT). InP HEMTs are used up to frequencies of 115 GHz. These have an additional layer of indium gallium arsenide, InGaAs, which has a different lattice constant, inserted between the doped AlGaAs and the GaAs buffer. In the InGaAs layer, enhanced electron transport compared to the GaAs is possible. Thus there is a higher electron density and higher current, as well as better confinement in the potential well than with conventional HEMTs. Since InGaAs has a different crystal lattice constant, the layer must be kept to less than 200 Å thick to insure that lattice strain is taken up coherently by the surrounding material. All of this is mounted on a carrier layer of GaAs, which serves as a buffer.

The InP based Indium-Aluminum-Arsenide (InAlAs/InGaAs) material heterostructure with a InGaAs-channel of 53–65% Indium has the advantage of higher bandgap discontinuity and higher saturation velocity, which leads to better performance at higher frequencies compared to GaAs-based PHEMTs. However, producing large numbers of these devices is difficult due to the brittle nature of InP substrates and small available wafer size. In addition, increasing the Indium composition in the device channel generally leads to a decrease in breakdown voltage due to enhanced impact ionization in the smaller bandgap material. In 1999, a GaAs-based metamorphic HEMT, or MHEMT, technology has emerged as a low cost alternative to InP-HEMTs. MHEMT technology has the potential to eventually displace

the InP HEMTs in millimeter-wave applications. In this new approach metamorphic buffers are used to accommodate the lattice mismatch between the GaAS-substrate and the active layers. Using the metamorphic buffer concept, it is expected that unstrained InAlAs/InGaAs heterostructures can be grown with approximately any InAs fraction. These metamorphic buffers are based on the controlled relaxation of the strain due to the mismatch between the layer and the substrate. A controlled relaxation is obtained by growing an approximately one μm thick alloy-like InAlAs and by varying the indium content with a lower value towards the substrate.

For low noise IF amplifiers, 4–8 GHz IF systems using GaAs HEMTs with 5 K noise temperature and more than 20 db of gain have been built. With InP HEMTs on GaAs-substrates, even lower noise temperatures are possible. As a rule of thumb, one expects an increase of 0.7 K per GHz for GaAs, while the corresponding value for InP HEMTs is 0.25 K per GHz. For front ends, noise temperatures of the amplifiers in the 18–26 GHz range are typically 12 K. High performance Metamorphic HEMTs (MHEMTs) are supplied by Raytheon, Filtronics and UMS. These may eventually replace InP HEMTs. High performance Pseudomorphic HEMTs are supplied by Mitsubishi and Fujitsu. See Fig. 5.9 for the photograph of such an IF amplifier.

At higher frequencies, the SEQUOIA receiver array uses Microwave Monolithic Integrated Circuits (MMIC's) in 32 front ends for a 16 beam, two polarization system. This development was pioneered by S. Weinreb. The MMIC is a complete amplifier on a single semiconductor, instead of using lumped components. The MMIC's have excellent performance in the 80–115 GHz region without requiring tuning adjustments. The simplicity makes MMIC's better suited for multi-beam systems. The noise temperatures of individual array elements are not as low as the very
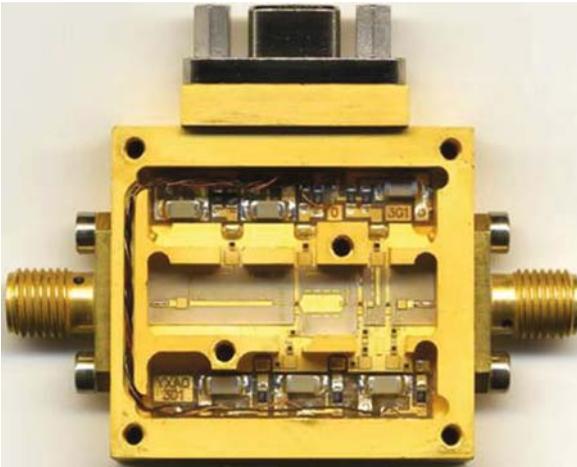
Fig. 5.9 An ALMA HEMT amplifier for 4–8 GHz. This low noise amplifier (LNA) was built by OAN, the Spanish National Observatory. This will be used as the first IF section after the SIS front end built at IRAM Grenoble. The receiver noise temperature is in the range of 5–7 K. The input is on the *right*, the output is on the *left*

best SIS devices, but the large number of beams compensates for this for the imaging larger regions.

5.2.6 Superconducting Mixers

Very general, semi-classical considerations indicate that the slope of the *I-V* curve shown in Fig. 5.7 changes gently. This is because the energy band gap energies are ≈ 1 V, much too large compared to the input energies of photons even at submillimeter wavelengths. This leads to a relatively poor noise figure, since much of the input signal is not converted to a lower frequency.

A significant improvement can be obtained if the junction is operated in the superconducting mode. Then the gap (see Fig. 5.10) between filled and empty states is ≈ 1 mV, and this is comparable to the photon energies at about 300 GHz. In addition, the local oscillator power requirements are ≈ 1000 times lower than are needed for conventional mixers. Finally, the physical layout of such devices is simpler since the mixer is a planar device, deposited on a substrate by lithographic techniques. SIS mixers consist of a superconducting layer, a thin insulating layer and another superconducting layer. A diagram of the energy levels is shown in Fig. 5.10. There is a gap between the filled states and the allowed unfilled states. In the filled states, the electrons are paired ("Cooper Pairs") and act as bosons which give rise to the Josephson phenomenon. The Josephson Effect increases the noise in an SIS mixer, so must be supressed. Thus, in addition to the mixer DC bias and LO signal, at frequencies above 120 GHz, one must apply a steady magnetic field to eliminate the Josephson Effect. SIS mixers depend on single carriers; a longer but more accurate
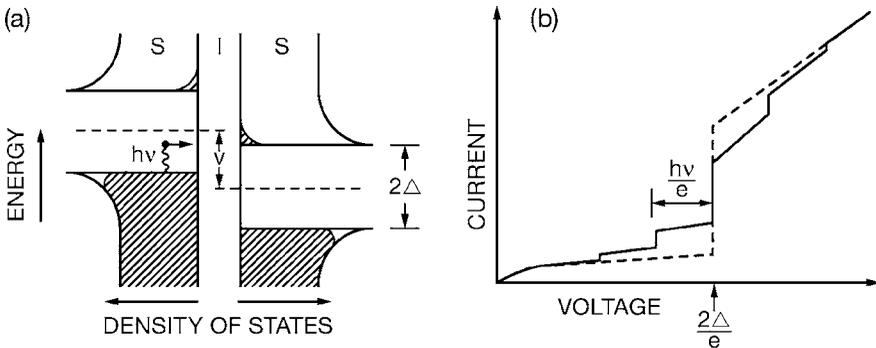
Fig. 5.10 (**a**) A sketch of the energy bands of a superconducting-insulating-superconducting (SIS) layer. The gap between the filled states (*below, shaded*) and the empty states (*above*) is 2Δ. On the left we sketch the process in which an electron absorbs a photon, gaining energy which allows it to tunnel through the insulating barrier. (**b**) The *I-V* curve of the SIS junction. The dashed line is the behavior when there is no local oscillator (LO) power; the solid line shows the behavior with LO power supplied. When DC biased at $2\Delta/e$, the SIS mixer efficiently converts photons to a lower frequency

description of SIS mixers is "single quasiparticle photon assisted tunneling detectors". When the SIS junction is biased to a value of $2\Delta/e$, the filled states on the left (see Fig. 5.10) reach the level of the unfilled band shown on the right, and the electrons can quantum mechanically tunnel through the insulating strip. In the *I-V* curve for a SIS device (Fig. 5.10) the sudden jumps in the *I-V* curve are typical of quantum-mechanical phenomena. For low noise operation, the SIS mixer must be DC biased at an appropriate voltage and current. If, in addition to the mixer bias, there is a source of photons of energy $h\nu$, then the tunneling can occur at a lower bias voltage, $h\nu/e$. If one then biases an SIS device and applies an LO signal at a frequency ν, the *I-V* curve becomes very sharp. There are other jumps in that curve at sub-harmonic frequencies, due to multiple photon absorptions. These can be minimized by filtering and proper biasing. For a weak (astronomical) signal present at frequency ν, the conversion of such photons to lower frequency is much more effective than with a classical mixer. We show a sketch of SIS junction mounted in a waveguide in Fig. 5.11. Under certain circumstances, SIS devices can produce gain. If the SIS mixer is tuned to produce substantial gain the SIS device is unstable, somewhat like the instability found with parametric amplifiers. Thus, this not useful in radio astronomical applications. In the mixer mode, that is, as a frequency converter, SIS devices can have a small amount of gain. This tends to balance inevitable losses, so SIS devices have losses that are lower than Schottky mixers. SIS mixers have performance that is unmatched in the millimeter region. Improvements to existing designs include *tunerless* and *single sideband* SIS mixers. Tunerless mixers have the advantage of repeatability when returning to the same frequency. Usually single sideband mixers require 2 backshorts. SIS mixers with a suppressed sideband use 2 or more identical junctions and a more complex LO system and electronics. For the Atacama Large Millimeter Array (ALMA) new SIS mixer designs have been developed. These are wideband, tunerless, single sideband devices with extremely low mixer noise temperatures. we show a photo of such a device in Fig. 5.12.

An increase in the gap energy, to allow the efficient detection of higher energy photons. This is done with Niobium superconducting materials with geometric junction sizes of 1 μm by 1 μm. For frequencies above 900 GHz, one uses niobium nitride
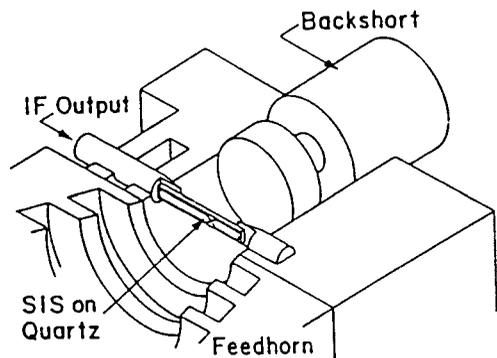
Fig. 5.11 A sketch of an SIS junction placed in a waveguide. Both the astronomical and LO signals enter through the waveguide; the difference frequency is present at the IF output. This response is optimized by tuning the back short
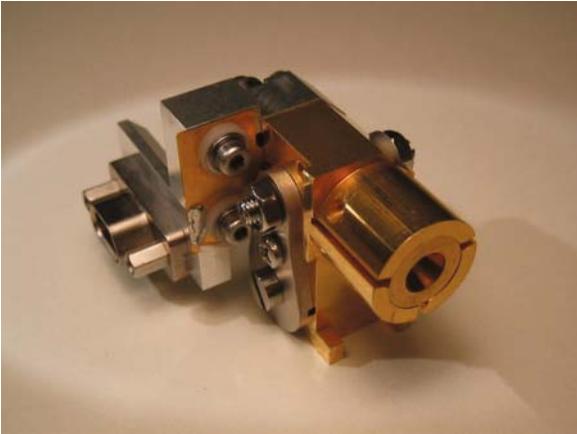
Fig. 5.12 [An ALMA Band 9 SIS mixer] A photo of the feed horn on the right, and the SIS mixer holder on the left. This was built at the Space Research Organization of the Netherlands (SRON). This receiver covers the frequency range 610–720 GHz. This is a tunerless double sideband mixer. This is typical for the sub-mm SIS mixers used in the ALMA instrument

junctions. Variants of such devices, such as the use of junctions in series, can be used to reduce the capacitance. An alternative is to reduce the size of the individual junctions to 0.25 μm.

SIS mixers are the front ends of choice for operation between 150 GHz and 900 GHz because:

- these are low-noise devices;
- the IF bandwidths can be >1 GHz;
- these are tunable over $\sim$30% of the frequency range
- the local oscillator power needed is <1 μW.

5.2.7 Hot Electron Bolometers

Superconducting Hot Electron Bolometer-mixers (HEB) are heterodyne devices, in spite of the name. These mixers make use of superconducting thin films which have sub-micron sizes. In an HEB mixer excess noise is removed either by diffusion of hot electrons out the junction, or by an electron-phonon exchange. The first HEBs operating on radio telescopes and used to take astronomical data were HEB's which made use of electron-phonon exchange. The HEB junctions were of μm size, consisting of Niobium Nitride (NbN), cooled to 4.2 K. Junctions using AlTiN have provided lower receiver noise temperatures. The IF center frequency was 1.8 GHz, and a had a full width of 1 GHz. Gershenzon et al. (1990) pioneered this development. The first astronomical measurements using an HEB device were carried out at 0.5 mm and 0.35 mm by the Blundell group from the Harvard-Smithsonian Center

for Astrophysics. A similar system was used to measure the $J = 9$–8 carbon monoxide line at 1.037 THz and later by the Köln University group using the Atacama Pathfinder EXperiment (APEX) telescope to measure the [N II] line at 1.5 THz.

HEB devices have the following advantages:

- the IF frequencies are >1 GHz, so the IF bandwidths can be >1 GHz;
- they can operate at wavelengths shorter than 0.3 mm, where present-day SIS mixer devices are approaching theoretical band-gap limits;
- these are low-noise devices;
- the local oscillator power needed is $<1 \mu$W;
- these are essentially resistive devices with $R = 20\Omega$ to 200Ω and with R independent of wavelengths to $\lambda = 2 \mu$m;
- At 1.3 mm, HEB devices have higher noise temperatures than SIS devices, however, for $\lambda < 0.3$ mm, HEBs have a clear advantage.

5.3 Summary of Front Ends Presently in Use

5.3.1 Single Pixel Receiver Systems

Devices that provide the lowest noise front ends are:

- for $\nu < 115$ GHz, High Electron Mobility Transistors (HEMT) and Microwave Monolithic Integrated Circuits (MMIC)
- for $72 < \nu < 800$ GHz, Superconducting Mixers (SIS)
- for $\nu > 900$ GHz, Hot Electron Bolometers (HEB)

See Fig. 5.13 for a comparison of front end receiver noise temperatures. For $\lambda > 3$ mm, HEMT front ends have now replaced just about all other types of systems. In the future, the performance of HEMTs may be extended to $\lambda = 1.3$ mm. In the cm range, Maser receivers may be somewhat more sensitive, but are much more complex systems. As a result, these are used only in very special circumstances. SIS mixers provide the lowest receiver noise in the mm and sub-mm range. SIS mixers are much more sensitive than classical Schottky mixers, and require less local oscillator power, but must be cooled to 4 Kelvin. All millimeter mixer receivers are tunable over 10–20% of the sky frequency. From the band gaps of junction materials, there is a short wavelength limit to the operation of SIS devices. For spectral line measurements at wavelengths, at $\lambda < 0.3$ mm, superconducting Hot Electron Bolometers (HEB), which have no such limit, have been developed. At frequencies above 2 THz there is a transition to far-infrared and optical techniques. The highest frequency heterodyne systems in radio astronomy are used in the Herschel-HIFI satellite. These are SIS and HEB mixers.

In addition to the front end mixers and amplifiers, the connections between feed and receiver are also often cooled. For some receivers sections of the feed horn with the coupling probes are cooled.
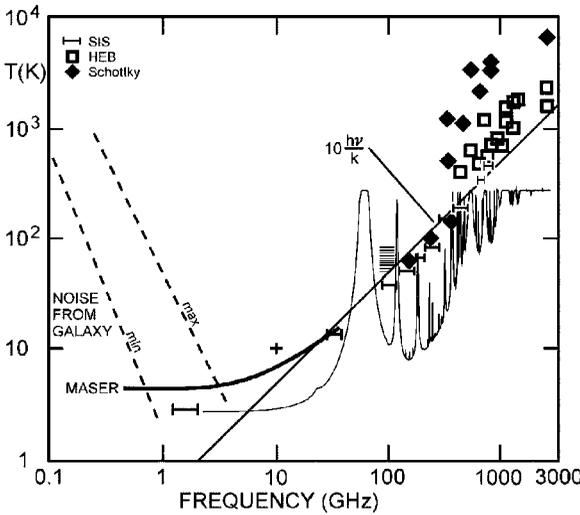
Fig. 5.13 Receiver noise temperatures for coherent amplifier systems compared to the temperatures from different astronomical sources and the atmosphere. The atmospheric emission is based on a model of zenith emission for 0.4 mm of water vapor (plot from B. Nikolic (Cambridge Univ.) from the "AM" program of S. Paine (CfA)). This does not take into account the absorption corresponding to this emission. In the 1–26 GHz range, the horizontal bars represent the noise temperatures of HEMT amplifiers (priv. comm. H. Mattes, MPIfR). The shaded region between 85 and 115.6 GHz is the receiver noise for the SEQUOIA array which is made up of monolithic millimeter integrated circuits (MMIC), at Five College Radio Astronomy Observatory. The meaning of the other symbols is given in the upper left of the diagram (partially taken from Rieke 2002). For the SIS mixers, we have used the ALMA specifications. These are single sideband mixers covering the frequency range shown by the horizantal bars. The mixer noise temperatures given as double sideband (DSB) values were converted to single sideband (SSB) temperatures by increasing the receiver noise by a factor of 2. The ALMA mixer noise temperatures are SSB. The HEMT values are SSB

The SIS or HEB mixers convert the RF frequency to the fixed IF frequency, where the signal is amplified by the IF amplifiers. Most of the amplification is done in the IF. The IF should only contribute a negligible part to the system noise temperature. Because some losses are associated with frequency conversion, the first mixer is a major source for the system noise. Two ways exist to decrease this contribution:

- using either an SIS or HEB mixer to convert the input to a lower frequency, or
- at lower frequencies using a low-noise amplifier before the mixer.

5.3.2 Multibeam Systems

Since HEMT front ends are rather simple receiver systems, there has been a trend to build many receivers in the focal plane. An array of N such receivers allows one

to map a given region N times faster than with a single receiver, although at the cost of more complexity. For spectral line mapping this involves both spectrometer hardware and software. Compared to single pixel receivers, such array systems are more complex but make more efficient use of observing time. These systems are usually mounted in the secondary focus because of weight and to avoid optical distortions. A 13 beam system for $\lambda = 21$ cm, using HEMT receivers, has been installed in the prime focus of the Parkes 64-m telescope in Australia. More than 300 pulsars have been discovered with this system. For spectroscopy, back ends with many channels are needed. At Parkes,one program involved blind searches for gas-rich-star-poor galaxies in the $\lambda = 21$ cm line . A 4-beam system for 21 cm, also using HEMT receivers, has been installed in the 76-m Lovell telescope at Jodrell Bank to complement the Parkes measurements. Another 21 cm system with 7 beams is ALFA, installed on the Arecibo 305-m radio telescope; this is used for both H I and pulsar measurements. At 3 mm, the SEQUOIA array receiver with 32 MMIC front ends connected to 16 beams had been used on 14-m telescope of the FCRAO for the last few years.

Multibeam system that use SIS front ends are rare. A 9 beam Heterodyne Receiver Array of SIS mixers at 1.3 mm, HERA, has been installed on the IRAM 30-m millimeter telescope to measure spectral line emission. To simplify data taking and reduction, the HERA beams are kept on a Right Ascension- Declination coordinate frame. HARP-B is a 16 beam SIS system in operation at the James-Clerk-Maxwell telescope. The sky frequency is 325–375 GHz. The beam size of each element is 14″, with a beam separation of 30″, and a FOV of about 2′. The total number of spectral channels in a heterodyne multi-beam system will be large. In addition, complex optics is needed to properly illuminate all of the beams. In the mm range this usually means that the receiver noise temperature of each element is larger than that for a single pixel receiver system, unless great care is taken.

For single dish continuum measurements at $\lambda < 2$ mm, multi-beam systems make use of bolometers. In comparison to incoherent recievers, heterodyne systems are still the most efficient receivers for spectral lines in the range $\lambda >0.3$ mm, although systems such as SPIFI may be competitive for some projects. Presently, the cooled GeGa bolometers are the most common systems and the best such systems have a large number of beams. In the future, TES bolometers seem to have great advantages.

5.4   Back Ends: Correlation Receivers, Polarimeters and Spectrometers

In the following, to the end of this chapter, we describe the basic functions of the back ends which are used to extract information about polarization, spectra, and pulses in the data.

5.4.1   Correlation Receivers and Polarimeters

Dicke switching is only one possible method to stabilize a receiver system; another involves the correlation of signals. The block diagram for a correlation receiver is shown in Fig. 5.14. The signals from the antenna and from the reference are input to a 3 dB hybrid, a four-port device with two input and two output ports. If the signals at the inputs are $x(t)$ and $y(t)$, the outputs are $1/2[U_A(t) + U_{ref}(t)]$ and $1/2[U_A(t) - U_{ref}(t)]$. Such hybrids can be built using various techniques, from coaxial to stripline and waveguide, and in general the increase of noise and loss of signal in such a device is lower than for a ferrite microwave switch, as in a Dicke receiver. The two outputs of the hybrid are amplified by two independent radiometer receivers which share a common local oscillator, and the IF signals then are correlated (Fig. 5.15). If the input voltages to the correlator are

$$U_1 = \sqrt{G_1}\left[(U_A + U_{ref})/\sqrt{2} + U_{N1}\right],$$
$$U_2 = \sqrt{G_2}\left[(U_A - U_{ref})/\sqrt{2} + U_{N2}\right],$$

where U_A is the voltage from the antenna and U_{ref} that from the reference load, the instantaneous output voltage is then

$$U = \sqrt{G_1 G_2}\left[(U_A^2 - U_{ref}^2)/2 + U_{N1}(U_A - U_{ref})/\sqrt{2}\right.$$
$$\left. + U_{N2}(U_A + U_{ref})/\sqrt{2} + U_{N1}U_{N2}\right],$$

where U_{N1} is the noise voltage from amplifier 1, and U_{N2} that from amplifier 2.

Since the stochastic signals U_A, U_{ref}, U_{N1} and U_{N2} are all uncorrelated, the time average of all mixed products will average zero and only

$$\langle U \rangle = \tfrac{1}{2}\sqrt{G_1 G_2}\left[\langle U_A^2 \rangle - \langle U_{ref}^2 \rangle\right] \tag{5.28}$$

remains. Gain fluctuations therefore affect only this difference signal; the stability of the correlation receiver is therefore the same as that of a Dicke receiver. For the limiting sensitivity we obtain
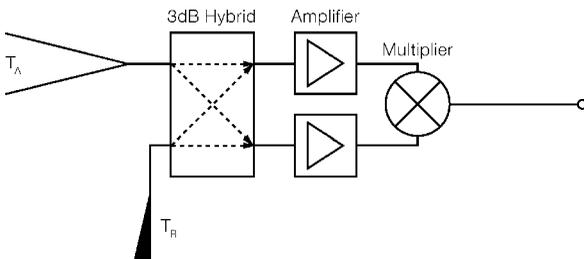
Fig. 5.14 A schematic of an analog correlation receiver. The operation of the "3 dB hybrid" is described in the text
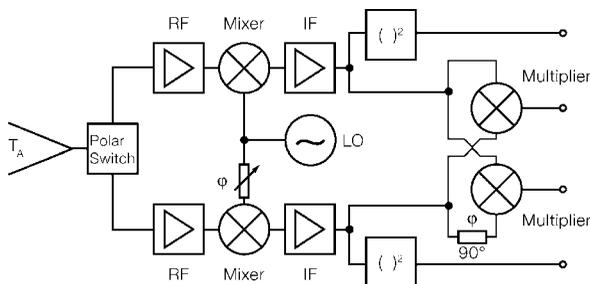
Fig. 5.15 An analog polarization receiver with four outputs, which are the four Stokes parameters. This is a Dicke system, with switching between the two senses of polarization. The "Polar Switch" could be a polarization transducer that allows a separation of the input into 2 senses of polarization

$$\frac{\Delta T}{T_{\text{sys}}} = \frac{\sqrt{2}}{\sqrt{\Delta \nu \, \tau}} . \tag{5.29}$$

A similar type of receiver can be used to measure the polarization of a wave field as defined by the Stokes parameters (3.52). The orthogonal linear polarization modes of a partially polarized wave field as collected by a circular horn are coupled by orthogonal probes into the two input ports of a correlation receiver. A fairly complex cross-correlation device then processes four output signals from the two IF signals:

- output z_1 is IF 1 detected by a square-law detector,
- output z_2 is IF 2 detected by a square-law detector,
- output z_3 is the correlation of IF 1 and IF 2,
- output z_4 is the correlation of IF 2 and IF 2 with a phase delay of $\pi/2$ in one of the channels.

Comparing these outputs with the definition of the Stokes parameters (3.52) we have

$$\begin{aligned}
I &= \text{const}(z_1 + z_2), \\
Q &= \text{const}(z_1 - z_2), \\
U &= 2\,\text{const}\,z_3, \\
V &= 2\,\text{const}\,z_4.
\end{aligned} \tag{5.30}$$

The output signals z_3 and z_4 come from the cross-correlation of both IF channels; they will therefore be fairly immune to amplification fluctuations. A polarimeter of the type described in (5.30) allows accurate measurements of U and V, that is, in wave fields with circular polarization. Usually the circular polarization of astronomical sources is exceedingly small, so one is more interested in measuring linear polarization. The polarimeter is easily converted for this purpose by including a $\lambda/4$ phase shifter in the waveguide section of the horn, so that the probes collect the left- and right-handed circular polarization components. If these are fed into the polarimeter, we now have

$$\begin{aligned} I &= \text{const}\,(z_1 + z_2)\,, \\ V &= \text{const}\,(z_1 - z_2)\,, \\ Q &= 2\,\text{const}\,z_3\,, \\ U &= 2\,\text{const}\,z_4\,, \end{aligned} \tag{5.31}$$

so that the linear polarization components Q and U are now derived by correlated outputs with the corresponding immunity to amplification fluctuations. This method is used with interferometer systems (see Chap. 9).

5.4.2 Spectrometers

Of the many different receiver back ends that have been designed for specialized purposes, spectrometers are probably the most widely used. Such spectrometers are designed to measure the power spectral density (PSD). These follow the principles shown in Fig. 4.1. Usually this is carried out in especially designed hardware, but recently there have been devices based on general purpose digital computers. For use with bolometers, one could use an analog Michelson or Fourier transform interferometer, or perhaps a Fabry-Perot system.

In designing spectrometers, emphasis is placed on the spectral information contained in the radiation field. To accomplish this the receivers must be single sideband and the frequency resolution Δv is usually small; perhaps in the kHz range, and the time stability must be high. For spectroscopy, SSB receivers are desirable.

If a resolution of Δv is to be achieved for the spectrometer, all those parts of the system that enter critically into the frequency response have to be maintained to better than $0.1\,\Delta v$. This applies in particular to the local oscillator; in this respect the same demand is set on each local oscillator frequency in a double or triple conversion super heterodyne receiver. If possible therefore, the local oscillator signal should be obtained from a master oscillator, such as a rubidium clock or a hydrogen maser by direct frequency multiplication. If this is difficult as e.g. for frequencies $> 10\,\text{GHz}$, frequency stabilization by applying phase lock schemes have been used. In all modern installations oscillator frequencies are computer controlled.

A frequency resolution in the kHz-range is required if narrow spectral features are to be resolved. The limiting sensitivity of the spectrometer is given by a slight generalization of Eq. (4.41):

$$\frac{\Delta T}{T_{\text{sys}}} = \frac{K}{\sqrt{\Delta v\,\tau}} \tag{5.32}$$

(a list of values of K for different receiver configurations is given in Table 4.3). The integration times, τ, needed to reach a given $\Delta T/T_{\text{sys}}$ can be quite long, since Δv is small. For this reason spectrometer back ends must have a very high stability, since *any* systematic errors will lead to fluctuations larger than that given by (5.32). We will only discuss the those spectrometer types that are used with heterodyne receivers. Recent descriptions of wide band spectrometers are to be found in Baker et al. eds. (2007).

5.4.2.1   Multichannel Filter Spectrometers

The time needed to measure the power spectrum for a given celestial position can be reduced by a factor n if the IF section with the filters defining the bandwidth Δv, the square-law detectors and the integrators are built not merely once, but n times. Then these form n separate channels that simultaneously measure different (usually adjacent) parts of the spectrum. Filter spectrometers are analog devices. In Fig. 4.1 these devices transform the input voltage (in the upper left of the diagram) to the PSD (in the lower right) by a path across the top and then down on the right. The technical details of how such a multichannel spectrometer is built may differ from one instrument to another, but experience has shown that the following design aims are essential:

1) The shape of the bandpass $G_i(v)$ for the individual channels must be identical. It is not sufficient that only the bandwidths Δv_i are the same. In interferometer systems, the phase characteristics of the filters must also be identical.
2) The square-law detectors for the channels must have identical characteristics. This refers both to the mean output power level and to any deviations from an ideal transfer characteristic.
3) Thermal drifts of the channels should be as identical as is technically feasible.

Goals 1 and 2 are determined by the need to detect weak spectral features in a few spectral channels. Condition 3 must be met if the long term behavior of the different channels is to be the same. For goal 3 the stability requirements for the individual channels are determined by the condition that stability times from (4.49) are $\gg$ the time interval between the measurement of signal and reference.

The fundamental limitation of filter spectrometers is that these are analog devices. As such these are sensitive to changes in the ambient temperature, as well as other environmental factors. Another limitation is the lack of flexibility: varying the frequency resolution, Δv is quite complex. The simplest solution is to build a number of separate filter banks. In the millimeter range, one usually has filter banks of 256 or 512 contiguous filters, each of width 100 kHz, 250 kHz and 1 MHz. For these reasons, alternatives, such as autocorrelation spectrometers, have been employed.

5.4.3   Fourier and Autocorrelation Spectrometers

The PSD can be determined using the *Wiener-Khinchin theorem*. Referring to Fig. 4.1, there are two paths to obtain the PSD. In the lower right side of the diagram, from $v(t)$, in the upper left of the diagram. These are presented in the next two subsections.

5.4.3.1   Fourier Spectrometers

One method is to Fourier Transform (FT) the input, $v(t)$, to obtain $v(v)$ and then square $v(v)$ to obtain the PSD. In Fig. 4.1 this is equivalent to moving across the top of the diagram, from left to right, then down on the right to obtain the PSD

(4.13). From the Nyquist theorem (see Fig. 4.4 and discussion), it is necessary to sample at a rate equal to twice the bandwidth. Then the FT's can be carried out using Fast Fourier Transform algorithms (FFT). FT spectrometers using this approach have been used at the Nobeyama Radio Observatory with notable success. These are referred to as "FX" autocorrelators. Recent developments at the Jodrell Bank Observatory have led to the building of COBRA (Coherent Baseband Receiver for Astronomy). This system uses high speed computers and sophisticated software in a flexible system which can be used as a spectrometer with a 100 MHz bandwidth, and also as a pulsar de-disperser (see Sect. 5.4.4). A filter limits the IF input frequency band. When used as a spectrometer, the analog input $v(t)$ is mixed to the video band (starting close to 0 Hz), digitized with 8 bit A/D converters, sampled at 200 MHz, the Nyquist rate, for a 100 MHz bandwidth transformed to $v(v)$ using FFT's, then squared to produce a detected signal and averaged.

5.4.3.2   Autocorrelation and Cross Correlation Spectrometers

The input $v(t)$ is correlated, and this result is FT'ed to obtain the PSD. In Fig. 4.1 this is the path down the left side and then across the bottom. The autocorrelation function $R(\tau)$ function is evaluated in hardware, then the FT is performed in a general purpose digital computer. $R(\tau)$ is calculated by a multiplication of the current sample with a sample delayed by a time τ. The first digital autocorrelation spectrometer used in astronomy was designed and built by S. Weinreb (1963). A description of the instrument and its theory is given in his thesis. Autocorrelation can also be carried out with the help of analog devices as shown in Sect. 5.4.1 on correlation receivers, using a series of cable delay lines. A recent development is WASP (Wideband Autocorrelation Spectrometer), a broadband autocorrelation *analog* spectrometer. Presently WASP has a total bandwidth of 3.6 GHz in which 128 channels provide a frequency resolution of 33 MHz. WASP has been used to measure extragalactic carbon monoxide rotational transitions.

   For narrower bandwidths, digital techniques offer more stability and flexibility. In the following, we describe the most used type of autorcorrelation spectrometer, the "XF" digital autocorrelator. XF processing is shown in Fig. 5.16. The hardware of an XF autocorrelator spectrometer shifts the digitized and sampled input at the Nyquist frequency into a shift register which holds each delay. By comparing the shift register content delayed by $\Delta\tau$ steps with the current sample, the contribution to the counters is then proportional to $R_y(\tau)$. Dividing $R_y(\tau)$ by $R_y(0)$, normalized values for the ACF will be obtained.

   The two significant advantages of digital spectrometers are: (1) flexibility and (2) a noise behavior that follows $1/\sqrt{t}$ after many hours of integration. The flexibility allows one to select many different frequency resolutions and bandwidths or even to employ a number of different spectrometers, each with different bandwidths, simultaneously. The second advantage follows directly from their digital nature. Once the signal is digitized, it is only mathematics. Tests on astronomical sources have shown that the noise follows a $1/\sqrt{Bt}$ behavior for integration times $>100\,$h; in these aspects, analog spectrometers are more limited.
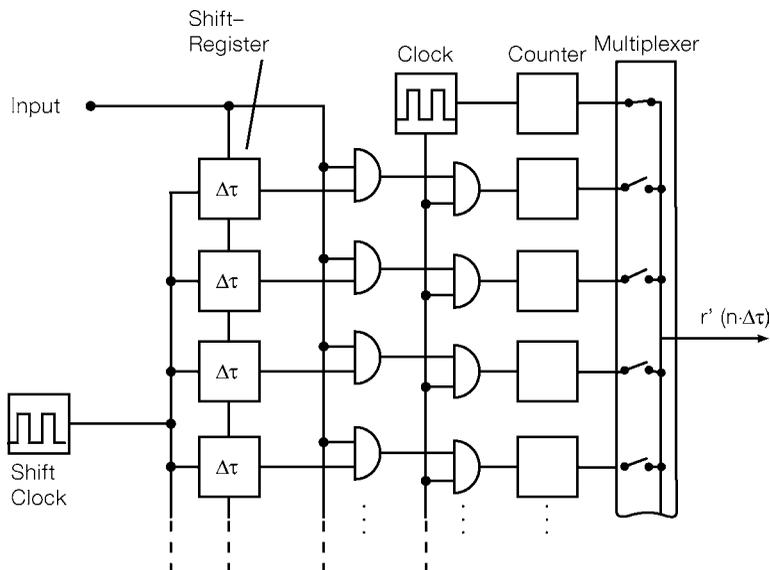
Fig. 5.16 Schematic showing the essential functional blocks of an XF autocorrelation spectrometer

Corrections for one-bit ("hard clipping") quantization can be expressed in a closed form; the details are presented in Appendix D. Corrections for 3 level and 4 level (2 bit) digitization (see Fig. 4.3 for a sketch of such a 3 level clipping scheme) follow similar procedures, but 3 and 4 level corrections cannot be expressed in a closed form. The PSD of the received signal is then calculated following the Wiener-Khinchin theorem (Eq. 4.14) by computing the FT of the measured autocorrelation function. Note that the limits of the integral in (4.14) extend to $\pm\infty$; in an actual instrument, however, $R(\tau)$ can be measured only up to a maximum delay τ_m. The measured ACF $R(\tau)$ can thus be considered to be the product of two functions: the true ACF $R(\tau)$ and a function describing the lag window

$$w(\tau) = \begin{cases} 1 \text{ for } |\tau| \leqq \tau_m \\ 0 \text{ for otherwise} \end{cases}. \tag{5.33}$$

The convolution theorem the measured PSD $\tilde{S}(v)$ is the convolution of the true PSD $S(v)$ and a filter with the frequency response

$$\tilde{S} = S(v) \otimes W(v). \tag{5.34}$$

so that

$$W(v) = 2\,\tau_m \,\text{sinc}(2\,\pi\,v\,\tau_m), \tag{5.35}$$

The response $W(v)$ determines the resolution of the autocorrelation spectrometer. If we define the frequency resolution of the spectrometer by the half width of (5.35)

we find that

$$\Delta v = \frac{0.605}{\tau_m} \, . \tag{5.36}$$

If the spectral region which is analyzed by the N channel spectrometer has the total bandwidth $N\Delta v$, the interval for the stochastic time series $x(t)$ must be $\Delta \tau = 1/2\Delta v$ according to the sampling theorem. Provided that the autocorrelator has N delay steps, multipliers and counters, then

$$\tau_m = N/2\Delta v \, ,$$

resulting in a frequency resolution

$$\Delta v = 1.21 \frac{\Delta v}{N_0} \tag{5.37}$$

or

$$\Delta v = \frac{0.605}{N\Delta \tau} \, . \tag{5.38}$$

For the autocorrelator discussed (N fixed), the frequency resolution can be changed simply by changing the sampling time step $\Delta \tau$, i.e. by changing the clock frequency. In order to satisfy the sampling theorem, the total bandwidth accepted by the spectrometer has to be simultaneously adjusted also, since $\Delta v = 1/2\Delta \tau$.

Using the lag window (5.33) results in a filter function (5.35) with high side-lobes. These will decline only slowly, since these vary as $1/2\tau_m$. If narrow, strong features occur in the spectrum $S(v), \tilde{S}(v)$ will be distorted. The sidelobes can be reduced by using a lag window different from (5.33). The window first introduced by J. von Hann ("hanning") is given by

$$w_H(\tau) = \begin{cases} \cos^2 \left(\dfrac{\pi \tau}{2\tau_m} \right) & \text{for } |\tau| \leqq \tau_m \, , \\ 0 & \text{otherwise} \, . \end{cases} \tag{5.39}$$

The corresponding filter frequency response is

$$W_H = \tau_m \left[\text{sinc}(2v\pi\tau_m) + \frac{2v\tau_m}{\pi[1-(2v\pi\tau_m)^2]} \sin(2\pi v\tau_m) \right] \, . \tag{5.40}$$

The frequency resolution corresponding to this lag window is

$$\Delta v = \frac{1}{\tau_m} = \frac{2\Delta v}{N_0} = \frac{1}{N_0\Delta\tau} \, ; \tag{5.41}$$

that is, the frequency resolution is 40 % less than using the window (5.33). The first side lobe, however, is now at only 2.6 % of the peak, while for (5.33) it is 22 %.

Multiplying the time series $x(t)$ with the lag window $w(\tau)$ is equivalent to convolving $S(v)$ with $W(v)$. Then introducing the lag window (5.39) can be done even

after performing the FT of $R(\tau)$ by convolving with (5.40). For a spectrum $S(\nu)$ given at equidistant frequencies with $\Delta \nu = 1/2\tau_{\mathrm{m}}$, this has the effect of forming a new spectrum consisting of the running average of the original spectrum with the weights $1/4, 1/2, 1/4$. This is called *hanning*. In the spectral realm this operation is equivalent to introducing the lag window (5.39); for lines of width comparable to the spectrometer resolution it is good practice to smooth spectra obtained by an autocorrelation spectrometer in this way.

In order to obtain the PSD $S(\nu)$ from the measured $\tilde{R}(\tau)$ an FT has to be performed. Since $\tilde{R}(\tau)$ is obtained for a series of equidistant power-of-two or even a factorable set of delays, τ_i, this transformation is best done by the Fast Fourier transform (FFT) algorithm of Cooley and Tukey. The use of the FFT considerably speeds up computations.

A serious drawback of digital auto and cross correlation spectrometers had been limited bandwidths. Previously 50–100 MHz had been the maximum possible bandwidth. This was determined by the requirement to meet Nyquist sampling rate, so that the A/D converters, samplers, shift registers and multipliers would have to run at a rate equal to twice the bandwidth. The speed of the electronic circuits was limited. However, advances in digital technology in recent years have allowed the construction of autocorrelation spectrometers with several 1000 channels covering bandwidths of several 100 MHz. One can obtain larger analyzing bandwidths by two methods. First, one can position a number, N, of individual autocorrelators side-by-side. Each would have a fairly small bandwidth, but the total analyzing bandwidth would be N times the bandwidth of each individual autocorrelator. In this arrangement, the first part of the system is analog and the second part is digital. Thus, this type of system is referred to as a *hybrid* system. In order to prevent unequal drifts in the analog part of the system, so-called *platforming* of the spectral shape, the connections between digital and analog parts of the system are periodically exchanged by a control computer. A second method to increase the bandwidth of autocorrelation spectrometers makes use of a single analog part, with a sampler which takes data at a rate $\Delta t = 1/2B$, but this output is then fed into M different shift register-correlator digital sections. In each, the autocorrelation analysis can be carried out at a rate which is M times slower.

Another improvement is the use of *recycling* auto and cross correlators. These spectrometers have the property that the product of bandwidth, B times the number of channels, N, is a constant. Basically, this type of system functions by having the digital part running at a high clock rate, while the data are sampled at a much slower rate. Then after the sample reaches the Nth shift register (Fig. 5.16) it is reinserted into the first register and another set of delays are correlated with the current sample. This leads to a higher number of channels and thus higher resolution. Such a system has the advantage of high-frequency resolution, but is limited in bandwidth. This has the greatest advantage for longer wavelength observations. Both of these developments have tended to make the use of digital spectrometers more widespread. This trend is likely to continue.

Autocorrelation systems are used in single telescopes, and make use of the symmetric nature of the ACF (4.11). Thus, the number of delays gives the number of

spectral channels. For cross-correlation, the current and delayed samples refer to different inputs. As will be shown in Chap. 9, cross-correlation systems are used in interferometers. This is a generalization of (4.11). In the simplest case of a two-element interferometer, the output is *not* symmetric about zero time delay, but can be expressed in terms of amplitude and phase at each frequency, where both the phase and intensity of the line signal are unknown. Thus, for interferometry the zero delay of the ACF is placed in channel $N/2$ and is in general asymmetric. The number of delays, N, allows the determination of $N/2$ spectral intensities, and $N/2$ phases. The cross-correlation hardware can employ either an XF or a FX correlator. The FX correlator has the advantage that the time delay is just a phase shift, so can be introduced more simply.

5.4.3.3   Acousto-Optical Spectrometers

Since the discovery of molecular line radiation in the mm wavelength range there has been a need for spectrometers with bandwidths of several hundred MHz. At 100 GHz, a velocity range of $300\,\mathrm{km\,s^{-1}}$ corresponds to 100 MHz, while the narrowest line widths observed correspond to 30 kHz. Autocorrelation spectrometers can reach such bandwidths only if complicated methods are used. Thus multichannel filter spectrometers are more common in the mm and sub-mm ranges. As pointed out in Sect. 5.4.2.1 these are rather inflexible and often have differential drift and calibration problems, and thus there was a need for a wide band system with reasonable stability that could be used to obtain different frequency resolutions and bandwidths easily. It now seems that acousto-optical spectrometers (AOS) can meet most of these requirements.

The AOS makes use of the diffraction of light by ultrasonic waves. This effect had been predicted in 1921 by Brillouin: Sound waves cause periodic density variations in the medium through which it passes. These density variations in turn cause variations in the bulk constants ε and n of the medium, so that a plane electromagnetic wave passing through this medium will be affected. In "Principles of Optics" by Born and Wolf (1965, p. 596*ff*) it is shown that using Maxwell's equations such a medium will cause a plane monochromatic electromagnetic wave (wave number $k = 2\pi/\lambda$ and frequency $\omega = 2\pi\nu$) to be dispersed. The emergent field can be described by the superposition of a sequence of waves with the wave number $k\sin\theta + lK$ and frequency $\omega + l\Omega$ where K and Ω are wave number and frequency of the sound wave, θ the angle between electric and acoustic wave, and l an index $l = 0, \pm1, \ldots$. The amplitudes of the different emerging waves can then be determined by recursion relations.

For a proper understanding of this mechanism, a series expansion of Maxwell's equations or the equivalent integral equation method [Born and Wolf (1965), Sect. 12.2] must be used. We will use a more intuitive approach based on an analogy. The plane periodic variations of the index of refraction n can be considered to form a 3-dimensional grating that causes diffraction of the electromagnetic wave.

Fig. 5.17 Diffraction of light
by an acoustic wave. Λ is the
wavelength of the acoustic
wave, and the *arrows* show
the path of the light waves
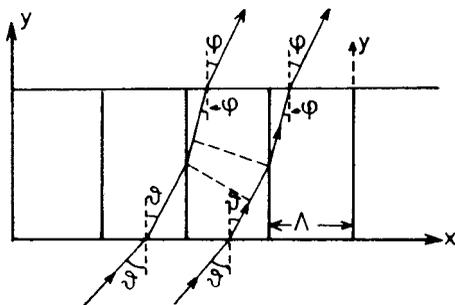
Let a monochromatic light wave of angular frequency ω and wavelength λ make
an angle θ with the y axis, and let the angle of the diffracted ray be ϕ (Fig. 5.17).
Since the velocity v of the compression wave is always much smaller than the ve-
locity of light we can consider the periodic structure in the matter to be stationary.
The permitted angles ϕ are then determined by the condition that the optical path
difference from neighboring acoustic wave planes should be integral multiples of λ.
With a spacing Λ between adjacent acoustic wave crests, thus (see Fig. 5.17)

$$\Lambda \left(\sin\phi - \sin\theta \right) = l\lambda, \qquad l = 0, \pm 1, \pm 2 \ldots . \tag{5.42}$$

This is the Bragg condition met in the diffraction of X-rays in crystals; for this
reason, the device is referred to as a Bragg cell. Because of the interactions, an
acoustic wave affects the index of refraction. This produces a traveling wave, which
can be detected by illuminating the cell with a monochromatic light beam.

The practical problem is to find a transparent material with a low sound velocity
so that the acoustic wavelength Λ is small for a given sound frequency v_s. A second
problem is how to couple the transducer that converts the electric signal v_s into an
acoustic wave with a reasonably constant conversion factor over a wide bandwidth.
And finally, an absorber of acoustic waves has to be provided so that no standing
wave pattern develops, since such a pattern will always have resonances and there-
fore is not suitable for broadband applications.

If the light beam is provided by a monochromatic laser and if the acoustic wave
intensity is small (in order to avoid problems with saturation) the intensity of the
diffracted light is proportional to the acoustic power. In the linear range, different
acoustic frequencies v_s can be superposed, resulting in different diffracted angles
ϕ_s. Differentiating (5.42) and substituting $\Lambda_s v_s = v_c$ we obtain

$$\cos\phi \, \delta\theta = \frac{l\lambda}{v_c} \Delta v_s \qquad . \tag{5.43}$$

The block diagram of an AOS is shown in Fig. 5.18. The light source is a laser;
the beam is expanded to match the aperture of the Bragg cell and the distribution of
the light intensity in the focal plane is detected by a CCD array. After an integration
time of some milliseconds, the counts recorded by the photo diodes are sampled,
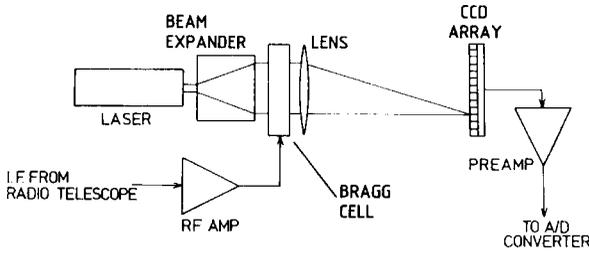
Fig. 5.18 A block diagram of an acousto optical spectrometer (AOS)

read out and transferred to a computer where the final integrations are carried out. The maximum number of channels that can be resolved by such an instrument can be determined by considering how well the wave front of light emerging from a monochromatic grating can be determined. This uncertainty is

$$\Delta \theta \cong \lambda / L \tag{5.44}$$

if L is the aperture of the Bragg cell. Therefore a total bandwidth $\Delta \nu$ can at most be resolved into

$$N_0 = \frac{\delta \theta}{\Delta \theta} = \frac{L}{v_c \cos \phi} \Delta \nu = \frac{\tau_c}{\cos \phi} \Delta \nu \tag{5.45}$$

channels. $\Delta \nu$ is limited by the condition that adjacent orders of the diffraction ($\Delta l = \pm 1$) should not overlap. τ_c is the time it takes the acoustic wave to pass through the Bragg cell: $\tau_c = L/v_c$. A typical value for the total bandwidth possible for a single Bragg cell is presently 1–2 GHz. The dynamic range is the ratio between the largest and the smallest signal that can be measured with any certainty. There are several effects that put limits on the dynamic range, the stability and consequently on the achievable sensitivity of the AOS. The first limit is determined by nonlinearities of the response, the second limit by the dark current of the CCD array and other internal noise sources. In order to reach a large dynamic range, the response of the Bragg cell to the radio frequency input must be linear, but a linear response means low RF power, otherwise intermodulation effects occur in the deflector.

Often, practical experience with AOSs shows that another effect, *laser speckles*, is present. This adds considerable instrumental noise and shows strong, narrow spikes which vary with channel position and time. Any variation of the optical path length due to mechanical or thermal fluctuations in the order of fractions of the laser wavelength causes spatial and amplitude fluctuations, giving rise to these speckles. As a consequence both the dynamic range and the noise performance of the spectrometer can be degraded by more than one order of magnitude. The main cause of these spikes is light scattered from the undeflected laser beam that is measured by the photo detectors. One way to reduce the level of the speckles is to confine the laser beam so that only a deflected signal reaches the CCD array. Another innovation makes use of a polarized laser beam. If the deflector is operating in the acoustical shear mode, there is a change in the polarization of the deflected light, while

the polarization of the undeflected and scattered light is not changed. Therefore a polarization filter in front of the CCD array can reduce the level of the scattered light by more than 20 dB. Today the integration time for an AOS is more than 100 s. After this time, a reference measurement is needed. By alternately measuring signal and reference positions, one may be able to carry out measurements for much longer periods. Due to their compactness and simplicity, AOS's have been used in a number of astronomcal satellites.

5.4.3.4   Chirp Transform Spectrometers

For not too wide bandwidths, an alternative to the AOS is the chirp transform spectrometer, CTS. The principle of the CTS was first given by Darlington (1964). The CTS makes use of radio technology only, in contrast to the AOS that makes use of both radio and optical technologies. As in an AOS, the radio signal is converted into an acoustic wave in a delay line, but one that intentionally has a strong dispersion. Therefore an input pulse is converted into an output signal with a sweeping frequency. This property can be used to form a FT-ing device using only electronic means. The principles of CTS operation follow from Fourier transforms (see Appendix B). These relate the frequency spectrum $F(v)$ to the time behavior $f(t)$ by

$$F(v) = \int_{-\infty}^{\infty} f(t) e^{-i 2\pi v t} dt . \tag{5.46}$$

Setting $v = \mu \tau$, we have

$$F(\mu \tau) = \int_{-\infty}^{\infty} f(t) e^{-i 2\pi \mu \tau t} dt . \tag{5.47}$$

Using the identity

$$2t\tau = t^2 + \tau^2 - (t - \tau)^2 , \tag{5.48}$$

we obtain

$$F(v) = \int_{-\infty}^{\infty} f(t) e^{-i \pi \mu (t^2 + \tau^2 - [t-\tau]^2)} dt . \tag{5.49}$$

Factoring terms out of the integral, we have the expression

$$F(v) = e^{-i \pi \mu \tau^2} \int_{-\infty}^{\infty} \left[f(t) e^{-i \pi \mu t^2} \right] \left[e^{i \pi (t-\tau)^2} \right] dt . \tag{5.50}$$

The term outside the integral, forming a constant phase shift, is not relevant for radio-astronomical applications. There are two terms inside the integral. The first square bracket contains the input signal, $f(t)$, modulated by a term $e^{-i \pi \mu t^2}$. This term is referred to as the chirp and the whole unit is called the compressor. The multiplication of the input signal with the chirp is carried out by using a mixer and it results in a conversion of a stationary frequency dependent signal into a time-varying signal. The $e^{-i \pi \mu t^2}$ waveform is produced in a dispersive delay line.

The second square bracket is a convolution. This is produced after mixing by passing the signal through a matched filter and it is called the expander. A single CTS has a duty cycle of about 50% which results from the ratio of the bandwidths of compressor to expander. In practical systems, two or more such networks are combined to produce a device with an effective 100% duty cycle. The operations are carried out with analog electronics. The final output is digitized so that summations can be made in a digital computer.

A prototype CTS was successfully used for astronomical measurements (Hartogh and Oslerschek 1998). The total bandwidth of this system was 178 MHz, with 4000 spectral points and a channel resolution of 44 kHz.

5.4.4   Pulsar Back Ends

Back ends for pulsar observations differ from others used in radio astronomy because the pulsar signals change rapidly with time, although in a strictly periodic fashion. Pulsars were first detected in 1967 when Hewish et al. (1968) employed a receiving system which could respond to short time intensity variations (the system was designed to measure the scintillation of small diameter radio sources which passed behind the sun). Modern pulsar back ends are optimized to measure the properties of pulsar radiation. Therefore we will first give a short summary of the relevant properties of this radiation. Pulsar radiation consists of short bursts of radiation which repeat with remarkable precision; pulsars have periods that vary from a few milliseconds to more than 4 s. The pulse width is usually only a small fraction (of the order 10^{-3}) of the total pulse period; the amplitude of the pulse can vary considerably from one pulse to the next. Some pulsars show strong linear polarization with systematic variations across the pulse; this is caused by the pulsar emission. For all pulsars the radiation arrival times are strongly frequency dependent; this s caused by propagation effects of the electromagnetic waves in the intervening interstellar medium.

Two commonly employed pulsar measurement processes are the determination of average pulse shapes and searches for periodic pulses with unknown periods.

In order to be able to measure the shape of the pulsar radiation, rather short receiver time constants, τ, must be used in the back ends. Because of the frequency dependence of the pulse arrival time, the IF bandwidth $\Delta \nu$ must be kept small. As a consequence the Dicke formula (4.41) the $\Delta T / T_{\mathrm{sys}}$ is fairly large. In order to suppress receiver noise, use is made of the strict pulse repetition rate; that is many individual pulses are added together to obtain an average pulse profile. The main part of most pulsar back ends therefore consists of a multichannel filter bank and signal averager, in which the pulsar signal is sampled, digitized and averaged. The fast sampling is controlled by a clock synchronized with the pulsar repetition rate. Then mean profiles can be accumulated over hours, and thus good signal-to-noise ratios can be obtained even if narrow bandwidths are used.

An alternative to a multichannel backend for Pulsars is COBRA developed at Jodrell Bank. With COBRA, the voltage from the receiver is sampled, digitized and Fourier transformed. Then this is convolved with a chirp function, to remove the frequency dependent delay in the interstellar medium, and FT to return to the time domain. The profile is folded to produce the grand average. With this coherent de-dispersion, one can analyze a 100 MHz bandwidth in all 4 Stokes parameters. Such a process is equivalent to a filter bank with a very large number of very narrow channels, so it allows a more accurate measurement of the details of the pulse shape. A system such as COBRA is very useful for pulsar timing experiments, since the instrumental broadening of the measured pulse shapes is less than those measured using filter banks.

5.4.4.1   Pulse Dispersion and Dispersion Removal

The dispersion of the pulse arrival time with frequency has a profound influence on the response of a receiver on such a signal. A thorough understanding of this requires a complex analysis of the transfer properties of the receiver which is beyond the scope of this book; a short version can be found in the review article by Phinney and Kulkarni (1994) or the article by Backer (1988).

Pulse dispersion in the interstellar medium can be described by a transfer function in the time or frequency domain. A filter can be constructed that removes this dispersion for a limited frequency range either by hardware or software techniques. This *predetection removal* is necessary when the bandwidth of the receiver must be widened in order to detect short time intensity variations in single pulses.

For a more intuitive description of the effect of a dispersed pulse on the receiver output, the effect on the time resolution of the receiver is of importance. Suppose the interstellar medium has a dispersion measure DM according to (2.85). The pulse then is received with a frequency sweep rate α, resulting in a pulse duration of at least

$$\frac{t_{\mathrm{s}}}{\mathrm{s}} = \frac{B}{\alpha} = 0.830 \times 10^4 \, \frac{\mathrm{DM}}{\mathrm{cm}^{-3}\,\mathrm{pc}} \, \frac{B/\mathrm{MHz}}{(\nu/\mathrm{MHz})^3} \,. \tag{5.51}$$

Therefore, if no corrections are made, a small bandwidth must be used when a high time resolution is required with a corresponding loss of amplitude resolution.

This situation can be improved by dividing the front end bandwidth into several contiguous bands that are detected separately and then appropriately combined after each signal has been delayed by the time given by (2.84). Conceptually the simplest such system is a filter bank. Such *postdetection dispersion removers* are used at most observatories where pulsar observations are carried out on a routine basis.

Comparing predetection and postdetection dispersion removal techniques, the advantage of the first is the high time resolution which can be obtained even at low frequencies and high dispersion measures. Hardware filters needed for this are, however, useful only for a single dispersion measure and it is difficult to readjust them for a different DM. Software filtering cannot be done in real time, so that usually predetection dispersion removal is only applied when the highest time resolution

is needed. On the other hand, postdetection dispersion removal can be done in real time, that is on-line, as the data is taken. The dispersion remover is easily reconfigured for different DM, and therefore such an approach is usually used on a routine basis.

5.4.4.2 Pulsar Searches

The first pulsars were detected by noting that there were periodic spikes on a chart record; however weaker and more sporadic pulsars cannot be seen on individual chart records. These have to be found by making use of the distinct signature of pulsar radiation: that is, their regular pulse period and dispersion in frequency of the pulse arrival time. The optimum detection of pulsar radiation in the presence of Gaussian noise is obtained by convolution of the received signal with a matching filter whose impulse response is given by rectangular functions spaced at the assumed period. This is referred to as *rail filtering*, but the use of this method requires a known period. In pulsar searches, the signal must be convolved with a whole series of *rail filters* covering a specified range of periods. Usually this is done using software techniques. The signal is sampled at regular time intervals and stored digitally. There exist two methods for the analysis of the presence of periodic signals in data: a *fast folding algorithm* (FFA) and the *fast Fourier transform* (FFT) method. Both can investigate the data in real time provided fast computers are used. Usually some kind of dispersion removal is also used, so that surveys will be most sensitive to pulsars within a certain range of dispersions.

Problems

1. What is the minimum noise possible with a coherent receiver operating at 115 GHz? At 1000 GHz, at 10^{14} Hz?

2. Coherent and incohent receivers are fundamentally different. However one can determine the equivalent noise temperature of a coherent receiver T_n which corresponds to the NEP of a bolometer. This can be determined by using the relation

$$\text{NEP} = 2kT_n \sqrt{\Delta v} \, .$$

For $\Delta v = 50$ GHz, determine T_n for NEP $= 10^{-16}$ W Hz$^{-1/2}$. A bolometer receiver system can detect a 1 mK source in 60 s at the 3σ level. The bandwidth is 100 GHz. How long must one integrate to reach this RMS noise level with a coherent receiver with a noise temperature of 50 K, and bandwidth 2 GHz?

3. In the millimeter and sub-millimeter range, the y factor (see Fig. 4.10) usually represents a double-sideband system response. For spectral lines, one wants the

single-sideband receiver noise temperature. If the sideband gains are equal, what is the relation of the y factor for a single- and double-sideband system?

4. The definition of a *decibel*, db, is

$$\mathrm{db} = 10\log\left(\frac{P_{\mathrm{output}}}{P_{\mathrm{input}}}\right).$$

If a 30 db amplifier with a noise temperature of 4 K is followed by a mixer with a noise temperature of 100 K, what is the percentage contribution of the mixer to the noise temperature of the total if

5. (a) In Fig. 5.5, the upper sideband (USB) frequency is 115 GHz, and the lower sideband frequency is 107 GHz. What is the intermediate frequency? What is the local Oscillator (LO) frequency?
(b) When observing with a double-sideband coherent receiver, an astronomical spectral line might enter from either upper or lower sideband. To distinguish between these two possibilities, one uses the following procedure. To decide whether the line is actually in the upper or lower sideband, the observer increases the local oscillator frequency by 100 kHz. The signal moves to *lower* frequency. Is the spectral line from the upper or lower sideband?

6. The same situation as in Problem 6, but after the first mixer is a second mixer with an LO frequency which is *higher* than the intermediate frequency of the first mixer. The spectral line is known to be in the upper sideband. To eliminate unwanted spectral lines, someone tells you to move the LO higher frequencies in steps of 100 kHz, and at the same time, move LO2 to lower frequencies by the same step. After repeating this procedure for 10 steps of 100 kHz, the result is added. Will this procedure eliminate spectral lines in the lower sideband? If the unwanted lower sideband spectral line has a width of 100 kHz, by how much is this line reduced in intensity?

7. In Fig. 5.6, is the schematic of a *single-sideband mixer*. In such a system, the image and signal bands are separated in the output if the input is $f(t) = \cos\omega_s t$. Use an analysis for this input signal to show that such a mixer is feasible. Repeat for $f(t) = \sin\omega_s t$.

8. The input power of a receiver can be 10^{-16} W, while the power at the output of a receiver must be about a milli Watt. What must be the power amplification of such a receiver? Express this in decibels. Suppose the gain stability of this receiver is 10^{-3} over 30 s. What is the change in the output power? Suppose that the system noise is 100 K and the bandwidth is 1 GHz. This is used to measure a source with a peak temperature of 0.01 K. What is the ratio of the signal intensity to that of gain fluctuations? The fluctuations can be reduced by periodic comparisons with a reference source; how often should one switch the receiver between the signal and a reference to stabilize the output power?

9. Laboratory measurements frequently make use of a data-taking method which involves a modulated signal. The output is then measured synchronously with the

modulation rate in both frequency and phase. We can measure a weak input signal, $S = T(\text{signal})e^{-\tau}$, in the presence of noise, $T(\text{cable})(1-e^{-\tau})$, by modulating the signal with a known frequency, f_1. The output is superimposed on noise background. What is the noise in the switched output? What is the signal-to-noise ratio? How will the signal-to-noise ratio change with time if only random noise is present?

10. If the bandwidth of a receiver is 500 MHz, how long must one integrate to reach an RMS noise which is 0.1% of the system noise with a total power system? Repeat for a Dicke switched system, and for a correlation system. Now assume that the receiver system has an instability described by (4.48). For a time dependence $(\Delta G/G)^2 = \gamma_0 + \gamma_1 \tau$ we take $\gamma_0 = 0$, $\gamma_1 = 10^{-2}$ and $K = 2$. On what time scale will the gain instabilities dominate uncertainties caused by receiver noise? If one wants to have the noise decrease as $1/\sqrt{t}$, what is the lowest frequency at which one must switch the input signal against a comparison?

11. At 234 MHz, the *minimum* sky noise is $\sim$100 K. For use as a first stage amplifier at 234 MHz should you buy an expensive receiver for use at a sky frequency of 234 MHz which has a noise temperature of 10 K, if a similar receiver has a noise temperature of 50 K but costs 10% of the price of the lower-noise receiver? Explain your decision by considering observational facts.

12. An all-sky continuum survey covering 41 252 square degrees, is carried out with a $40'$ beam at 234 MHz. Three spatial samples are taken for each beamwidth. These samples are used to image the sky at 234 MHz.
(a) Compare the sampling procedure to the Nyquist sampling rate using the example of the sampling of sine or cosine waves. What is the total number of samples?
(b) Next, assume that the sky noise dominates the receiver noise. If the bandwidth B is 10 MHz and the integration time is 10 s per position, what is the RMS noise as a fraction of T_{source}, the sky noise? How many data points are needed to completely characterize the resulting map? If one needs 20 s of time for measuring each position, how long will this survey require?
(c) Repeat this estimate for a survey at 5000 MHz carried out with a $3'$ beam, for a receiver with noise temperature 50 K, 500 MHz bandwidth, 10 s integration per point. Note that the sky background contributes only a small amount of the receiver noise at 5 GHz. How much observing time is needed for this survey?

Chapter 6
Fundamentals of Antenna Theory

6.1 Electromagnetic Potentials

Analytic solutions of Maxwell's equations (2.4, 2.5, 2.6, 2.7) are rather simple for plane harmonic waves, but are very complex for realistic configurations. As a simplification of the mathematics, we introduce new functions, the electrodynamic potentials Φ and A , which can be determined from given current and charge densities J and ϱ. These potentials give both E and B in a straightforward way. In electromagnetic theory, potential functions were first used by Green 1828, but this was noted by the scientific community only in 1846, when Lord Kelvin directed attention to this paper. Independently, one year before Franz Neumann in Königsberg had successfully used this method.

According to Maxwell's equation (2.5), we always have $\nabla \cdot B = 0$. From Stokes' theorem (A 22), we can write

$$\boxed{B = \nabla \times A} \quad , \tag{6.1}$$

so that (2.6) becomes

$$\nabla \times \left(E + \frac{1}{c}\dot{A} \right) = 0$$

where the order of time and spatial differentiation have been interchanged. But Gauss' theorem states that a vector whose curl vanishes can always be expressed as the gradient of a scalar, so that

$$E + \frac{1}{c}\dot{A} = -\nabla\Phi$$

or

$$\boxed{E = -\nabla\Phi - \frac{1}{c}\dot{A}} \quad . \tag{6.2}$$

Both B and E can be expressed in terms of A and Φ. If these expressions are to be useful, we require that the resulting fields B and E should obey Maxwell's

equations. To determine which additional restrictions this imposes on A and Φ, we introduce (6.1) and (6.2) into Maxwell's equations. For simplicity, we adopt free space conditions and set ε, μ and σ equal to 1.

From (2.7) we then obtain

$$\nabla \times (\nabla \times A) + \frac{1}{c} \frac{\partial}{\partial t} \left[\nabla \Phi + \frac{1}{c} \dot{A} \right] = \frac{4\pi}{c} J,$$

$$\nabla (\nabla \cdot A) - \nabla^2 A + \frac{1}{c} \frac{\partial}{\partial t} \left[\nabla \Phi + \frac{1}{c} \dot{A} \right] = \frac{4\pi}{c} J,$$

$$\nabla^2 A - \frac{1}{c^2} \ddot{A} - \nabla \left(\nabla \cdot A + \frac{1}{c} \dot{\Phi} \right) = -\frac{4\pi}{c} J. \tag{6.3}$$

Using (6.2) the remaining Eq. (2.4) gives

$$\nabla \cdot \nabla \Phi + \frac{1}{c} \nabla \cdot \dot{A} = -4\pi \varrho$$

or

$$\nabla^2 \Phi - \frac{1}{c^2} \ddot{\Phi} + \frac{1}{c} \frac{\partial}{\partial t} \left[\nabla \cdot A + \frac{1}{c} \dot{\Phi} \right] = -4\pi \varrho. \tag{6.4}$$

Neither A nor Φ are completely determined by the definitions (6.1) and (6.2). An arbitrary vector can be added to A without changing the resulting B provided this additive term has a zero value for the operation $(\nabla \times)$. This will be so if A is the gradient of a scalar function

$$\boxed{\hat{A} = A + \nabla \Lambda} \tag{6.5}$$

then B will be unchanged. According to (6.2), E will be affected, unless Φ is

$$\boxed{\hat{\Phi} = \Phi - \frac{1}{c} \dot{\Lambda}} \quad . \tag{6.6}$$

In (6.5) and (6.6) we are free in choosing Λ, so we can use this freedom in Λ to simplify Eqs. (6.3) and (6.4). An obvious choice is

$$\boxed{\nabla \cdot A + \frac{1}{c} \dot{\phi} = 0} \quad . \tag{6.7}$$

This is *Lorentz gauge*. This requires that the gauge function, Λ, satisfies (6.5), (6.6) and (6.7):

$$\nabla \cdot A + \nabla \cdot \nabla \Lambda + \frac{1}{c} \dot{\Phi} - \frac{1}{c^2} \ddot{\Lambda} = 0$$

$$\boxed{\nabla^2 \Lambda - \frac{1}{c^2}\ddot{\Lambda} = 0} \quad . \tag{6.8}$$

Electrodynamic potentials in Lorentz gauge satisfy the equations

$$\boxed{\begin{aligned} \nabla^2 \boldsymbol{A} - \frac{1}{c^2}\ddot{\boldsymbol{A}} &= -\frac{4\pi}{c}\boldsymbol{J} \\[1em] \nabla^2 \Phi - \frac{1}{c^2}\ddot{\Phi} &= -4\pi\varrho \quad . \end{aligned}}$$

$$\tag{6.9}$$
$$\tag{6.10}$$

Equations (6.9), (6.10) and (6.7) are equivalent to Maxwell's equations (2.4) to (2.7) together with the constitutive equations (2.1, 2.2, 2.3). These four equations are decoupled and have the form of an inhomogeneous wave equation.

6.2   Green's Function for the Wave Equation

The wave equations (6.9) and (6.10) have the form

$$\nabla^2 \psi - \frac{1}{v^2}\ddot{\psi} = -f(\boldsymbol{x},t), \tag{6.11}$$

where $f(\boldsymbol{x},t)$ is the given source distribution and c is the propagation velocity as derived in (2.32). Since the time dependence of (6.11) complicates the problem, it is useful to eliminate the time in (6.11) by taking the inverse Fourier transform. Substituting

$$\psi(\boldsymbol{x},t) = \int_{-\infty}^{\infty} \Psi(\boldsymbol{x},\omega)\,e^{\,i\omega t}\,d\omega,$$

$$f(\boldsymbol{x},t) = \int_{-\infty}^{\infty} F(\boldsymbol{x},\omega)\,e^{\,i\omega t}\,d\omega, \tag{6.12}$$

into (6.11) we find that $\Psi(\boldsymbol{x},\omega)$ obeys the time-independent Helmholtz wave equation

$$\boxed{(\nabla^2 + k^2)\,\Psi(\boldsymbol{x},\omega) = -F(\boldsymbol{x},\omega)} \tag{6.13}$$

for brevity we have set

$$k = \omega/c. \tag{6.14}$$

The left-hand side of (6.13) is linear in Ψ, but the function F on the right-hand side prevents the application of a superposition principle to the complete equation. An arbitrary linear combination of solutions of the *homogeneous* equation can always be added to any particular solution of (6.13). A convenient method to construct a particular solution of (6.13) that fulfills the given initial or boundary conditions is

provided by Green's functions. These are solutions of an inhomogeneous differential equation in the form of (6.13) with a convenient form on the right hand side. This is chosen such that the general function F can be expanded into a linear combination of these special functions. The solution Ψ of the general equation (6.13) is then formed by the same kind of linear superposition as F.

The Green's function $G(\boldsymbol{x},\boldsymbol{x}')$ therefore is defined as the solution of

$$\left(\nabla^2 + k^2\right) G(\boldsymbol{x},\boldsymbol{x}') = -\delta(\boldsymbol{x} - \boldsymbol{x}') \qquad . \tag{6.15}$$

$G(\boldsymbol{x},\boldsymbol{x}')$ must be a solution to (6.15), this expression also has the symmetries specific to the problem and satisfies the initial or boundary conditions. The inverse FT of $G(\boldsymbol{x},\boldsymbol{x}')$,

$$g(\boldsymbol{x},t,\boldsymbol{x}',t') = \frac{1}{2\pi} \int G(\boldsymbol{x},\boldsymbol{x}') e^{i\omega t}\, d\omega\,, \tag{6.16}$$

is then a solution of the inverse FT of (6.15), that is of

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) g(\boldsymbol{x},\boldsymbol{x}',t,t') = -\delta(\boldsymbol{x} - \boldsymbol{x}')\,\delta(t - t')\,. \tag{6.17}$$

The Green's function method will now be applied to the case of spherical waves emitted from a point source. A spherical coordinate system (r, ϑ, φ) is appropriate in this case, so that (6.15) becomes [see Appendix (A.27)]

$$\frac{1}{r}\frac{d^2}{dr^2}(rG) + k^2 G = -\delta(r)\,. \tag{6.18}$$

For $r \neq 0$ the solution is

$$G = \frac{1}{4\pi r} e^{\pm ikr}\,, \tag{6.19}$$

It can be shown that this solution also applies to $r \to 0$. The corresponding Green's function for the time dependent problem is then obtained by the inverse FT of (6.15), that is, by

$$g(\boldsymbol{x},\boldsymbol{x}';t) = \frac{1}{4\pi\,|\,\boldsymbol{x} - \boldsymbol{x}'\,|}\frac{1}{2\pi} \int\limits_{-\infty}^{\infty} e^{i(\omega t \pm k|\boldsymbol{x} - \boldsymbol{x}'|)}\, d\omega\,, \tag{6.20}$$

or, introducing a new, retarded (or advanced) time t'

$$t' = t \mp \frac{k}{\omega}\,|\,\boldsymbol{x} - \boldsymbol{x}'\,| = t \mp \frac{|\,\boldsymbol{x} - \boldsymbol{x}'\,|}{v}\,, \tag{6.21}$$

by

$$g(\boldsymbol{x},\boldsymbol{x}',t,t') = \frac{\delta\left(t' + \dfrac{|\,\boldsymbol{x} - \boldsymbol{x}'\,|}{v} - t\right)}{4\pi\,|\,\boldsymbol{x} - \boldsymbol{x}'\,|}\,. \tag{6.22}$$

In (6.21) two choices of the sign are in principle possible; here the upper sign, which represents the *retarded* potentials is selected because only this results in the proper causal relation. Selecting the retarded and not the advanced solution is an indication of the *arrow of time*.

The solution for the wave equation (6.11) is then, in the absence of boundaries,

$$\psi(\mathbf{x},t) = \frac{1}{4\pi} \int\!\!\int \frac{f(\mathbf{x}',t')\,\delta\left(t' + \frac{|\mathbf{x}-\mathbf{x}'|}{v} - t\right)}{|\mathbf{x}-\mathbf{x}'|}\, d^3x'\, dt'. \tag{6.23}$$

If the integration over t' is performed we finally arrive at the result

$$\boxed{\psi(\mathbf{x},t) = \frac{1}{4\pi} \int \frac{f\left(\mathbf{x}',t - \frac{|\mathbf{x}-\mathbf{x}'|}{v}\right)}{|\mathbf{x}-\mathbf{x}'|}\, d^3x'} \quad . \tag{6.24}$$

A short hand version of this is

$$\boxed{\psi(\mathbf{x},t) = \frac{1}{4\pi} \int \frac{[f(\mathbf{x}',t')]_{\text{ret}}}{|\mathbf{x}-\mathbf{x}'|}\, d^3x'} \quad , \tag{6.25}$$

for $[\ ]_{\text{ret}}$, t is the retarded time $t' = t - |\mathbf{x}-\mathbf{x}'|/v$.

If we use the expression (6.24) or (6.25) for the retarded Green's function in the wave equation for the electrodynamic potential (6.9) and (6.10), we can write any reasonable solution of Maxwell's equation as

$$A(\mathbf{x},t) = \frac{\mu}{c} \int\!\!\int\!\!\int \frac{\mathbf{J}\left(\mathbf{x}',t - \frac{|\mathbf{x}-\mathbf{x}'|}{v}\right)}{|\mathbf{x}-\mathbf{x}'|}\, d^3x', \tag{6.26}$$

$$\Phi(\mathbf{x},t) = \frac{1}{\varepsilon} \int\!\!\int\!\!\int \frac{\varrho\left(\mathbf{x}',t - \frac{|\mathbf{x}-\mathbf{x}'|}{v}\right)}{|\mathbf{x}-\mathbf{x}'|}\, d^3x'. \tag{6.27}$$

To determine the electrodynamic potentials we must know the distribution of the currents $\mathbf{J}$ and electric charges ϱ over the whole volume. The actual situation is more complicated, since specifying A and Φ (or E and B) result in currents and charges. Thus this is a coupled problem, requiring a self-consistent field.

For investigations of the radiation fields of a system of oscillating charges and currents, there is no loss of generality considering only quantities that vary sinusoidally with time. Therefore we adopt

$$\varrho(\mathbf{x},t) = \varrho(\mathbf{x})\,\mathrm{e}^{-\mathrm{i}\omega t},$$
$$\mathbf{J}(\mathbf{x},t) = \mathbf{J}(\mathbf{x})\,\mathrm{e}^{-\mathrm{i}\omega t}. \tag{6.28}$$

The amplitudes $\varrho(x)$ and $J(x)$ can be complex quantities, so that the phases of the oscillations will be dependent on the position x. According to (6.26) the vector potential generated by these currents is

$$A(x,t) = A(x)\,e^{-i\omega t}, \tag{6.29}$$

where

$$A(x) = \frac{1}{c} \iiint_V J(x') \frac{e^{ik|x-x'|}}{|x-x'|}\, d^3x' \tag{6.30}$$

and

$$k = \frac{\omega}{c} = \frac{2\pi}{\lambda}. \tag{6.31}$$

Here V is the volume in which the current J flows.

For analytic solutions, the problem is often greatly simplified if the currents J (and charges, if present) are confined to a finite region (for example the antenna proper) and we postulate a given distribution for J. If one can assume that reactions of fields on these currents can be neglected, the integrals in (6.26) and (6.27) can be computed. Usually one can also take $\mu=\varepsilon=1$.

6.3   The Hertz Dipole

Next, we analyze the Hertz dipole as an example of a simple antenna. Here the volume over which the integrals must be computed is that of an infinitesimal dipole with a length Δl and a cross section q. H. Hertz calculated the solution for this configuration and then performed the experiment to demonstrate the existence of electromagnetic waves in 1888.

If a current I is flowing in this dipole, the current density is $|J| = I/q$ in the dipole, and $J = 0$ outside. Then the integration volume is only that of the dipole $dV = q\Delta l$. If the rectangular coordinate system (x,y,z) is oriented such that the dipole extends from $z = -\Delta l/2$ to $z = +\Delta l/2$ on the z axis, then $J_x = J_y = 0$, and

$$J_z = \frac{I}{q}\,e^{-i\omega t}.$$

Following (6.26), the vector potential A has only the component

$$A_z = \frac{1}{c} \int_{-\Delta l/2}^{\Delta l/2} \frac{I}{q} \cdot \frac{q}{r} \cdot e^{-i\omega\left(t-\frac{r}{c}\right)}\, dl,$$

resulting in

$$\boxed{A_z = \frac{1}{c}\frac{I\Delta l}{r}e^{-i(\omega t - kr)}} \quad. \tag{6.32}$$

Fig. 6.1 The geometry and coordinate system used for the treatment of radiation from an electric dipole (Hertz dipole)
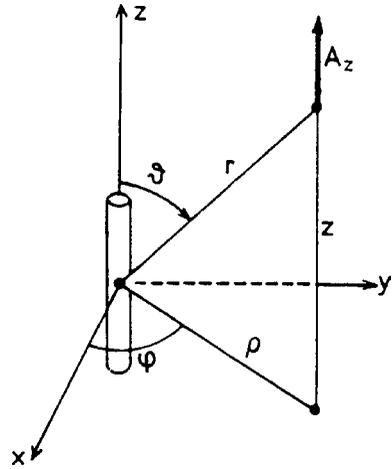
Thus A_z is constant on concentric spheres $r^2 = x^2 + y^2 + z^2$. Introducing cylindrical coordinates $(\varrho, \varphi, z,$ see Fig. 6.1) we derive from (6.1)

$$B_\varphi = (\nabla \times A)_\varphi = \frac{\partial A_\varrho}{\partial z} - \frac{\partial A_z}{\partial \varrho}.$$

But since $A_\varrho \equiv 0$, we find that

$$B_\varphi = -\frac{\partial A_z}{\partial \varrho} = -\frac{\partial A_z}{\partial r}\frac{\partial r}{\partial \varrho}.$$

Since

$$r^2 = \varrho^2 + z^2,$$

$$\frac{\partial r}{\partial \varrho} = \frac{\varrho}{r} = \sin \vartheta,$$

and

$$H_\varphi = B_\varphi,$$

we find that

$$H_\varphi = -\mathrm{i}\,\frac{I\Delta l}{2\lambda}\frac{\sin \vartheta}{r}\left[1 - \frac{1}{\mathrm{i}kr}\right]\mathrm{e}^{-\mathrm{i}(\omega t - kr)} \tag{6.33}$$

where we have used (6.31). The other components of H are zero, because $A_\varrho \equiv A_\vartheta \equiv 0$. For the electric field, we again make use of Maxwell's equations. According to (2.7)

$$\nabla \times \boldsymbol{H} = \frac{1}{c}\dot{\boldsymbol{D}} + \frac{4\pi}{c}\boldsymbol{J}.$$

Outside the region occupied by the dipole, $\sigma = 0$, so $\boldsymbol{J} = 0$. For a harmonic wave, from (2.35), $\dot{\boldsymbol{D}} = -\mathrm{i}\omega\boldsymbol{D}$ so

$$\boldsymbol{E} = \frac{\mathrm{i}c}{\omega}(\nabla \times \boldsymbol{H}).$$

Returning to a spherical coordinate system (r, ϑ, φ) we find that

$$E_\vartheta = \frac{\mathrm{i}}{\omega}(\nabla \times \boldsymbol{H})_\vartheta.$$

since $H_r \equiv 0$, from (A.26),

$$(\nabla \times \boldsymbol{H})_\vartheta = -\frac{1}{r}\frac{\partial(rH_\varphi)}{\partial r}$$

so that

$$E_\vartheta = -\mathrm{i}\frac{I\Delta l}{2\lambda}\frac{\sin\vartheta}{r}\left[1 - \frac{1}{\mathrm{i}kr} + \frac{1}{(\mathrm{i}kr)^2}\right]\mathrm{e}^{-\mathrm{i}(\omega t - kr)}.$$

$$(6.34)$$

Finally, since $H_\vartheta = 0$, we find that

$$(\nabla \times \boldsymbol{H})_r = \frac{1}{r\sin\vartheta}\frac{\partial(\sin\vartheta H_\varphi)}{\partial\vartheta}$$

and thus

$$E_r = \mathrm{i}\frac{I\Delta l}{2\lambda}\frac{2\cos\vartheta}{r}\left[\frac{1}{\mathrm{i}kr} - \frac{1}{(\mathrm{i}kr)^2}\right]\mathrm{e}^{-\mathrm{i}(\omega t - kr)}.$$

$$(6.35)$$

$E_\varphi \equiv 0$ since $H_r \equiv 0$ and $H_\vartheta \equiv 0$. Therefore (6.33), (6.34) and (6.35) are the only non-vanishing components of the electromagnetic field of an electric dipole. Forming the scalar product of $\boldsymbol{E}$ and $\boldsymbol{H}$ we find

$$\boldsymbol{E} \cdot \boldsymbol{H} = 0 \qquad\qquad (6.36)$$

As in the case of plane electromagnetic waves $\boldsymbol{E}$ and $\boldsymbol{H}$ of a radiating dipole are perpendicular everywhere. However, the expressions for $\boldsymbol{E}$ and $\boldsymbol{H}$ contain different powers of the distance r. Near and far field of an oscillating Hertz dipole are shown
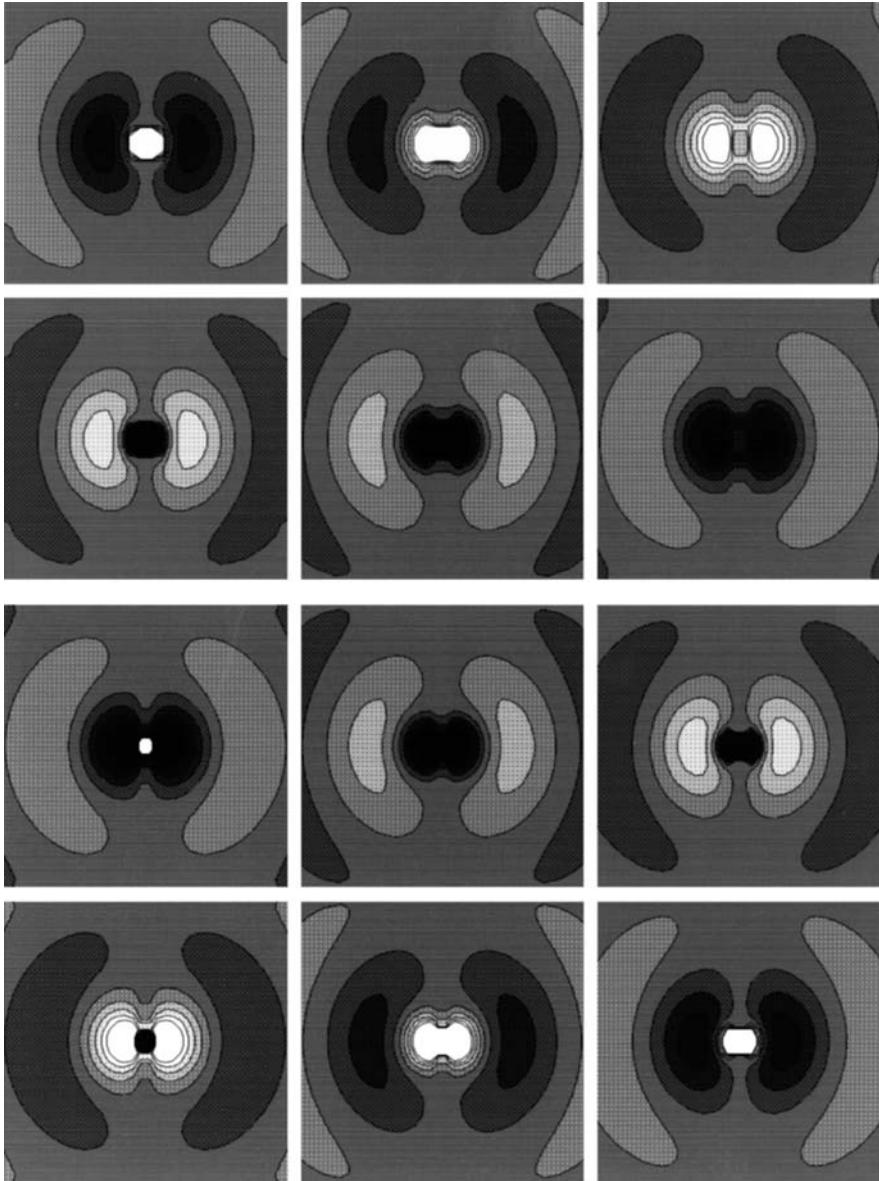
Fig. 6.2 The field for an oscillating Hertz dipole for the region close to the dipole

in Figs. 6.2 and 6.3. The $1/r^2$ terms in (6.33), (6.34) and (6.35) represent the *induction field* of a quasistationary electric dipole for slow oscillations. In addition, for $\boldsymbol{E}$ there is the $1/r^3$ field of the *static dipole*. Most important for $r \gg l$ is the *radiation field* which has a $1/r$ dependence. This has components

$$H_\varphi = -\mathrm{i}\,\frac{I\Delta l}{2\lambda}\frac{\sin\vartheta}{r}\mathrm{e}^{-\mathrm{i}(\omega t - kr)} \tag{6.37}$$

$$\tag{6.38}$$
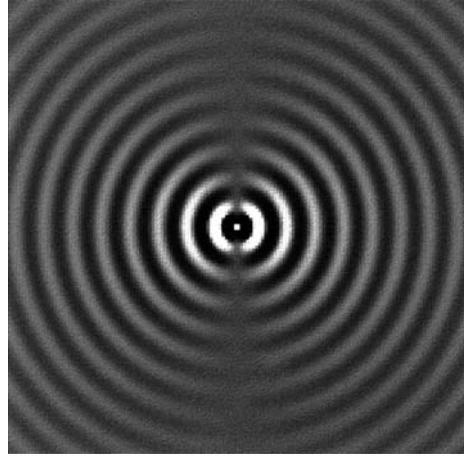
$$E_\vartheta = -\mathrm{i}\,\frac{I\Delta l}{2\lambda}\frac{\sin\vartheta}{r}\mathrm{e}^{-\mathrm{i}(\omega t - kr)}$$

As in the case of plane electromagnetic waves, we have

$$\frac{|\boldsymbol{E}|}{|\boldsymbol{H}|} = 1. \tag{6.39}$$

The Poynting vector for the radiation field is directed radially outward, its time average value is, according to (2.21),

$$|\langle\boldsymbol{S}\rangle| = \frac{c}{4\pi}\,|\,\mathrm{Re}\,(\boldsymbol{E}\times\boldsymbol{H}^*)\,| = \frac{c}{4\pi}\left(\frac{I\Delta l}{2\lambda}\right)^2\frac{\sin^2\vartheta}{r^2} \quad . \tag{6.40}$$

Thus the total radiated power is

$$P = \int\limits_0^{2\pi}\int\limits_0^{\pi} |\langle\boldsymbol{S}\rangle|\,r^2\sin\vartheta\,\mathrm{d}\vartheta\,\mathrm{d}\varphi\,,$$

using

$$\int\limits_0^{\pi}\sin^3\vartheta\,\mathrm{d}\vartheta = \tfrac{4}{3}$$

this becomes

$$P = \frac{2c}{3}\left(\frac{I\Delta l}{2\lambda}\right)^2 \qquad . \qquad (6.41)$$

In MKS units, with I in amperes, and λ in meters, this is

$$\left[\frac{P}{\text{Watts}}\right] = 395\left(\frac{I\Delta l}{\lambda}\right)^2$$

This expression has the same general form as that giving the ohmic losses of a resistor

$$[P] = \tfrac{1}{2}RI^2$$

so that we are led to introduce a *radiation impedance*, R_S, of the Hertz dipole

$$R_S = \frac{c}{3}\left(\frac{\Delta l}{\lambda}\right)^2 . \qquad (6.42)$$

For MKS units, this is

$$\left[\frac{R_S}{\text{Ohms}}\right] = 790\left(\frac{\Delta l}{\lambda}\right)^2$$

6.3.1 Arrays of Emitters

An important extension of the case of a single emitting element is the case of an array. A particular case is one in which one has a number of identical emitters placed parallel to each other in a plane, at a distance, D. Each element of the array produces the same E field. From the principle of superposition, the E fields are additive. Furthermore, we assume that the E field of each dipole does not affect any of the others. In two dimensions, for a distant observer at r, two elements have a combined field of

$$E = E_1 + E_1 e^{i\Phi} \qquad (6.43)$$

The geometric arrangement is given in Fig. 6.4. with $\Phi = \frac{2\pi}{\lambda}D\sin\phi$ and $\frac{2\pi}{\lambda} = k$, we have

$$E = E_1\left[1 + e^{ikD\sin\phi}\right] \qquad (6.44)$$

If we call the term in square brackets $\hat{S}$, we have

$$\hat{S} = 1 + e^{ikD\sin\phi} \qquad (6.45)$$
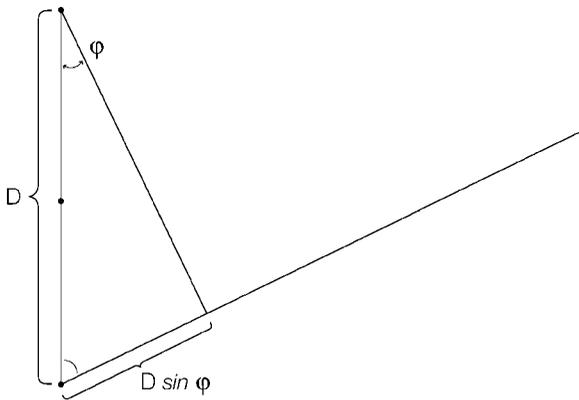
Fig. 6.4 The arrangement for two (at 0 and D) and three $(0, D/2$ and $D)$ emitters. The emitters are placed on the vertical axis

for two emitters. For three emitters spaced equally between 0 and D, this is

$$\hat{S} = 1 + e^{i\,kD/2\sin\phi} + e^{i\,kD\sin\phi} \tag{6.46}$$

A plot of $|\hat{S}|^2$ for a 2 and 3 element system is shown in Fig. 6.5. The 2 element system has a set of equally spaced maxima and minima. For the 3 element system some of the maxima are replaced by weaker secondary maxima. In antenna jargon, these secondary maxima are referred to as "sidelobes".
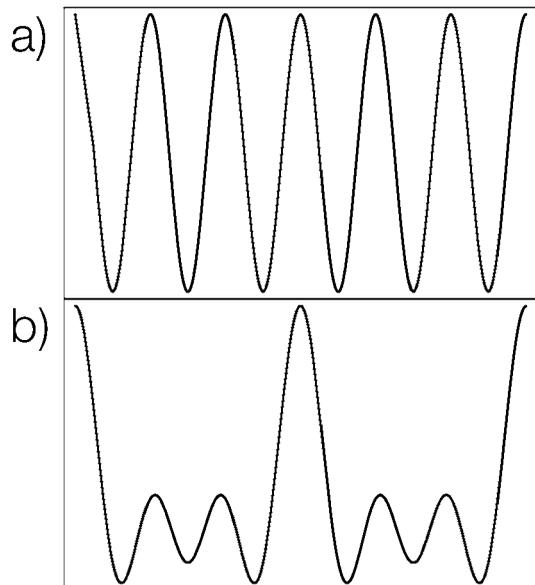
Fig. 6.5 Power patterns for different configurations. Panel (**a**) shows two elements and panel (**b**) shows the superposition of three emitters

For a set of N emitters spaced equally by D, we have:

$$\hat{S} = \Sigma_{n=0}^{N} e^{i\,knD\,\sin(\phi)} \tag{6.47}$$

The sum is

$$\hat{S} = e^{i\,kD\,\sin(\phi)} \cdot e^{-i\,(N-1)kD/2\,\sin(\phi)} \cdot \left[\frac{\sin\frac{kND}{2}\sin(\phi)}{\sin\frac{kD}{2}\sin(\phi)} \right] \tag{6.48}$$

(see problem 1 at the end of this chapter). The radiated power is proportional to $|\hat{S}|^2$. Then for (6.48), the radiated power is proportional to

$$|\hat{S}|^2 = \left[\frac{\sin\left(\frac{kND}{2}\sin(\phi)\right)}{\sin\left(\frac{kD}{2}\sin(\phi)\right)} \right]^2 \tag{6.49}$$

On axis, $\phi = 0$, which gives a maximum value. At an angle

$$\sin(\phi) = \frac{\lambda}{ND}$$

the sum of the contributions reaches zero. Using $ND = \hat{D}$, the expression for the angular distance to the zero point is

$$\theta = \frac{\lambda}{\hat{D}}$$

This is usually referred to as the *Rayleigh Criterion*, since Rayleigh first showed that when a second point source is separated from a point source by such an angle, these two can be distinguished. That is, the second source is located at the first null point of the diffraction pattern of the first source.

6.3.2 Arrays of Hertz Dipoles

For the particular case where each element is a Hertz dipole, the radiation field of each dipole is given by (6.38). The sum of the E fields is the product of (6.38) and (6.49). If each dipole antenna is connected a single source, the currents in the dipoles are related, each with a definite phase, ϕ, and amplitude, I_0. For the case of two dimensions, we align the dipoles along the y axis, with dipoles parallel to the z axis. Setting $\theta = 90°$ in (6.38), the E_ϑ vector is in the $x - y$ plane. Then the total E is:

$$E_\vartheta = \left(-i\frac{I\Delta l}{2\lambda}\frac{1}{r} \right) e^{-i(\omega t - kr)}\,\hat{S} \tag{6.50}$$

where $\hat{S}$ is given by (6.48).

One can change the phase of each dipole to alter the direction of the maximum. This so-called *electronic steering* is used to direct arrays of dipoles, which are referred to as phased arrays. In radio astronomy such concepts are used in instruments such as LOFAR and the Allen Array. This will be used in the Square Kilometer Array (SKA).

6.4 Radiation Fields of Filled Antennas

6.4.1 Two Dimensional Far Field

We start with (6.50) but change the $\hat{S}$ factor by keeping the total size of the array nD constant, while increasing the number of dipoles, n and simultaneously decreasing the distance, D between these. Then the summation in (6.47) becomes an integral with variable x'. If we replace $\hat{S}$ by the symbol $g(x')$ and change the integration variable, the expression becomes:

$$g(x') = \int_0^N J(x')\,e^{i k x' \sin(\phi)} dx' \tag{6.51}$$

The expression for the $\boldsymbol{E}$ field is

$$dE_y(\phi) = -i\,J_0\,g(x')\,\frac{1}{r}\,e^{-i(\omega t - kr)}\,dx' \;. \tag{6.52}$$

The current grading $g(x)$ takes into account changes in the currents across the aperture; $g(x)$ and the far field pattern are FT pairs. Given the importance of this expression, we present a few one dimensional FT pairs in graphical form in Fig. 6.6. The most commonly used expression for a grading function is the Gaussian, since the FT of a Gaussian is another Gaussian.
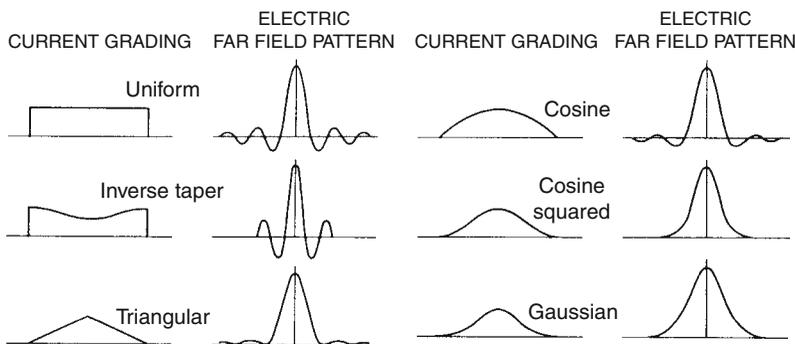
Fig. 6.6 Pairs of current gradings and corresponding electric far field pattern. For each pair, the sketch on the *left* is a current grading across the one-dimensional aperture, g. On the *right* is the corresponding electric field pattern in the far field [adapted from Kraus (1986)]

6.4.2   Three Dimensional Far Field

The clearest difference between the case of two and three dimensions is the inclusion of coordinate systems. For two dimensions, either an $x - y$ or $r - \theta$ coordinate system is a reasonable choice. For three dimensions, there can be many coordinate systems, depending on geometry, and, in addition, the coordinates of the observer and the aperture will be rather different. So, for example, a rectangular aperture, an $x - y - z$ coordinate system is the choice, while for a circular aperture the choice is a cylindrical coordinate system.

For a two-dimensional antenna structure, the current distribution $\boldsymbol{J}(\boldsymbol{x})$ must be specified for the aperture. We assume that this is a plane aperture. We choose the coordinate system such that the aperture is a finite area of the plane $z' = 0$; this aperture is assumed to be a surface of constant wave phase and the unit vector of the current density is chosen to be $\boldsymbol{J}_0 = (0, J_0, 0)$.

The current in a surface element $\mathrm{d}x'\,\mathrm{d}y'$ at $\boldsymbol{x}'$ in the aperture $\mathscr{A}$ is then $J_0\,g(\boldsymbol{x}')$ $\mathrm{d}x'\,\mathrm{d}y'$. If we take $\boldsymbol{x}$ to be approximately perpendicular to $\mathscr{A}$ the only component of the electric field induced by this current element in $\boldsymbol{x}$ is, according to (6.38),

$$\mathrm{d}E_y = -\frac{\mathrm{i}}{2}\lambda J_0\,g(\boldsymbol{x}')\frac{F_\mathrm{e}(\boldsymbol{n})}{|\,\boldsymbol{x} - \boldsymbol{x}'\,|}\,\mathrm{e}^{-\mathrm{i}\,(\omega t - k\,|\boldsymbol{x} - \boldsymbol{x}'|)}\,\frac{\mathrm{d}x'}{\lambda}\,\frac{\mathrm{d}y'}{\lambda}\ . \tag{6.53}$$

Again we make use of an extension of the case of a Hertz dipole. Here $F_\mathrm{e}(\boldsymbol{n})$ is the field pattern of the current element for the direction $\boldsymbol{n} = \boldsymbol{x}/|\boldsymbol{x}|$ which, for the Hertz dipole, is $\sin\vartheta$ where ϑ is the angle between $\boldsymbol{n}$ and $\boldsymbol{J}_0$. The total field in $\boldsymbol{x}$ is then obtained by integrating (6.53) over the full aperture.

For the far field, at distance r, we have assumed that the extent of the aperture is small compared to r, the integral can be simplified considerably by introducing the *Fraunhofer Approximation* (see Fig. 6.7), $|\,\boldsymbol{x} - \boldsymbol{x}'\,| \approx r - \boldsymbol{n}\cdot\boldsymbol{x}'$, where $r = |\,\boldsymbol{x}'\,|$. Because $r \gg |\boldsymbol{n}\cdot\boldsymbol{x}'|$, we can neglect $\boldsymbol{n}\cdot\boldsymbol{x}'$ compared to r everywhere except in the exponent. There the term $k\,\boldsymbol{n}\cdot\boldsymbol{x}'$ appears. We assumed that the aperture is larger than λ, that is $|\,\boldsymbol{x}'\,| > \lambda$,

$$k\,(\boldsymbol{n}\cdot\boldsymbol{x}') = \frac{2\pi}{\lambda}(\boldsymbol{n}\cdot\boldsymbol{x}') \gtrless 1\ .$$
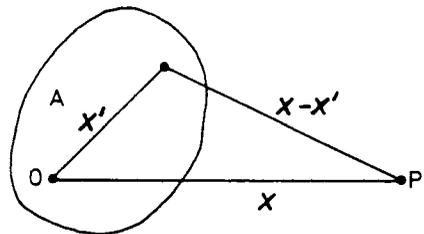
Fig. 6.7 The geometry of the Fraunhofer approximation

Then we have

$$E_y = -\frac{i}{2} \lambda J_0 \frac{F_e(\boldsymbol{n})}{r} e^{-i(\omega t - kr)} \iint\limits_{\mathcal{A}} g(\boldsymbol{x}') e^{-ik\boldsymbol{n}\cdot\boldsymbol{x}'} \frac{dx'}{\lambda} \frac{dy'}{\lambda} \qquad (6.54)$$

or

$$E_y = -i\lambda J_0 \pi \frac{F_e(\boldsymbol{n})}{r} f(\boldsymbol{n}) e^{-i(\omega t - kr)} \qquad (6.55)$$

where

$$\boxed{f(\boldsymbol{n}) = \frac{1}{2\pi} \iint\limits_{-\infty}^{\infty} g(\boldsymbol{x}') e^{-ik\boldsymbol{n}\cdot\boldsymbol{x}'} \frac{dx'}{\lambda} \frac{dy'}{\lambda}} \qquad . \qquad (6.56)$$

Thus the integral over the aperture $\mathcal{A}$ in (6.54) has been formally replaced by the two dimensional Fourier integral (6.56) by setting $g(\boldsymbol{x}') = 0$ for $\boldsymbol{x}' \notin \mathcal{A}$. The expression for the magnetic field strength $\boldsymbol{H}$ is similar to (6.55). The normalized power pattern P_n (see 6.40 and following) is then

$$P_n(\boldsymbol{n}) = \frac{P(\boldsymbol{n})}{P_{\max}} = \frac{|\boldsymbol{E}\cdot\boldsymbol{E}^*|}{|\boldsymbol{E}\cdot\boldsymbol{E}^*|_{\max}} ,$$

so that

$$\boxed{P_n = \frac{|f(\boldsymbol{n})|^2}{|f_{\max}|^2}} \qquad . \qquad (6.57)$$

As will be shown in Chap. 9, the FT relations also play a fundamental role in interferometry.

An excellent illustration of the application of (6.56) and (6.57) is given by the example below.

The Normalized Power Pattern of a Rectangular Aperture with Uniform Illumination

If the linear dimensions of the aperture are L_x and L_y, then the current grading can be written as

$$g(x, y) = \begin{cases} 1 \text{ for } |x| \leqq L_x/2, \ |y| \leqq L_y/2 \\ 0 \text{ otherwise} \end{cases} . \qquad (6.58)$$

The components of the unit vector are $\boldsymbol{n} = (l, m, n)$ with $l^2 + m^2 + n^2 = 1$. If the aperture is part of the plane $z' = 0$, then (6.56) becomes

$$f(l, m) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} g(x', y') \exp\left\{ -i\frac{2\pi}{\lambda}(lx' + my') \right\} \frac{dx'}{\lambda} \frac{dy'}{\lambda} .$$

With (6.58) this becomes

$$f(l, m) = \frac{\sin(\pi l L_x/\lambda)}{\pi l L_x/\lambda} \frac{\sin(\pi m L_y/\lambda)}{\pi m L_y/\lambda} \qquad (6.59)$$

and the normalized power pattern is

$$P_{\mathrm{n}}(l,m) = \left[\frac{\sin(\pi l L_x/\lambda)}{\pi l L_x/\lambda} \frac{\sin(\pi m L_y/\lambda)}{\pi m L_y/\lambda} \right]^2 \quad . \tag{6.60}$$

The main beam is the solid angle between the first nulls of P_{n} at

$$l_0 = \pm \lambda / L_x ; \qquad m_0 = \pm \lambda / L_y \tag{6.61}$$

The full width to half power (FWHP), i.e. the angle between those points of the main beam for which $P_{\mathrm{n}} = 1/2$, is

$$\mathrm{FWHP}_x = 0.88 \frac{\lambda}{L_x} \mathrm{rad} = 50.3° \frac{\lambda}{L_x} , \tag{6.62}$$

with a similar expression for FWHP_y. The first side lobes are located at

$$\frac{1}{2} \frac{\lambda}{L_x} = 28.0° \frac{\lambda}{L_x}$$

from the axis; these have an intensity (relative to the main beam) of $P_{\mathrm{n}} = 0.0472$ corresponding to an attenuation of 13.3 dB. This is a rather high side-lobe level. For rectangular apertures, the far field patterns are the products of the one-dimensional FT pairs.

The full width to half power (FWHP), relative gain and sidelobe levels depend on the shape of g. These quantities can be changed by altering the illumination or grading $g(x,y)$. This will be investigated in more detail for the case of circular apertures because these are widely used for large antennas in radio astronomy. Usually small antennas with a rectangular apertures are used as feed horns at wavelengths shorter than 30 cm to efficiently couple the receiver to the free space waves focussed by the radio telescope. Large horn antennas are also used for calibration purposes. Well known examples of radio telescopes with a rectangular aperture are the Bell Laboratories Horn antenna, at Holmdel, N. J., used to discover the 2.7 K background radiation and the "Little Big Horn" of the National Radio Astronomy Observatory in Greenbank W.Va., USA, that was used to establish the time variation of the flux density of the supernova remnant Cassiopeia A.

6.4.3   Circular Apertures

Circularly symmetric paraboloids are the most commonly used antennas. For a circular aperture it is convenient to introduce polar coordinates ϱ, φ by

$$\begin{aligned} x &= \lambda \varrho \cos \varphi \\ y &= \lambda \varrho \sin \varphi . \end{aligned} \tag{6.63}$$

If we now assume that the aperture is defined by $\varrho \leqq D/2\lambda$ and that the current grading g depends on ϱ only, then the resulting beam pattern will also show circular symmetry; instead of two directional cosines l and m, only a single value, u, the sine of the angle between $\boldsymbol{n}$ and the direction of the main beam is needed. Substituting (6.63) into (6.56) we obtain

$$f(u) = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} g(\varrho) e^{-2\pi i u \varrho \cos \varphi} \varrho \, d\varrho \, d\varphi \, . \qquad (6.64)$$

Since the integral representation of the Bessel function of order zero is

$$J_0(z) = \frac{1}{2\pi} \int_0^{2\pi} e^{iz \cos \varphi} \, d\varphi \, , \qquad (6.65)$$

(6.64) can be written as

$$f(u) = \int_0^{\infty} g(\varrho) J_0(2\pi u \varrho) \varrho \, d\varrho \qquad (6.66)$$

For the case of circular symmetry the electric and magnetic field strength is thus the Hankel transform of the current grading. For the normalized beam pattern we then obtain

$$P_{\mathrm{n}}(u) = \left[\frac{\int_0^{\infty} g(\varrho) J_0(2\pi u \varrho) \varrho \, d\varrho}{\int_0^{\infty} g(\varrho) \varrho \, d\varrho} \right]^2 \qquad (6.67)$$

because $J_0(0) = 1$.

For a circular aperture with uniform illumination, that is for

$$g(\varrho) = \begin{cases} 1 \text{ for } \varrho \leqq D/2\lambda \\ 0 \text{ else} \end{cases} \qquad (6.68)$$

(6.67) then becomes

$$P_{\mathrm{n}}(u) = \left[\frac{\int_0^{D/2\lambda} J_0(2\pi u \varrho) \varrho \, d\varrho}{\int_0^{D/2\lambda} \varrho \, d\varrho} \right]^2 = \left[\frac{2\lambda}{\pi u D} \int_0^{\pi u D/\lambda} J_0(z) z \, dz \right]^2 \, . \qquad (6.69)$$

For Bessel functions, the relation

$$\frac{d}{dz}\{z^n J_n(z)\} = z^n J_{n-1}(z) \tag{6.70}$$

[cf. Abramowitz and Stegun (1964), Eq. (9.1.30)] applies so that

$$x^n J_n(x) = \int_0^x z^n J_{n-1}(z)\,dz,$$

and we obtain for the normalized power pattern

$$P_n(u) = \left[\frac{2J_1(\pi u D/\lambda)}{\pi u D/\lambda}\right]^2 = \Lambda_1^2\,(\pi u D/\lambda) \tag{6.71}$$

where

$$\Lambda_1(u) = \frac{2}{u}J_1(u). \tag{6.72}$$

If the region up to and including the first nulls of $P_n(u)$ at $\pi u D/\lambda = 3.83171$ is included in the main beam region, the *full beam width between the first nulls, BWFN*, that is from one null to the other, is:

$$\text{BWFN} = 2.439\,\frac{\lambda}{D}\,\text{rad} \simeq 139.8°\,\frac{\lambda}{D} \tag{6.73}$$

and the *full width to half power beam width, FWHP*, that is, from one half power point to the other, is:

$$\text{FWHP} = 1.02\,\frac{\lambda}{D}\,\text{rad} \simeq 58.4°\,\frac{\lambda}{D} \tag{6.74}$$

For an aperture with a nonuniform illumination or grading the antenna pattern will be different from (6.71), see Fig. 6.8; the relation between grading and antenna
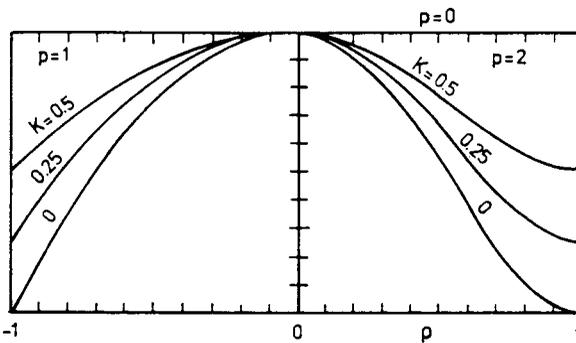
Fig. 6.8 A representative set of illumination tapers $g(\varrho) = K + (1 - \varrho^2)^p$. Note that $p = 0$ is a horizontal line at the *top* of the *right* panel

pattern is given by (6.67). Depending on the choice of $g(\varrho)$, the integral (6.67) may be difficult to evaluate. However the qualitative dependence of P_n on the grading can be obtained by selecting a convenient interpolation formula. If g is chosen such that (6.67) can be evaluated in a closed form, we can obtain analytical expressions. Such a family of functions is

$$g(\varrho) = \left[1 - \left(\frac{2\lambda\varrho}{D}\right)^2\right]^p + K.$$ (6.75)

Because

$$\int_0^1 (1 - r^2)^p J_0(qr)\, r\, dr = \frac{2^p\, p!\, J_{p+1}(q)}{q^{p+1}}$$ (6.76)

[Gradshteyn and Ryzhik (1965), Eq. (6.567.1)], (6.67) can be evaluated in terms of J_p;

$$P_n(u) = \left[\frac{2^{p+1}\, p!\, J_{p+1}(\pi u D/\lambda)}{(\pi u D/\lambda)^{p+1}}\right]^2.$$ (6.77)

The values of BWFN and FWHP are given in Table 6.1.

Table 6.1 Normalized power pattern characteristics produced by aperture illumination following (6.75)

p	K	FWHP (rad)	BWFN (rad)	Relative gain	First side lobe (dB)
0		1.02	2.44	1.00	-17.6
1		1.27	3.26	0.75	-24.6
2		1.47	4.06	0.56	-30.6
1	0.25	1.17	2.98	0.87	-23.7
2	0.25	1.23	3.36	0.81	-32.3
1	0.50	1.13	2.66	0.92	-22.0
2	0.50	1.16	3.02	0.88	-26.5

6.4.4 Antenna Taper Related to Power Pattern

By selecting a taper, one can influence the properties of an antenna, to some extent. A strong gradation results in a lower side-lobe level but lower relative gain. Antennas used for transmission to distant space probes must be designed for high relative gain; that is, one must use a low gradation. However, this results in a high sidelobe level. Since from the prime focus, the far side lobes are directed towards the hot (300 K) ground, low-noise antennas should have a beam pattern with low side-lobe levels. A compromise is to adjust the illumination. A better approach is to make use of the secondary focus, so that the sidelobes are directed at cold sky

(see, e.g. Fig. 7.6). Then for a given sidelobe level, the unwanted noise power will be lower.

The radiation properties of antennas can be changed, somewhat, using the grading function, $g(x)$, as shown in Figs. 6.6 and 6.8. In many applications, the gradations are between the extremes of $p = 2$ and $p = 0$ in Table 6.1. For low noise antennas used in satellite communication, $p = 0$ has advantages since the relative gain should be as large as possible for such point-like sources. However the prime focus should not be used, since the receiver noise is raised substantially due to sidelobes coupled to the 300 K ground radiation. In this case the receiver is mounted in the Cassegrain with sidelobes coupled to cold sky.

However the relative (on-axis) gain is lower. This is achieved by using a strong gradation for $g(\varrho)$.

6.5   The Reciprocity Theorem

So far, we have discussed antennas used to emit radiation. However, in general, the parameters of an antenna when used to receive or transmit radiation are the same. For the Hertz dipole with $\lambda \gg \Delta l$, this is rather simple to prove. However, there is a proof of the general case of the reciprocity theorem that provides a general solution for this problem. The details of this derivation are presented in Appendix D.

6.6   Summary

The Hertz dipole has been used to introduce the expression for power radiated; this will be used later in presentations of spectral line radiation. Arrays of dipoles will enhance the power radiated in a given direction. These dipole arrays can be used to steer beams in direction. This feature is used in the LOFAR and Allen Array instruments, and will be used in the Square Kilometer Array. A collection of such elements can also be used to illuminate a filled aperture.

The basic results for scalar wave diffraction in 2 dimensions can be obtained from a number of starting points. Many are given in optics texts such as Jenkins and White (2001), where the sum of waves passing through an aperture of size D is used to obtain a one dimensional diffraction pattern for the far field or Fraunhofer case. This analysis is used to obtain the well-know result that the angular resolution for a uniformly illuminated aperture is $\theta = \lambda / D$. However, a grading function is difficult to introduce in this analysis.

In presentations of diffraction in 3 dimensions, Rossi (1957) has used the Huygens Principle. Other authors (e.g. Slater and Frank 1933) simply postulated the mathematical expression (6.19), choosing a time delayed expanding spherical wave. For one dimension, geometry and coordinate systems are less of a concern but in two dimensions these must be addressed consistently. We have shown that the expression

for diffraction in one dimension can be derived from the limit of an array of dipoles. We then related the 3 dimensional case to that in 2 dimensions; some details are to be completed (see Problem 9). The two dimensional expression can also be obtained directly from the vector potential, A (see Eq. 6.29). Such a derivation is complex, but even this rather lengthy development has been simplified since polarization was neglected. An excellent summary of the diffraction problem is given in Stratton (pp. 460–470)

Problems

1. Complete the mathematical details of summing the expression in Eq. (6.48). First, show that

$$\hat{S} = \Sigma_{n=0}^{N} q = \frac{1 - q^{n+1}}{1 - q}$$

(multiply by q to obtain one variant, then subtract from the relation above). With $q = e^{ikD\sin(\phi)}$, show that we obtain (6.48).

From this, one can obtain the power pattern:

$$\hat{S}^2 = \left[\frac{\sin\left[k(n+1)D/2\right]\sin(\phi)}{\sin\left[kD/2\right]\sin(\phi)} \right]^2$$

Use limits to show that the square of the x component of (6.60) can be obtained from the expression above.

2. You read that there are antennas without sidelobes. That is, *all* of the energy is contained in the main lobe. Should you believe the report? Comment using qualitative arguments, but *not* detailed calculations.

3. If the size of the pupil of the human eye, D is 0.5 cm, what are the number of wavelengths across this aperture for light of $\lambda = 500$ nm? Compare this to the number of wavelengths across the aperture of a 100 m radio telescope for a wavelength of 2 m, 1 cm. Repeat for the ALMA radio telescope, with a diameter of 12 m, for $\lambda = 1$ cm, 3 mm, 0.3 mm. Discuss the implications of these results.

4. Derive the increase in the radiated power for an array of N dipoles, for the case of phases set to zero in Eq. (6.50). Compare this to the maximum power radiated in a given direction by a Hertz dipole.

5. The full width half power (FWHP) angular size, θ, in radians, of the main beam of a diffraction pattern from an aperture of diameter D is $\theta \approx 1.02\lambda/D$.
(a) Determine the value of θ, in arc min, for the human eye, where $D = 0.3$ cm, at $\lambda = 5 \times 10^{-5}$ cm.
(b) Repeat for a filled aperture radio telescope, with $D = 100$ m, at $\lambda = 2$ cm, and for the very large array interferometer (VLA), $D = 27$ km, at $\lambda = 2$ cm.
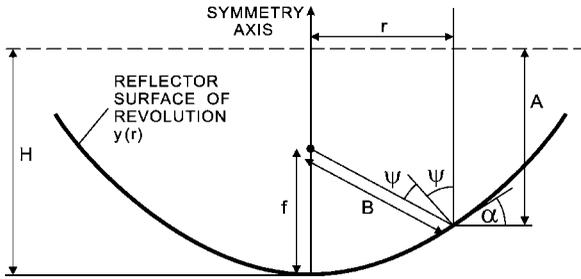
Fig. 6.9 A sketch of a parabola showing angles used in Problem 9

(c) Show that when λ has the units of millimeters, and D the units of kilometers and θ the units of arc seconds, then $\theta = 0.2\lambda/D$. Is this consistent with (6.74)?

6. Hertz used $\lambda \approx 26$ cm for the shortest wavelength in his experiments.
(a) If Hertz employed a parabolic reflector of diameter $D \approx 2$ m, what was the FWHP beam size? (See Problem 3.)
(b) If the $\Delta l \approx 26$ cm, what was the radiation resistance, from equation (6.42)?
(c) Hertz's transmitter was a spark gap. Suppose the current in the spark was 0.5 A. What was the average radiated power?

7. Over the whole world, there have been (on average) 100 radio telescopes of (average) diameter 25 m operating since 1960. Assume that the power received by each is 10^{-15} W over this period of time. What amount of energy has been received in this period of time? Compare this to the energy released by an ash (taken to be 1 g) from a cigarette falling a distance of 2 cm in the earth's gravity.

8. Refer to Fig. 6.9; the surface is $y(x) = (1/4f)x^2$.
(a) Find a general expression for the path from the pupil plane (dashed line) to the focus, f.
(b) If an on-axis plane wave is in the pupil plane, show that for a paraboloid, there is a single focus.
(c) Is there such a relation for a circle, $y(x) = \sqrt{R_0^2 - x^2}$?

9. Show that the two dimensional equation (6.52) and one of the factors in the three dimensional diffraction equation (6.53) are related by equating $(|x - x'|) = \frac{1}{r}$ and identifying J_0 as $I\Delta l/2\lambda$, the current density.

10. If two dipoles spaced by $\lambda/4$ are connected to a coherent input, what is the far field radiation pattern if the phases of the dipoles differ by $\lambda/4$?

11. Suppose you have a single dipole at $\lambda/4$ in front of a perfectly conducting plate. Determine the far field radiation pattern. Compare this to the result of problem 10.

Chapter 7
Practical Aspects of Filled Aperture Antennas

7.1 Descriptive Antenna Parameters

In general, most antenna systems, especially those with high gain and directivity used in radio astronomy and communications must be analyzed using detailed numerical models such as GRASP. The most important antennas used in these applications are fully steerable paraboloids whose properties are treated at greater length in the monograph by Baars (2007). If one wants an accurate but rather simple description of antenna properties, one must use the concepts presented in the following Sections, which allow one to characterize the antenna properties based on astronomical measurements. In the following we provide details of realistic antennas, starting with some necessary details.

7.1.1 The Power Pattern $P(\vartheta, \varphi)$

Often, the *normalized power pattern*, not the power pattern is measured:

$$\boxed{P_{\mathrm{n}}(\vartheta, \varphi) = \frac{1}{P_{\max}} P(\vartheta, \varphi)} \quad . \tag{7.1}$$

The reciprocity theorem provides a method to measure this quantity. The radiation source can be replaced by a small diameter radio source. The flux densities of such sources are determined by measurements using horn antennas at centimeter and millimeter wavelengths. At short wavelengths, one uses planets, or moons of planets, whose surface temperatures are determined from infrared data.

If the power pattern is measured using artificial transmitters, care should be taken that the distance from antenna A to antenna B is so large that B is in the far radiation field of A. This requires that the curvature of a wavefront emitted by B is much less than a wavelength across the geometric dimensions of A. From geometry, this curvature must be $k \ll 2D^2/\lambda$, for an antenna of diameter D and a wavelength λ.

Consider the power pattern of the antenna used as a transmitter. If the total spectral power, $\mathscr{P}_\nu$ in [W Hz^{-1}] is fed into a lossless isotropic antenna, this would transmit P power units per solid angle per Hertz. Then the total radiated power at frequency ν is $4\pi P_\nu$ In a realistic, but still lossless antenna, a power $P(\vartheta,\varphi)$ per unit solid angle is radiated in the direction (ϑ,φ). If we define the directive gain $G(\vartheta,\varphi)$ as the

$$P(\vartheta,\varphi) = G(\vartheta,\varphi)P$$

or

$$G(\vartheta,\varphi) = \frac{4\pi P(\vartheta,\varphi)}{\iint P(\vartheta,\varphi)\,d\Omega} \quad . \tag{7.2}$$

Thus the gain or directivity is also a normalized power pattern similar to (7.1), but with the difference that the normalizing factor is $\int P(\vartheta,\varphi)\,d\Omega/4\pi$. This is the gain relative to a lossless isotropic source. Since such an isotropic source cannot be realized in practice, a measurable quantity is the gain relative to some standard antenna such as a half-wave dipole whose directivity is known from theoretical considerations.

7.1.2 The Main Beam Solid Angle

The *beam solid angle* Ω_A of an antenna is given by

$$\Omega_A = \iint\limits_{4\pi} P_n(\vartheta,\varphi)\,d\Omega = \int\limits_0^{2\pi}\int\limits_0^{\pi} P_n(\vartheta,\varphi)\sin\vartheta\,d\vartheta\,d\varphi \tag{7.3}$$

this is measured in steradians (sr). The integration is extended over the full sphere 4π, such that Ω_A is the solid angle of an ideal antenna having $P_n = 1$ for all of Ω_A and $P_n = 0$ everywhere else. Such an antenna does not exist; for most antennas the (normalized) power pattern has considerably larger values for a certain range of both ϑ and φ than for the remainder of the sphere. This range is called the main beam or main lobe of the antenna; the remainder are the side lobes or back lobes (Fig. 7.1). For actual situations, the properties are well defined up to the shortest operating wavelengths. At the shortest wavelength, there is indeed a main beam, but much of the power enters through sidelobes. In addition, the main beam efficiency may vary significantly with elevation. Thus, the ability to accurately calibrate the radio telescope at the shortest wavelengths may be challenging.

In analogy to (7.3) we define the *main beam solid angle* Ω_{MB} by

$$\Omega_{MB} = \iint\limits_{\substack{\text{main}\\\text{lobe}}} P_n(\vartheta,\varphi)\,d\Omega \quad . \tag{7.4}$$

Fig. 7.1 A polar power pattern showing the main beam, and near and far side lobes. The weaker far side lobes have been combined to form the stray pattern
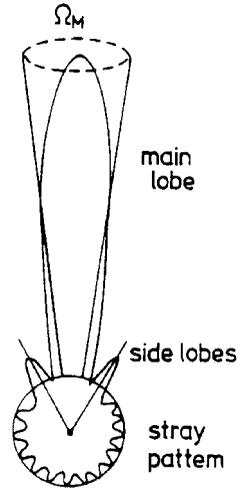
The quality of an antenna as a direction measuring device depends on how well the power pattern is concentrated in the main beam. If a large fraction of the received power comes from the side lobes it would be rather difficult to determine the location of the radiation source, the so-called "pointing".

It is appropriate to define a *main beam efficiency* or (usually) *beam efficiency*, η_B, by

$$\eta_B = \frac{\Omega_{MB}}{\Omega_A} \qquad . \tag{7.5}$$

The main beam efficiency is not related to the angular size of the main beam. A small antenna with a wide main beam can have a high beam efficiency: η_B is an indication of the fraction of the power is concentrated in the main beam. The main beam efficiency can be modified (within certain limits) for parabolic antennas by a choice grading function (6.58) of the main reflector. This can easily be accomplished by the choice of primary feeds and foci. If the FWHP beamwidth is well defined, the location of an isolated source is determined to an accuracy given by the FWHP divided by the S/N ratio. Thus, it is possible to determine positions to a small fraction of the FWHP beamwidth.

Substituting (7.3) into (7.2) it is easy to see that the maximum directive gain G_{max} or *directivity* $\mathscr{D}$ can be expressed as

$$\mathscr{D} = G_{max} = \frac{4\pi}{\Omega_A} \qquad . \tag{7.6}$$

The angular extent of the main beam is usually described by the *half power beam width* (HPBW), which is the angle between points of the main beam where the normalized power pattern falls to $1/2$ of the maximum (Fig. 7.2). This is also referred to as the *full width to half power* (FWHP). Less frequently used definitions are the
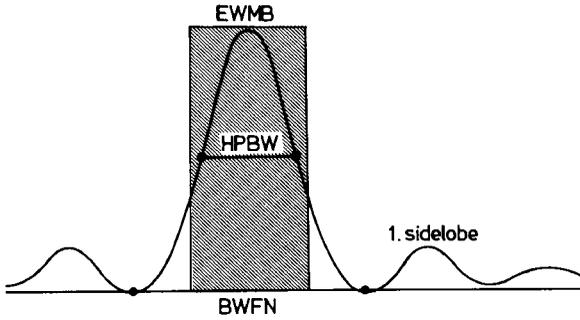
Fig. 7.2 A sketch of the telescope beamwidth, together with commonly used measurements of beam size for a one-dimensional power pattern. EWMB is the equivalent width of the (*full*) half power beam width. The HPBW is sometimes referred to as FWHP, or full width to half power. BWFN denotes the beam width between first nulls. This is indicated by the two dots

beam width between first nulls (BWFN) or the *equivalent width of the main beam* (EWMB). The latter quantity is defined by

$$\text{EWMB} = \sqrt{\frac{12}{\pi}\Omega_{\text{MB}}} \; . \tag{7.7}$$

For elliptically shaped main beams, values for widths in orthogonal directions are needed. The beamwidth is related to the geometric size of the antenna and the wavelength used; the exact beamsize depends on grading functions and illumination.

7.1.3 The Effective Aperture

Let a plane wave with the power density $|\langle S \rangle|$ be intercepted by an antenna. A certain amount of power is then extracted by the antenna from this wave; let this amount of power be P_{e}. We will then call the fraction

$$A_{\text{e}} = P_{\text{e}} / |\langle S \rangle| \tag{7.8}$$

the *effective aperture* of the antenna. A_{e} is a quantity very much like a cross-section in particle physics, A_{e} has the dimension of m^2. Comparing this to the *geometric aperture* A_{g} we can define an aperture efficiency η_{A} by

$$\boxed{A_{\text{e}} = \eta_{\text{A}} A_{\text{g}}} \quad . \tag{7.9}$$

For some antennas, such as the Hertz dipole there is no clearly defined geometric aperture; in such cases there is no simple expression for the aperture efficiency η_{A}. For a calculation of the effective aperture, the peak value of A_{e} is used; this is the direction of the telescope axis. Directivity is related to A_{e} by

$$\mathscr{D} = G_{\text{max}} = \frac{4\pi A_e}{\lambda^2}$$

(7.10)

which according to (7.6) is equivalent to

$$A_e \, \Omega_A = \lambda^2 \quad .$$

(7.11)

Often derivations of (7.10) or (7.11) are given by computing $\mathscr{D}$ and A_e for some simple antennas and then generalizing the result. To obtain a result that is generally applicable, we follow the derivation of (7.10) given by Pawsey and Bracewell (1954) which makes use of thermodynamic considerations. Let antenna, receiver and a radiating surface C all be enclosed by a black body at the temperature T. Let us assume thermodynamic equilibrium for the whole system. Then the antenna will radiate power into the black enclosure, and this power will be absorbed there. The black body will also radiate, and part of this radiation will be received by the antenna. Let the radiation surface C subtend the solid angle Ω_A as seen from the antenna (Fig. 7.3), whose directivity is $\mathscr{D}$, effective aperture A_e and receiver bandwidth Δv. According to the Rayleigh-Jeans relation, the surface C radiates with the intensity

$$I_v = \frac{2kT}{\lambda^2} \Delta v$$

in units of $\text{W m}^{-2}\text{Hz}^{-1}$ per unit solid angle. Then the antenna collects a total power of

$$W = A_e \frac{kT}{\lambda^2} \Delta v \, \Omega_A$$

(7.12)

since, according to (7.8) only one polarization component is recorded.
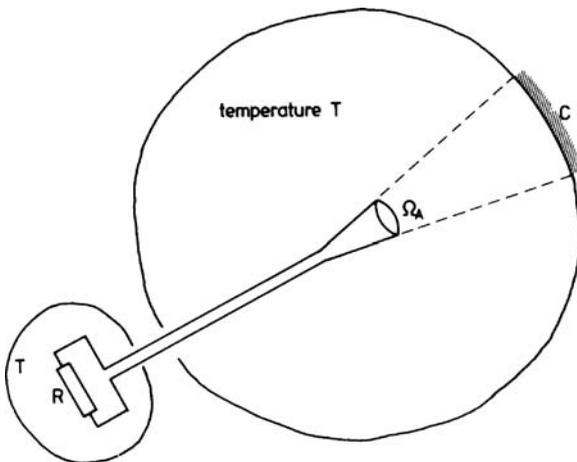
Fig. 7.3 A sketch to illustrate the relation between effective aperture and directivity

If the whole system is in thermal equilibrium, the principle of detailed balance holds. Then the antenna must reradiate the same amount of power that it receives. If the antenna terminals are connected by a matched resistor R, then the transmitted power according to the Nyquist theorem (1.42) is

$$L\Delta\nu = kT\,\Delta\nu\,.$$

According to the definition (7.6) the fraction

$$L'\Delta\nu = kT\,\Delta\nu\,\mathscr{D}\,\frac{\Omega_{\mathrm{A}}}{4\pi} \tag{7.13}$$

is intercepted by the surface C. As stated, relation (7.12) and (7.13) are equal if thermodynamic equilibrium prevails; thus

$$A_{\mathrm{e}}\,\frac{kT}{\lambda^2}\,\Delta\nu\,\Omega_{\mathrm{A}} = kT\,\Delta\nu\,\mathscr{D}\,\frac{\Omega_{\mathrm{A}}}{4\pi}$$

so that

$$\mathscr{D} = \frac{4\pi A_{\mathrm{e}}}{\lambda^2}\,.$$

Although this relation has been derived under the assumption of thermodynamic equilibrium, this relates quantities which do not involve thermodynamics, so will always be valid.

7.1.4 The Concept of Antenna Temperature

Consider a receiving antenna with a normalized power pattern $P_{\mathrm{n}}(\vartheta,\varphi)$ that is pointed at a brightness distribution $B_\nu(\vartheta,\varphi)$ in the sky. Then at the output terminals of the antenna, the total power per unit bandwidth, $\mathscr{P}_\nu$ is

$$\mathscr{P}_\nu = \tfrac{1}{2} A_{\mathrm{e}} \iint B_\nu(\vartheta,\varphi)\,P_{\mathrm{n}}(\vartheta,\varphi)\,\mathrm{d}\Omega\,. \tag{7.14}$$

By definition, we are in the Rayleigh-Jeans limit, and can therefore exchange the brightness distribution by an equivalent distribution of brightness temperature. Using the Nyquist theorem (1.42) we can introduce an equivalent *antenna temperature* T_{A} by

$$\mathscr{P}_\nu = kT_{\mathrm{A}}\,. \tag{7.15}$$

This definition of *antenna temperature* relates the output of the antenna to the power from a matched resistor. When these two power levels are equal, then the antenna temperature is given by the temperature of the resistor. Instead of the effective aperture A_{e} we can introduce the beam solid angle Ω_{A}, from (7.11). Then (7.14) becomes

$$T_A(\vartheta_0, \varphi_0) = \frac{\int T_b(\vartheta, \varphi) P_n(\vartheta - \vartheta_0, \varphi - \varphi_0) \sin \vartheta \, d\vartheta \, d\varphi}{\int P_n(\vartheta, \varphi) \, d\Omega} \qquad (7.16)$$

which is the *convolution* of the brightness temperature with the beam pattern of the telescope. The brightness temperature $T_b(\vartheta, \varphi)$ corresponds to the thermodynamic temperature of the radiating material only for thermal radiation in the Rayleigh-Jeans limit from an optically thick source; in all other cases T_b is only an convenient quantity that in general depends on the frequency.

The quantity T_A in (7.16) was obtained for an antenna with no ohmic losses, and no absorption in the earth's atmosphere. In terms of the definitions in Sect. 8.2.5, the expression T_A in (7.16) is actually T_A', that is, a temperature corrected for atmospheric losses. We will use the term T_A' in Sect. 8.2.5. Since T_A is the quantity measured while T_b is the one desired, (7.16) must be inverted. (7.16) is an integral equation of the first kind, which in theory can be solved if the full range of $T_A(\vartheta, \varphi)$ and $P_n(\vartheta, \varphi)$ are known. In practice this inversion is possible only approximately. Usually both $T_A(\vartheta, \varphi)$ and $P_n(\vartheta, \varphi)$ are known only for a limited range of ϑ and φ values, and the measured data are not free of errors. Therefore usually only an approximate deconvolution is performed. A special case is one for which the source distribution $T_b(\vartheta, \varphi)$ has a small extent compared to the telescope beam. Given a finite signal-to-noise ratio, the best estimate for the upper limit to the actual FWHP source size is one half of the FWHP of the telescope beam. This will be described further in Chap. 8, where the steps necessary to calibrate an antenna are discussed.

7.2   Primary Feeds

In the preceding paragraphs we indicated how the antenna pattern depends on the current grading across the aperture, but we have not specified how this grading is achieved in practical situations. Generally a receiving antenna can be considered as a device which transforms an electromagnetic wave in free space into a guided wave. The reflector transforms the plane wave into a converging spherical wave. The primary feed accepts this converging wave and transforms its characteristics so that the power will reach the receiver.

For a successful antenna design many aims have to be met simultaneously; some of these may be contradictory, so that one can only be fulfilled at the expense of others. Antenna design therefore is more empirical than analytical; there is no general theory that covers all aspects simultaneously. Provided the current distribution $J(x')$ is given, the vector potential $A(x)$ and thus the electromagnetic fields $E(x)$ and $H(x)$ or $B(x)$ can be computed. However, these induce currents J so that we have the problem of self-consistency. These complications make rigorous analytic solutions so difficult to obtain that Sommerfeld's 1896 rigorous analytical solution of the diffraction of a plane wave by a perfectly conducting semi-infinite screen has not been markedly improved after 100 years. Methods involving numerical solutions are

therefore necessary. From the reciprocity theorem, the parameters of a given antenna are identical if used for transmission or for reception. Some concepts are more easily visualized if a receiving situation is assumed, while others are best understood in terms of transmitting antennas. A discussion of current grading as influenced by the feed is best given for the case of transmission. Thus we consider the primary feed of a transmitting antenna. At point x' of the reflector, the power will induce a surface current depending on the amplitude of the oscillating field strength. If the primary feed is sufficiently far from the reflector that far field conditions can be adopted, the relative distribution of the field strength (for both electric and magnetic fields) can be computed from the normalized power pattern of the feed and, in most cases, this can be used at least as a first approximation.

7.2.1 Prime Focus Feeds: Dipole and Reflector

As an introduction, we first discuss a simple feed which is no longer in common use. The simplest feed is formed by a short dipole in front of a reflecting disk $\lambda/4$ behind the dipole. Such designs were frequently used as primary feeds for deep parabolic dishes with a small f/D ratios ($f =$ focal length, $D =$ diameter). Simple dipoles are not very efficient primary feeds. Main reflectors with $f/D = 0.25$ the prime focus must illuminate the half sphere, 2π. There are very few other designs that can illuminate such a large angle, so dipoles are sometimes used. For reflectors with f/D ratios larger than 0.3, simple dipole disk feeds produce large spillover losses. In some cases, the illumination angle has been adapted to the reflector by using dipole arrays as prime feeds, but this is usually done only if the central focus position is occupied by a waveguide feed for some other frequency. A simple dipole feed is sensitive only to linear polarization with the electric field strength directed parallel to the dipole. The greatest disadvantage of a dipole feed is the nonuniform illumination. This results in a non circular main beam of the telescope (see Fig. 7.4). Since the phase of the electromagnetic waves vary rather strongly across the aperture, both the aperture efficiency and the beam efficiency of dipole-disk feeds are rather low.

7.2.2 Horn Feeds Used Today

The electric and magnetic field strengths at the open face of a wave guide will vary across the aperture. The power pattern of this radiation depends both on the dimension of the wave guide in units of the wavelength, λ, and on the mode of the wave. The greater the dimension of the wave guide in λ, the greater is the directivity of this power pattern. However, the larger the cross-section of a wave guide in terms of the wavelength, the more difficult it becomes to restrict the wave to a single mode. Thus wave guides of a given size can be used only for a limited frequency range.

Fig. 7.4 Equivalent current distribution in the aperture plane of a dipole fed paraboloid. The numbers give the absolute value of the current density, phases are not indicated, except for the loci of phase reversal [after Heilmann (1970)]
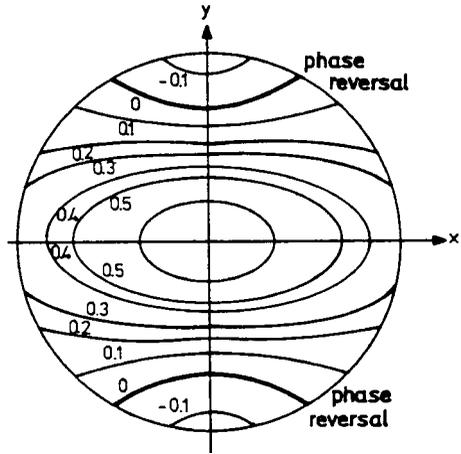
The aperture required for a selected directivity is then obtained by flaring the sides of a section of the wave guide so that the wave guide becomes a horn.

Simple pyramidal horns are usually designed to transmit only the lowest modes. However, then the electric and magnetic field strengths are distributed differently along the sides of the horn aperture.

Great advances in the design of feeds have been made since 1960, and most parabolic dish antennas now use hybrid mode feeds (Fig. 7.5). If a truly circular beam for an arbitrary polarization angle is wanted, more than TE modes are used; the electric field in the aperture has to be oriented in the direction of propagation. But then the conductivity of the horn in this direction has to be zero. This is achieved by constructing the walls of the circular horn from rings that form periodic structures

Fig. 7.5 A corrugated waveguide hybrid mode feed for λ 2 mm for the IRAM Plateau de Bure interferometer (courtesy of B. Lazareff)

which have a characteristic size of $\lambda/4$. For such corrugated horns, the theory is to be found in Love (1976). These are also referred to as *Scalar* or *Multi-Mode* feeds. Such feed horns are used on all parabolic antennas. These provide much higher efficiencies than simple single mode horn antennas and are well suited for polarization measurements.

At centimeter wavelengths, a more recent development is the use of phased arrays of individual feeds to synthesize a beam that has characteristics that are superior to that of the individual components. In Sect. 6.3.1, a scheme using a simplified description of individual elements was presented. This has been extended to the use of feed horns in the centimeter wavelength range, at a larger cost. In this scheme, after each feed horn there is a cooled receiver system, and the outputs of each receiver system are combined using phase shifters to produce a series of beams. The advantage of this method is that the properties of the beams can be varied until the desired result is obtained. With extra complexity, multiple beams can be produced.

7.2.3 Multiple Reflector Systems

If the size of a radio telescope is more than a few hundred wavelengths, designs similar to those of optical telescopes are preferred. For such telescopes Cassegrain, Gregorian and Nasmyth systems have been used. In a Cassegrain system, a convex hyperbolic reflector is introduced into the converging beam immediately in front of the prime focus. This reflector transfers the converging rays to a secondary focus which, in most practical systems is situated close to the apex of the main dish. A Gregorian system makes use of a concave reflector with an elliptical profile. This must be positioned behind the prime focus in the diverging beam. In the Nasmyth system this secondary focus is situated in the elevation axis of the telescope by introducing another, usually flat, mirror. The advantage of a Nasmyth system is that the receiver front ends remain horizontal while when the telescope is pointed toward different elevations. This is an advantage for receivers cooled with liquid helium, which become unstable when tipped.

There are several reasons why secondary focus systems are useful in radio astronomy. In small telescopes the weight of the secondary reflector is much less than that of receiver front ends, especially if these must be cooled. In addition, these are usually more easily mounted and are more accessible at the apex. The effective focal ratio f/D of Cassegrain or Gregorian systems (Fig. 7.6) is usually 5–10 times larger than that of primary focus systems. Then optical distortions such as coma are negligible for much larger fields than in primary focus configurations. For such foci several receiving systems can be placed at different positions, including some far off axis. In this way, systems at a number of different frequencies or an array of receivers for the same frequency range can be accommodated in the secondary focus.

Finally, it is much easier to build low noise systems using such a design. High aperture efficiency requires a current grading with a good illumination up to the
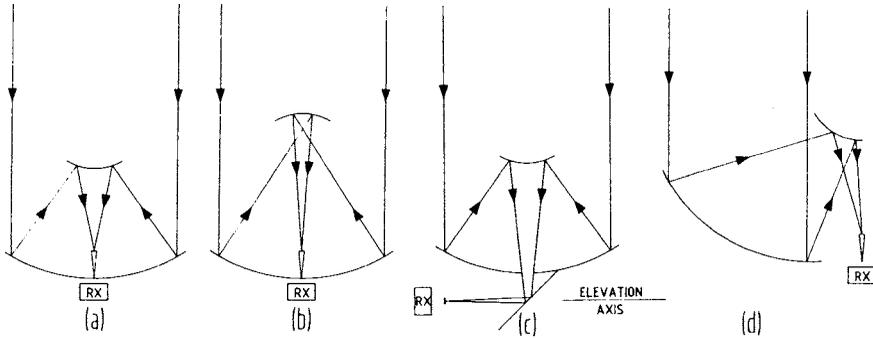
Fig. 7.6 The geometry of (**a**) Cassegrain, (**b**) Gregory, (**c**) Nasmyth and (**d**) offset Cassegrain systems

edge of the dish. If, however, in a prime focus configuration, the spillover side-lobe pattern of the feed extends beyond the edge of the dish, the feed will also receive 300 K thermal radiation of the ground. In a system with a secondary reflector, the power received by the feed from beyond the edge of the secondary reflector is radiation from the sky, which has a temperature of only a few K. For low-noise systems, this results in an overall system noise temperature that is significantly less than for prime focus systems, in which the power is received from the ground. This can be quantified in the so-called "G/T value", the ratio of the gain of the antenna to the system noise. Any telescope design must aim to minimize the excess noise at the receiver input while maximizing gain. For a specific antenna, this maximization may involve the design of feeds and the choice of foci. Naturally secondary focus systems also have disadvantages. While the angle that the telescope disk subtends as seen from the prime focus is usually between $100°$ and $180°$, a secondary reflector usually subtends only 10–15°. Then the secondary focus feed horns must have much larger directivity and consequently greater dimensions. The dimensions increase directly with wavelength and therefore there is usually a lower limit for frequencies measured at the secondary focus. For the Effelsberg 100 m dish secondary focus this limit is near $f = 2.3$ GHz. At this frequency, the primary feed horns have aperture diameters of 1.5 m and overall lengths of 3 m. Thus, at longer wavelengths, the prime focus must be used. At shorter wavelengths, the secondary focus can be used and is preferred. First, the Field of View, that is, the region of the sky that can be accurately measured, is larger. Second, the noise due to the reception of radiation from the ground is smaller. Third, it is possible to correct for large scale deformations of the primary reflector by deforming the subreflector appropriately.

   That the secondary reflector blocks the central parts in the main dish from reflecting the incoming radiation causes some interesting differences between the actual beam pattern from that of an unobstructed telescope. For the simple case of a circular annular aperture antenna with uniform illumination ($g \approx 1$) and an inner/outer radius of $d/2$ and $D/2$ the normalized illumination pattern can be computed by an expression similar to (6.69), when the lower limit of the integral, $d/2\lambda$ is used. If the blocking factor is given by

$$b = d/D$$

then the resulting normalized beam pattern is

$$P_{\mathrm{n}}(u, D, b) = \left[\frac{J_1(\pi u D/\lambda) - b J_1(\pi u b D/\lambda)}{\pi u D (1 - b^2)/2\lambda} \right]^2 \quad . \tag{7.17}$$

The main differences between this result, compared to the beam pattern of an unobstructed dish, are (1) the increased level of the first side lobe for a finite value for b and (2) a slightly lower angular resolving power. Effect 2 can be understood intuitively by considering that (7.17) is formed by subtracting the voltage produced by a circular reflector with the diameter bD from that of one with diameter D. For small values of u these contributions are in phase, while for larger u the phases may differ. Therefore those contributions that form the main beam will always be diminished by this process, while the contributions to the outer side lobes can have any phase. Normalizing the main beam contribution to unity therefore will increase the side-lobe level.

Realistic filled aperture antennas will have a beam pattern different from (7.17) for several reasons. First the illumination of the reflector will not be uniform but has a taper by 10 dB or more at the edge of the reflector. As seen in Table 6.1 the side-lobe level is strongly influenced by this taper, and aperture blocking will again produce a different result. Second, the secondary reflector must be supported by three or four support legs, which will produce aperture blocking and thus affect the shape of the beam pattern. In particular feed leg blockage will cause deviations from circular symmetry. For altitude-azimuth telescopes these side lobes will change position on the sky with hour angle. This may be a serious defect, since these effects will be significant for maps of low intensity regions if the main lobe is near an intense source. The side lobe response can also dependent on the polarization of the incoming radiation. Telescopes that employ only a primary focus will suffer the same effects, since the primary focus has to be supported by wide support legs if there is a massive prime focus to accommodate receivers. Such blocking is usually larger than for telescopes with receivers only in the secondary foci. The Effelsberg 100 m telescope has both foci, prime and Gregory, and four rather wide legs. The geometric blockage is 17%, a large value. Calculations show that the minimum blockage which might be achieved for an Effelsberg-type design is $\sim 7\%$.

Another disadvantage of on-axis systems, regardless of focus, is that they are often more susceptible to instrumental frequency baselines, so-called *baseline ripples* across the receiver band than primary focus systems. Part of this ripple is caused by multiple reflections of noise from source or receiver in the antenna structure. Ripples can arise in the receiver, but these can be removed or compensated rather easily. Telescope baseline ripples are more difficult to eliminate: it is known that large amounts of blockage and larger feed sizes lead to large baseline ripples. The effect is discussed in somewhat more detail in Sect. 8.4. This effect depends on many details, so can only be handled by experience. The influence of baseline

ripples on measurements can be reduced to a limited extent by appropriate observing procedures. A possible solution is the construction of off-axis systems. There are four reasons for favoring off axis telescopes. These have: (1) baseline ripples with smaller amplitudes, (2) lower side lobes and thus higher antenna and beam efficiencies, (3) higher G/T values and (4) since the sidelobe levels are lower, there will be less interference entering through sidelobes. However, there are drawbacks: (1) the reflecting surfaces must have a more complex shape, without axial symmetry, and (2) the polarization properties of the radiation in off-axis designs are also complex. However, such off-axis systems with active surface adjustment have great advantages.

7.3   Antenna Tolerance Theory

When the relation between aperture illumination and antenna pattern was derived in Sect. 6.4, the aperture was assumed to be a plane of constant phase. If there are deviations, some of the results must be modified. The modifications caused by phase variations across the aperture are the subject of this section. Most results will only be stated qualitatively; a more detailed treatment can be found in textbooks on antenna theory.

It is convenient to distinguish several different kinds of phase errors in the current distribution across the aperture of a two-dimensional antenna.

1) A phase error that varies linearly along some direction across the aperture is treated most simply by defining a new aperture plane oriented such that the phase remains constant. All directions then have to be measured relative to this new aperture plane. A linear phase error therefore results only in a tilt of the direction of the main beam.

2) A phase error which varies quadratically across the aperture is more complex. A treatment of this requires the introduction of Fresnel integrals, which describe the conditions of the electromagnetic field in a slightly defocussed state. We will not discuss this further here, but such errors can be avoided by properly focusing the telescope.

3) A third class of phase errors is caused by the fabrication tolerances of the reflector; such errors are avoidable only to some extent. The theoretical shape of the reflector can be approached only up to some finite tolerance ε. This will cause a phase error

$$\delta = 4\pi \frac{\varepsilon}{\lambda} \tag{7.18}$$

in the aperture plane. If δ is measured in radians, ε is the displacement of the reflector surface element in the direction of the wave propagation. We will discuss this error in some detail.

The current grading in the aperture plane according to (6.53) can then be written as

$$g(x) = g_0(x) e^{i\delta(x)}, \qquad g_0 \text{ real}. \tag{7.19}$$

The directivity gain of the reflector is, according to (7.2), (6.56) and (6.57)

$$G = \frac{4\pi}{\lambda^2} \frac{\left| \iint_{\mathscr{A}} g_0(x) e^{-i[k\mathbf{n}\cdot\mathbf{x}-\delta(x)]} d^2x \right|^2}{\iint_{\mathscr{A}} g_0^2(x) d^2x}. \tag{7.20}$$

Assuming that δ is small, the exponential function in (7.20) can be expanded in a power series including terms up to the second order

$$e^{i\delta} = 1 + i\delta - \frac{1}{2}\delta^2 \dots .$$

The ratio of the directivity gain of a system with random phase errors δ to that of an error-free system G_0 of identical dimensions then becomes

$$\frac{G}{G_0} = 1 + \bar{\delta}^2 - \overline{\delta^2}, \tag{7.21}$$

where

$$\bar{\delta} = \frac{\iint_{\mathscr{A}} g_0(x)\delta(x) d^2x}{\iint_{\mathscr{A}} g_0(x) d^2x} \tag{7.22}$$

and

$$\overline{\delta^2} = \frac{\iint_{\mathscr{A}} g_0(x)\delta^2(x) d^2x}{\iint_{\mathscr{A}} g_0(x) d^2x}. \tag{7.23}$$

$\bar{\delta}$ is the illumination weighted mean phase error. By selecting a suitable aperture plane, we can always force $\bar{\delta}$ to be zero. Then only the illumination weighted mean square phase error remains. This results in

$$\frac{G}{G_0} = 1 - \overline{\delta^2}, \tag{7.24}$$

For practical applications this series expansion has two drawbacks:

1) it is valid only for small δ while phase errors of $\delta \gtrsim 1$ and even larger will occur if antennas are used near their short wavelength limit, and

2) a more sophisticated antenna tolerance theory is needed because the phase errors $\delta(\boldsymbol{x})$ are not completely independent and randomly distributed across the aperture.

This second effect is the result of the following practical considerations. If at some point $\delta < 0$, chances are great that δ is also < 0 in an area surrounding this point. The reason for this is that the reflecting surface is smooth and has a certain stiffness, so that some kind of correlation distance for the phase errors has to be introduced. If this correlation distance d is of the same order of magnitude as the diameter of the reflector, part of the phase error can be treated as a systematic phase variation, either a linear error resulting only in a tilt of the main beam, or in a quadratic phase error which could be largely eliminated by refocussing. For $d \ll D$ the phase errors are almost independently distributed across the aperture, while for intermediate cases according to a good estimate for the expected value of the RMS phase error is given by [Ruze (1952, 1966)]

$$\bar{\delta}^2 = \left(\frac{4\pi\varepsilon}{\lambda}\right)^2 \left[1 - \exp\left\{-\frac{\Delta^2}{d^2}\right\}\right] \quad , \tag{7.25}$$

where Δ is the distance between two points in the aperture that are to be compared and d is the correlation distance. The gain of the system now depends both on $\bar{\delta}^2$ and on d. In addition, there is a complicated dependence both on the grading of the illumination and on the manner in which δ is distributed across the aperture. Ruze has given several approaches to such a theory. All of these lead to results which are basically similar. These results can be expressed by stating that the gain of a reflector containing phase errors can be approximated by an expression

$$G(u) = \eta\, e^{-\bar{\delta}^2} \left(\frac{\pi D}{\lambda}\right)^2 \lambda_1^2 \left(\frac{\pi D u}{\lambda}\right) + (1 - e^{-\bar{\delta}^2}) \left(\frac{2\pi d}{\lambda}\right)^2 \lambda_1^2 \left(\frac{2\pi d u}{\lambda}\right), \tag{7.26}$$

where

η is the aperture efficiency,
$u = \sin\vartheta$,
$\lambda_1(u) = \frac{2}{u} J_1(u)$ is the Lambda function,
D the diameter of the reflector, and
d the correlation distance of the phase errors.

There are now two contributions to the beam shape of the system. The first is that of a circular aperture with a diameter D as given by (6.71) but reduced in sensitivity due to the random phase error δ as given by (7.24). The second term is the so-called *error beam*. This can be described as equal to the beam of a (circular) aperture with a diameter $2d$, its amplitude multiplied by

$$(1 - e^{-\bar{\delta}^2})$$

The error beam contribution therefore will decrease to zero as $\bar{\delta} \to 0$.

The gain of a filled aperture antenna with phase irregularities δ cannot increase indefinitely with increasing frequency but reaches a maximum at $\lambda_m = 4\pi\varepsilon$, and this gain is 4.3 dB below that of an error-free antenna of identical dimensions. Then, if the frequency can be determined at which the gain of a given antenna attains its maximum value, the RMS phase error and the surface irregularities ε can be measured electrically. Experience with many radio telescopes shows reasonably good agreement of such values for ε with direct measurements, giving empirical support for the Ruze tolerance theory. The effect of surface errors on antenna efficiency is shown in Fig. 7.7. A plot of gain versus efficiency for a number of antennas is shown in Fig. 7.8.

Fig. 7.7 Aperture efficiency η_A (——) and beam efficiency η_{MB} ($- - -$) for different values of K in Table 6.1. The values for both an ideal reflector ($\delta = 0$) and one that introduces random phase errors of $\delta = 0.04\lambda$ are given [after Nash (1964)]
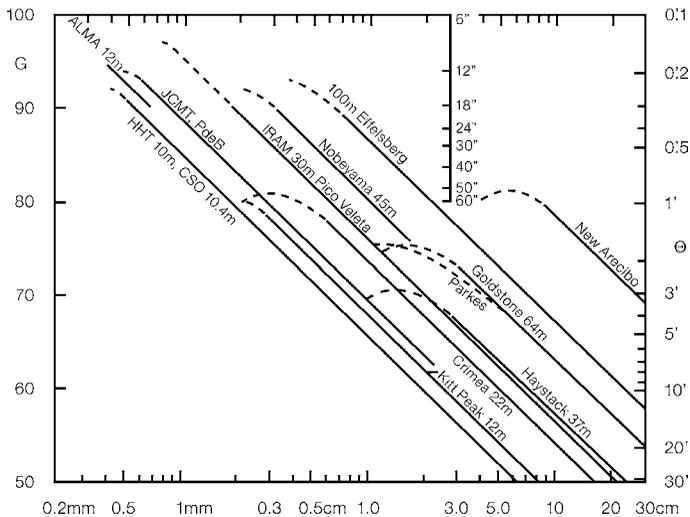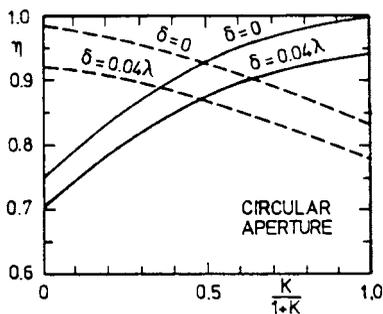
Fig. 7.8 The gain G in dB (*left* axis) of some high-precision filled aperture radio telescopes is plotted against wavelength λ. The Full Width to Half Power, θ is plotted on the *right* axis. The ALMA 12 m antennas will have the same properties as the APEX antenna. The curve for these antennas extends to 0.3 mm, and to longer wavelengths, where it joins the curve for the Kitt Peak 12 m antenna which has a lower surface accuracy. The 15 m diameter JCMT (James Clerk Maxwell Telescope) on Mauna Kea and PdeB (IRAM Plateau de Bure interferometer) antennas are also shown. Smaller but highly accurate antennas are a part of the SOFIA and Herschel observatories

7.4   The Practical Design of Parabolic Reflectors

7.4.1   General Considerations

Measurements of the *mechanical properties* of an antenna are of importance for its performance. This is especially true if the telescope deforms homologously. By *Homology* it is meant that, at various elevations, the main reflector deforms from one paraboloid into another. Today, homology is an intrinsic part of the design of all symmetric reflectors.

Non-homologous telescopes with diameters up to 12 m surfaces have been adjusted with the help of templates to accuracies of 100 μm. This is not possible for telescopes of larger diameter or if great precision is needed. Surveying and adjustment of the surface panels can be done by measuring the position of fiducial marks with surveying techniques. Previously one had used Invar tapes to measure distances along the surface and a theodolite to measure angles from the apex of the dish, gives sufficient precision. Other methods use laser ranging, as well as the theodolite. These methods have been replaced by *photogrametry*, in which optical measurements of small reflectors attached to theantenna surface are used to determine the positions of panels to accuracies of about 50 μm. Most recently, holographic methods have become popular. In such measurements, the usual source of coherent radiation is a signal transmitted from an earth satellite at 7 mm. A small telescope and the large dish to be measured receive the signal, which is then correlated, preserving both the relative phase and amplitude. The large dish is scanned, thus allowing an accurate measurement of amplitude and phase from the main beam and the side lobes. A second holographic method employs radiation from a giant planet such as Jupiter. This has been used to set the panels of the Caltech Submillimeter Observatory (CSO) 10 m dish to 17 μm. A third method, most commonly used today, is to use holography with a transmitter in the near field of the radio telescope. This is a more complex undertaking, since the near radiation field is important, and the distance to the transmitter must be accurately determined. The advantage in using this method is that it allows a very high S/N ratio. As a result of such near field measurements, the surface of the IRAM 30 m telescope has been set to a precision of 70 μm and that of the James Clark Maxwell Telescope (JCMT) to 50 μm.

Pointing errors rather than surface inaccuracies have usually set the ultimate limit to telescope performance. Due to the diurnal rotation of the earth, all celestial objects rotate about the celestial pole. Therefore for prolonged measurements in a given direction, a mounting that permits compensation for this motion has to be provided. Even for small reflectors a straightforward adaptation of the classical equatorial mount for optical telescopes is seldom used today. In the 1950s, this mount was commonly used before adequate digital control systems were available. In an equatorial mount the telescope is turned with constant angular velocity around a polar axis which is parallel to the earth's axis of rotation. Different declinations can be reached by tilting the reflector about the declination axis, which is perpendicular to the polar axis. The advantage of this design is the simplicity

of the resulting telescope control; as long as the telescope is aimed at a point with fixed celestial coordinates the telescope must rotate only about the polar axis with a constant angular velocity. However, this mount has the great disadvantage that the forces due to the weight of the telescope act on bearings at an arbitrary angle, and for the case of the declination axis these angles are always changing in the course of the diurnal motion. For these reasons all modern telescopes make use of an altitude-azimuthal mounting; the altitude-azimuth to equatorial coordinate transformation is carried out with a computer. The azimuthal axis is vertical, the elevation axis horizontal and both remain so even when the telescope is turning. The gravity load of the telescope acts either parallel or perpendicular to these axes and, when the telescope is tilted, the resulting gravitational force vector will always remain in a plane as seen from the telescope, while for the equatorial mounting this force vector can point to any direction within a hemisphere. For celestial positions which pass through the zenith, the azimuthal angular velocity becomes singular, so that no observations are possible in a region surrounding the zenith. The size of this region depends on the maximum possible speed for azimuth, but usually a field with a diameter of not more than 2–5° has to be avoided. For altitude-azimuth mountings, the relation between celestial and telescope coordinates is constantly changing, so the polarization angle and position angles of side lobes caused by feed legs also change as a source is moving across the sky. This may be turned to advantage since this effect can be used to eliminate a large part of these side lobe effects.

Pointing corrections for small dishes can be determined by mounting a small optical telescope approximately parallel to the axis of the telescope and tracking stars. Large homology dishes use radio measurements. Usually most of the errors remain constant over weeks or months and can be incorporated as constants in a pointing model in the telescope control computer program. Other errors, such as the collimation error, which depend on how the receiver currently in use is mounted in the telescope, must be determined more frequently.

Determinations of pointing constants are carried out by dedicated measurements. These constants are then stored in the control computer. If the control program includes such corrections, there will be compensations for known flexure of the telescope, changes in focus position, etc. Then a considerable precision of telescope pointing can be reached. In this way an RMS pointing error of $\lesssim 10''$ in both coordinates is obtained for the 100 m telescope at Effelsberg, if no special efforts for obtaining positional accuracies are taken. With frequent position checks, observing only at night and in low wind, an RMS error of $4''$ is possible. Even better performance has been reached with modern large single dishes used in the mm and sub-mm range.

The discussion in the preceding sections showed how the radio frequency characteristics of parabolic reflectors depend on the electric properties of the design and the precision with which it can be built. This must be supplemented by some remarks about limits based on mechanical and practical restrictions.

7.4.2 Specific Telescopes

The sole purpose of the telescope backup structure is to support the main reflector surface. This surface may be perforated sheet metal or wire mesh as long as the mesh size is about 1/16th of the shortest wavelength. This is acceptable if the intended wavelength is more than a few cm. For shorter wavelengths solid surfaces are needed. Usually the surface consists of panels that can be adjusted individually to the desired paraboloid shape. Attempts to include the surface into the support structure of the dish have not been successful.

When considerations of the costs are of prime importance, savings are often possible by restricting the range of the motion for the telescope. The prime example for this is the fixed 305 m spherical reflector at Arecibo, Puerto Rico (Fig. 7.9). The spherical main reflector is completely stationary and its mesh surface is constructed in a depression in the ground. The telescope beam is directed by moving the structure containing the feed horns and receivers. The shape of the reflector was chosen to be spherical. A sphere has no single focus, only a focal line (or *caustic*), and therefore a feed must be used to compensate for the spherical aberration. For many years so-called *line feeds* were used with Arecibo. In the last years, to simplify the optics, the secondary reflector was installed. This is a 26 m reflector which feeds the radiation to a tertiary Gregorian reflecting surface were placed above the
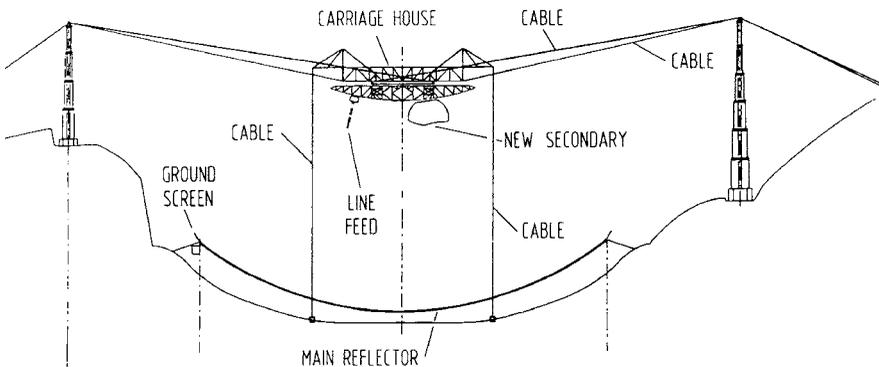
Fig. 7.9 A sketch of the Arecibo 305 m telescope in Puerto Rico, USA. The main reflector does not move; the telescope beam is steered by moving the *carriage house* suspended over the main reflector. On the left is the *line feed* used originally to correct the systematic phase errors caused by the spherical primary main reflector. The new arrangement to the right includes two reflectors which refocus the power from the main reflector and compensate for its large spherical aberrations. The 25 m diameter secondary and an 8 m diameter tertiary reflector (inside the structure marked "new secondary") direct the power to the receiver system. The receiver and the two reflectors are housed inside a radome on the right, below the carriage house. In addition to the new feed arrangement, a ground screen, 16 m high, was erected around the edge of the primary reflector (courtesy National Astronomy and Ionosphere Center). There are plans for an even larger Arecibo-type instrument in China. This would be one of the possible Square Kilometer Array (SKA) designs

primary reflector. With this, conventional horn feed illuminates approximately a 225 m aperture with very low spillover, high efficiency. The Arecibo antenna can track sources over a 20° range from zenith, enabling observations for declinations between 0 and +40°. A limiting frequency of about 10 GHz is the design goal. Arecibo has the largest collecting area of any radio telescope. There is a proposal from China to build a 500 m diameter version of Arecibo as a prototype for their version of the "Square Kilometer Array", SKA. In our presentation of antenna tolerance theory, it was shown that the shape of the reflecting surface must be $< \lambda/16$ in order to achieve a telescope gain within 67 % of an ideal reflector. For a 30 m radio telescope usable to $\lambda = 1$ mm the tolerance must be 0.06 mm RMS. This requirement must be met for all positions at which the telescope can be pointed. Any structure that is built from existing material will show flexure due to its own weight if tilted. This is caused by the finite maximum stress that the material can transmit, the modulus of elasticity and the density of the material used for the construction. The geometric shape introduces only a numerical factor; it cannot completely suppress the deformation. In order to obtain some idea of the size of these deformations it should be noted that the rim of the dish of the 100 m telescope at Effelsberg deforms by about 60 mm when the telescope is tilted from zenith to horizon. This should be compared to the required precision of about 0.5 mm, RMS, if the telescope is to be fully usable for $\lambda = 1$ cm. Closer scrutiny shows that what is needed is not the absence of deflections but only that the shape of the reflector remains a paraboloid of revolution; changes in both the shape and the position (apex, focal point and axis) of this paraboloid can be tolerated. Such deformations are called *homologous*, and it is imperative that only such deformations occur. The first large radio telescope designed specifically with a homology design was the 100 m telescope at Effelsberg (see Fig. 7.10). The success of this design was such that this telescope is usable at wavelengths as short as $\lambda = 3$ mm, since the error due to telescope deformations always remains smaller than 0.5 mm. More modern homology telescopes have a much stiffer design than Effelsberg, but the principle of homology has been incorporated in the design of symmetric dishes such as the IRAM 30 m. A larger version of the IRAM dish is the 50 m diameter Large Millimeter Telescope (LMT) located near Puebla, Mexico. The LMT is near completion.

One large telescope that does *not* have a homologous design is the GBT, since this has an asymmetric structure. Rather the individual panels of the GBT are equipped with an actuator system that adjusts the surface for maximum gain. In some sense one can compare the Effelsberg homology design to an analog computer, and the GBT system to a digital computer. The GBT design allows more freedom.

The use of homology insures that deformations due to gravity are compensated. The practical limits are now given by thermal and wind deformations. In order to minimize thermal deformations, and to protect the telescope from adverse weather conditions, these are often contained in radomes or astrodomes. Two examples are the 15 m James Clerk Maxwell Telescope (JCMT) and the 10 m CSO telescope, both at 4 km elevation on Mauna Kea, Hawaii. Another approach to reduce the effect of temperature fluctuations is to use structures made of low expansion materials. Since the surface tolerances are so critical, the support structure is often
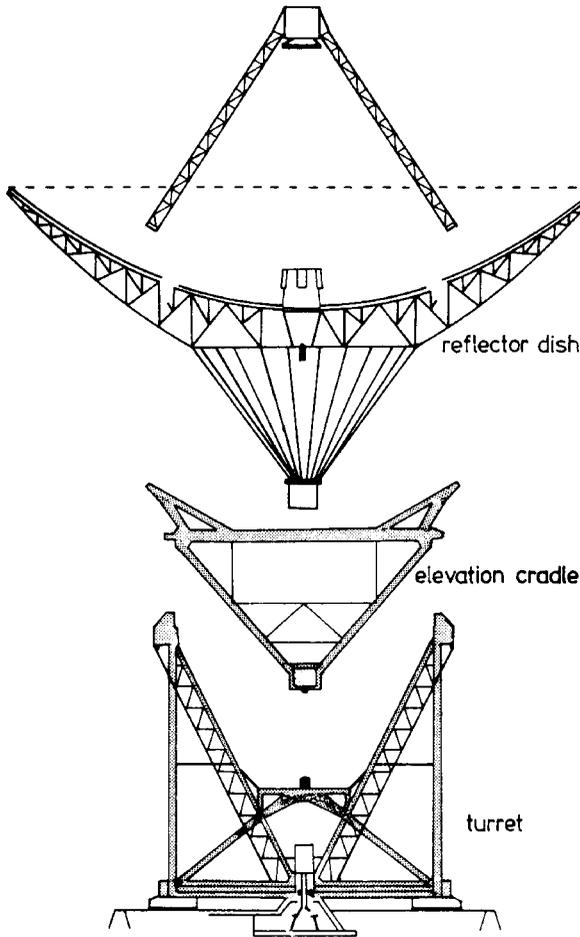
Fig. 7.10 A sketch showing the homology design principle of the 100 m-telescope at Effelsberg (after Grahl). The support for the surface is provided by a set of radial struts from the backup structure to the central hub. The support for the prime focus is provided by four legs. The surface is free to deform from one paraboloid to another. Such a design places strong constraints on the symmetry of the structure. For off-axis systems such as the GBT, active surface adjustment is used to replace the effect of homology

now made of a low expansion material, *carbon fiber reinforced plastic*, or CFRP. With CFRP, together with the Invar nodes joining the CFRP tubes, a space frame structure with a negligible temperature expansion coefficient over a fairly wide temperature range can be built. From this design, differences in temperature within the telescope structure will not influence the shape of the dish. The progress in telescope construction made in the last years may be seen by the fact that modern versions of telescopes do not require protection from weather in the form of an astrodome. The most remarkable example of such a design is the 12 m sub-mm telescope, the Atacama Pathfinder EXperiment (APEX). Although APEX operates at

a site of 5000 m elevation at Chajnantor/Chile. APEX, a telescope project operated jointly by MPIfR, ESO and the Onsala Space Observatory is now taking data. The individual antennas of the Atacama Large Millimeter/sub-mm Array (ALMA) are similar to APEX. The specifications of these instruments are: RMS surface accuracy better than 25 μm, pointing accuracy $2''$ over the whole sky, tracking accuracy $0.6''$. APEX and the ALMA dishes are the highest performance radio telescopes built. SOFIA (Fig. 7.12, see http://www.nasa.gov/mission_pages/SOFIA/index.html for a description) has a 2.5 meter reflector that has optical quality, while the Herschel Satellite Observatory has a 3.5 meter silicon-carbide reflector (Fig. 7.13, see http://sci.esa.int/science-e/www/area/index.cfm?fareaid=16 for more information).

For longer wavelength observations, equipment has to be placed near the position of the prime focus since feeds must have sizes of more than $\lambda/4$ to be efficient. In the prime focus, the telescope is illuminated by the primary feed. In order to avoid the losses of long transmission lines, the receiver front end is also mounted at the prime focus. At shorter wavelengths the secondary focus in Cassegrain, Gregory or Nasmyth foci can be used, a secondary reflector is then required. In symmetric telescopes both the supporting legs and the secondary reflector or the receiver cabin obstruct part of the aperture, giving rise to aperture blocking. Usually the loss of gain caused by blockage of the effective aperture is minor. Of much greater importance is the influence on the side-lobe level. A complete analytical treatment of this effect is rather difficult because of geometry, thus empirical estimates are used. Support structures with three and four legs have been used; the resulting side lobe structure shows a six-fold or a four-fold symmetry – at least to a first approximation.

Aperture blocking and all the problems connected with it, that is baseline ripple due to standing wave patterns of radiation reflected from the feed horn and the supporting legs (see Section 8.4.3), increased side-lobe level and an increased susceptibility to man-made interference, can be avoided if an off-axis construction is used. This is the design principle chosen for the new Green Bank Telescope, GBT (Fig. 7.11). The design of an offset paraboloid has, however, complications. Since the design has less symmetry homology is more difficult to achieve and therefore active, real time adjustments of the surface are needed if the design limit of 7 mm or perhaps even 3 mm wavelength is to be reached. This will be accomplished by an actuator system controlled in real time by a laser measuring system. But for a wavelength larger than 2 cm the GBT will not require active surface adjustment.

Considerations resulting in a high efficiency for the telescope are only some of the criteria in the design of radio telescopes. Of almost equal importance are features that result in an overall low-noise system. These refer mainly to the receiver design, but the telescope design can be important also, since not all the radiation that is received arises from the radio source at which the telescope is pointed. A large part of the signal, in some systems up to 50%, arises from the immediate telescope vicinity. This could be radiation from the ground, either leaking through the perforated reflecting surface or picked up by spillover lobes of the primary feed extending over the edge of the reflector dish. As already noted, the noise performance of Cassegrain or Gregorian systems is usually much better than that of prime focus systems because such spillover lobes. The influence of spillover noise can be decreased by

Fig. 7.11 The Green Bank Telescope (GBT) at West Virginia, U.S.A. The telescope is an off-axis paraboloid with a 110 m by 100 m diameter. This design was chosen in order to minimize side lobes and reflections in the telescope structure that lead to instrumental artefacts or "baseline ripples"

Fig. 7.12 The SOFIA facility will fly at about 12 km altitude, above most of the water in the earth's atmosphere. This will allow measurements in the sub-mm and far infrared range. The modified 747SP airplane contains a 2.5 m optical telescope. This is a joint project of NASA and the German Space Agency, DLR. The first test flights are now taking place

Fig. 7.13 The Herschel Space Observatory will observe at sub-mm and far infrared wavelengths with bolometer and heterodyne instruments. The antenna is a 3.5 m paraboloid made of silicon carbide. The structure behind the antenna is a sun shield; the structure below the main reflector is the cryostat containing the receiver systems. Herschel is equipped with two bolometer cameras, SPIRE (covering 250–520 µm) and PACS (covering 75–170 µm, and a single pixel heterodyne instrument, HIFI (covering 157–212 and 240–625 µm). The bolometers also have spectroscopic capability. Launch is planned for mid-2008

suitably placed screens that direct the side lobes towards cold sky and not ground. These have much the same purpose as baffle tubes in optical telescopes.

7.5 Summary

1) Fully steerable paraboloids of revolution have become the standard antennas in the centimeter, millimeter and sub-mm wavelength regions. In the mm and sub-mm ranges, these were housed in shelters, but recently designs have allowed high performance paraboloids to operate in the open air.
2) All symmetric paraboloids have designs that make use of the homology principle. That is, with changing elevation, the surfaces deform freely from one symmetric parabolic shape to another.
3) All modern designs use Altitude-Azimuth mounts. The control is carried out with digital computer systems.
4) For millimeter and sub-mm paraboloids, Carbon Fiber Reinforced Plastic (CFRP) rods are employed in the support structures and subreflector support legs. CFRP is needed to minimize the changes in structures due to changing temperatures.

5) The aim of all designs is to minimize the blockage of the primary reflector. This requires thin, "knife edge" subreflector support legs. Thus the weight of the sub-reflector must be kept to a minimum.

6) The Cassegrain and Nasmyth foci are preferred for millimeter and sub-mm telescopes, since these help to minimize the reception of noise from the ground, have additional reflecting surfaces that allow optimization of telescope gain, and provide larger amounts of space for receivers.

Problems

1. (a) Use Eqs. (7.3), (7.14) and (7.15) to show that for a source with an angular size $\ll$ the telescope beam, $T_A = S_v A_e/2k$. Use these relations and Eq. (7.16) to show that $T_A = \eta_B T_B$, where T_B is the observed brightness temperature.
(b) Suppose that a Gaussian-shaped source has an actual angular size θ_s and actual peak temperature T_0. This source is measured with a Gaussian-shaped telescope beam size θ_B. The resulting peak temperature is T_B. The flux density, S_v, integrated over the entire source, must be a fixed quantity, no matter what the size of the telescope beam. Use this argument to obtain a relation between temperature integrated over the telescope beam, T_B

$$T_B = T_0 \left(\frac{\theta_s^2}{\theta_B^2 + \theta_s^2} \right).$$

Show that when the source is small compared to the beam, the main beam brightness temperature $T_B = T_0(\theta_s/\theta_B)^2$, and further the antenna temperature $T_A = \eta_B T_0 (\theta_s/\theta_B)^2$.

2. Suppose that a source has $T_0 = 600\,\mathrm{K}$, $\theta_0 = 40''$, $\theta_B = 8'$ and $\eta_B = 0.6$. What is T_A? (Use the result of Problem **1(b)**.)

3. Suppose your television needs $1\,\mu\mathrm{W}$ of power at the input for good reception. The transmitter radiates $100\,\mathrm{kW}$ in all azimuthal directions, and within an angle $\pm10°$ about the horizontal direction, and is at $100\,\mathrm{m}$ elevation. Ignore reflections and assume that the earth is perfectly flat. Calculate the effective area, A_e, that your TV antenna must have if you live $30\,\mathrm{km}$ from the transmitter.

4. Suppose that your antenna has a normalized peak power, P, with the following values: $P = 1$ for $\theta < 1°$, $P = 0.1$ for $1° < \theta < 10°$, and $P = 0$ for $\theta > 10°$. What is Ω_A, from Eq. (5.51) in "Tools"? What is Ω_{MB} and η_B.

5. A scientist claims that for a very special antenna the source brightness temperature of a compact source exceeds the antenna temperature. Do you believe this?

6. You are told that there is a special procedure which allows the *measured* Gaussian source size (*not* the deconvolved size) to be smaller than the Gaussian telescope

beam. This can occur (so the claim goes) if the source is very intense. Do you believe this?

7. The Gaussian function considered in Chap. 4 was:

$$y(x) = A \exp\left(-\frac{x^2}{2\sigma^2}\right),$$

where A is a normalization constant. For radio astronomical applications, one usually takes the form of this function as

$$y(x) = A \exp\left(-\frac{4\ln 2 (x - x_0)^2}{\theta_{1/2}^2}\right).$$

Relate the parameters σ and $\theta_{1/2}$. The quantity $\theta_{1/2}$ is the FWHP, full width to half power. In the literature, the "width" of a Gaussian function is usually the FWHP.

8. The ground screen for the Arecibo telescope has a height of 15 m, and is mounted around the edge of the 305 m diameter radio telescope. Assume you could direct the entire ground screen so that the power is collected at a single location. **(a)** What is the geometric area of this ground screen? Take the antenna as a ring, with an inner radius of 305 m, the outer radius being 315 m. **(b)** Calculate the far-field antenna pattern. What are the location and intensity in the first sidelobe, relative to the main lobe? **(c)** Calculate the conversion factor, from Jy to K, for the antenna temperature if the antenna efficiency is 0.6.

9. Single telescope pointing is checked by scanning through the center positions of known sources by a few beamwidths in orthogonal directions. The positional error, $\Delta\theta$, caused by random noise, as measured with a beam of FWHP size θ_0 and signal-to-noise ratio of (S/N) is $\theta_0/(S/N)$. Neglect all systematic errors. What would have to be the (S/N) to determine a source position to 1/50 of the FWHP beamwidth of the telescope? Is there a contradiction between the angular resolution of a telescope, $\theta \sim \lambda/D$, and the positional accuracy?

10. Figure 7.6d represents the AT&T Bell Labs 7 m radio telescope. This has a beam efficiency of 0.95 at a wavelength of 3 mm. Assume that $K = 0$ in Eq. (6.75) and Table 6.1. From Fig. 7.7, what must be the surface accuracy? **(a)** What must be the antenna efficiency from Fig. 7.7? **(b)** At one time, this telescope was used for satellite tests at 28 GHz. The satellite is a point source in the beam of this telescope, so η_A should be optimized for a point source. Now what are the values of antenna and beam efficiency? What is the beam size?

11. Combine (7.5), (7.8) and (7.9), together with $\theta_{\text{geom}} = \frac{\lambda}{D}$ and $A_{\text{geom}} = \frac{\pi}{4}D^2$ to obtain the relation

$$\eta_B = \eta_A \frac{\pi D^2}{4\lambda^2} \left[\frac{\theta_B}{\theta_{\text{geom}}}\right]^2 \tag{7.27}$$

12. Use expression (6.40), to determine the normalized power pattern of the Hertz dipole. Use Eq. (7.2) to determine the gain of the Hertz dipole. For the Hertz dipole, $P(\theta) = P_0 \sin^2 \theta$. Use Eqs. (7.3), (7.5) and 7.11 to obtain Ω_A, Ω_{MB}, η_B and A_e.

13. What is the Rayleigh distance, $k = 2D^2/\lambda$, for an antenna of diameter $D = 100$ m and a wavelength $\lambda = 3$ cm.

14. For a 305 m diameter radio telescope with η_A=0.5, what is the ratio of antenna temperature to flux density for a point source? for an antenna of diameter $D = 100$ m and a wavelength $\lambda = 3$ cm.

Chapter 8
Single Dish Observational Methods

8.1 The Earth's Atmosphere

For ground–based radio telescopes, the signal entering the receiver has been attenuated by the earth's atmosphere. In addition to attenuation, there is atmospheric emission, the signal is refracted and there are changes in the path length. Usually these effects change slowly with time, but there can also be rapid changes such as scintillation and anomalous refraction. Thus the propagation properties of the atmosphere and detailed features of its radiation must be taken into account, if the astronomical measurements are to be interpreted properly. In Sect. 1.2 it was noted that the earth's atmosphere is fairly transparent to radio waves for frequencies above the cut-off given by the critical frequency of free electrons in the ionosphere. This cut-off frequency varies depending on the electron density but usually in the region below 10 MHz. Most radio astronomical measurements are made at frequencies well above this limit. At lower frequencies ionospheric effects can be important; these are of great intrinsic interest for geophysics, and must be compensated for in high angular resolution, low frequency astronomical images.

For the cm and mm wavelength range and especially in the submillimeter range, tropospheric absorption has to be taken into account. The various constituents of the atmosphere absorb by different amounts. Because the atmosphere can be considered to be in LTE, these constituents are also radio emitters. Clouds of water droplets absorb and scatter radio waves even at frequencies as low as 6 GHz – a large rain cloud will cause an attenuation as high as 1.5 dB, while the average value for clear sky at zenith is of the order of 0.2 dB. At higher frequencies the atmospheric absorption increases.

The dry atmosphere below 80 km is a mixture of gases with the principle constituents nitrogen (N_2: 78.09% by volume), oxygen (O_2 : 20.95%) and argon (Ar : 0.93%). This mixture is almost constant in the lower atmosphere, but there are several minor constituents whose relative percentage may vary both with altitude and time.

The most important of these is water vapor (H_2O). Its concentration, given by the mixing ratio r (in g/kg air) varies erratically with the local weather conditions and with altitude. Carbon dioxide (CO_2) with an average percentage of 0.03% shows

both seasonal variations and a secular trend. In recent years it has come to prominence in connection with the greenhouse effect.

Equally notable is ozone (O_3). This has maximum concentration at an altitude between 20 and 30 km, with a total number density of about $5 \times 10^{12} \, \text{cm}^{-3}$. Ozone shows strong seasonal and geographical variations, and in addition, the total amount has decreased dramatically in the last 10 to 15 years. Since ozone is responsible for the absorption of the near UV solar radiation, this decrease is of great practical importance. Spectral lines of ozone are present at 67.36 GHz and higher. The emission lines in the zenith reach $\Delta T_b = 60 \, \text{K}$ for dry air; with a total amount of $2 \, \text{g cm}^{-2} \, H_2O$, ΔT_b remains below 5–6 K.

The atmospheric pressure decreases roughly exponentially with the altitude h

$$P(h) = P_0 \, e^{-h/H} \, . \tag{8.1}$$

The determination of H, the scale height, is rather approximate, with a typical value of

$$H = \frac{\mathscr{R} \, T}{\mu \, g} \approx 7998 \, \text{m}, \tag{8.2}$$

where μ is the mean molecular mass of air, $\mathscr{R}$ the gas constant, g the gravitational acceleration and T the gas temperature.

The total amount of precipitable water (usually measured in mm) above an altitude h_0 is an integral along the line-of-sight. Frequently, the amount of H_2O is determined by radio measurements carried out at 225 GHz combined with models of the atmosphere. For excellent sites, measurements of the 183 GHz spectral line of water vapor can be used to estimate the total amount of H_2O in the atmosphere. For sea level sites, the 22.235 GHz line of water vapor is used for this purpose. The scale height $H_{H_2O} \approx 2 \, \text{km}$, is considerably less than $H_{air} \approx 8 \, \text{km}$ of dry air. For this reason, sites for submillimeter radio telescopes are usually mountain sites with elevations above $\approx 3000 \, \text{m}$.

The variation of the intensity of an extraterrestrial radio source due to propagation effects in the atmosphere is given by [see (1.17)]

$$I_v(s) = I_v(0) \, e^{-\tau_v(0)} + \int_0^{\tau_v(0)} B_v(T(\tau)) \, e^{-\tau} \, d\tau, \tag{8.3}$$

where

$$\tau_v(s) = \int_{s_0}^{s} \kappa_v(s) \, ds \, . \tag{8.4}$$

Here s is the (geometric) path length along the line-of-sight with $s = 0$ at the upper "edge" of the atmosphere and $s = s_0$ at the receiving telescope. Both the (volume) absorption coefficient κ and the gas temperature T will vary with s, introducing the mass absorption coefficient k_v by

$$\kappa_v = k_v \cdot \varrho \, , \tag{8.5}$$

where ϱ is the gas density; this variation of κ can mainly be traced to that of ϱ as long as the gas mixture remains constant along the line-of-sight.

Because the variation of ϱ with s is so much larger that that of $T(s)$, a useful approximation can be obtained by introducing an effective temperature for the atmosphere so that

$$T_A(s) = T_b(0)\,e^{-\tau(0)} + T_{Atm}(1 - e^{-\tau(0)}). \tag{8.6}$$

The first term is absorption in the earth's atmosphere, while the second is emission. The physics of the atmosphere is contained in the derivation of $\tau(0)$, the total opacity along the line-of-sight, and in T_{Atm}. In Fig. 8.1 we show a model of the atmosphere used to predict attenuation from O_2 and H_2O, and other constituents. This example gives an indication of the influence of the atmosphere at cm, mm and sub-mm wavelengths. High frequency resolution measurements of the atmospheric emission are possible and aid in improving models. In the cm region, there is some effect near 22.235 GHz from H_2O and a large effect at 50–70 GHz from O_2. If we assume the physical parameters of the atmosphere are independent of position (within a neighborhood of ≈ 100 km diameter) so that all variations of atmospheric mixture, density, pressure and temperature are only dependent on the height h in the atmosphere, then the total opacity $\tau(z)$ along the line-of-sight at the zenith angle z will be

$$\tau(z) = \tau_0 \cdot X(z), \tag{8.7}$$

where τ_0 is the optical depth for zenith and $X(z)$ is the relative air mass

$$\boxed{X(z) = \frac{1}{\int_0^\infty \varrho(h)\,dh} \int_0^\infty \frac{\varrho(h)}{\sqrt{1 - \left(\dfrac{R}{R+h}\dfrac{n_0}{n}\right)^2 \sin^2 z}}\,dh} \quad . \tag{8.8}$$

Here R is the earth radius, $\varrho(h)$ the gas density at the height $h = r - R$ in the atmosphere, $n(h)$ the index of refraction at this height, n_0 that at $h = 0$. In deriving this expression the surfaces of equal physical state in the atmosphere are assumed to be concentric spheres, and the curvature of the rays due to varying refraction has been taken into account. Tables of $X(z)$ have been computed by Bemporad [see Schoenberg (1929)], a Chebyshev fit up to $X = 5.2$ with a mean error of less than 6.4×10^{-4} is given by

$$X(z) = -0.0045 + 1.00672\sec z - 0.002234\sec^2 z - 0.0006247\sec^3 z. \tag{8.9}$$

If the atmosphere can be considered to be reasonably stable so that both T_{Atm} and τ_0 will not vary noticeably within several hours, these atmospheric parameters can be obtained by measuring a calibration source repeatedly so that its radiation enters the telescope after passing through different air masses X. The unknown atmospheric parameters then can be obtained from a least squares fit of T_A against X.
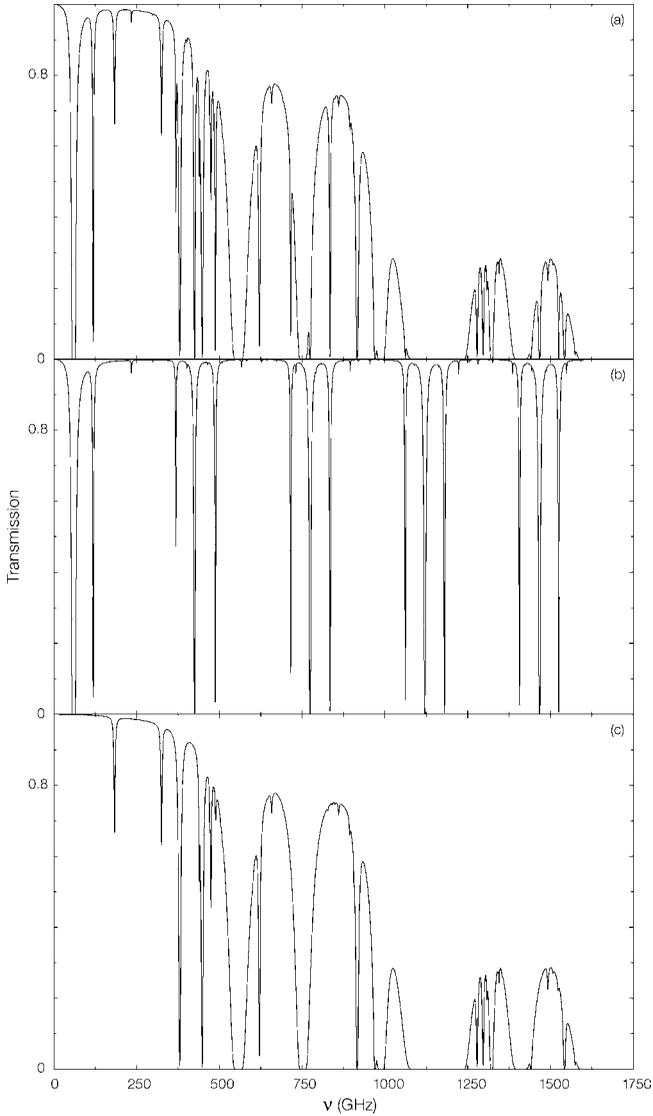
Fig. 8.1 A plot of the transmission properties of an atmospheric model for a precipitable water vapor of 0.2 mm, for an altitude of 4.2 km. In panel (**a**) is shown the total transmission, in (**b**) the transmission if only O_2 is present, and in (**c**) if only water vapor is present. The total in (**a**) also takes the effect of trace gases and the interaction with N_2 [produced using the AM program of S. Paine by B. Nikolic (unpublished)]

Refraction effects in the atmosphere depend on the real part of the (complex) index of refraction (2.33). Except for the anomalous dispersion near water vapor lines and oxygen lines, it is essentially independent of frequency and given by

$$(n-1) \times 10^6 = 77.6 \left(\frac{p}{\text{hPa}}\right)\left(\frac{T}{\text{K}}\right)^{-1} + 3.73 \times 10^5 \left(\frac{p_\text{w}}{\text{hPa}}\right)\left(\frac{T}{\text{K}}\right)^{-2}, \qquad (8.10)$$

where p is the total air pressure in hectoPascals (hPa = 100 mB), p_w the partial pressure of water vapor (in hPa) and T the gas temperature in K. Therefore the refraction will depend on the humidity of the air. For $z < 80°$

$$\Delta z = \beta \tan z, \qquad (8.11)$$

where

$$\beta = (n-1).$$

A rapidly time variable effect is *anomalous refraction* (Altenhoff et al. 1987). This has been found at 1.3 cm at Effelsberg and at 3 mm and 1.3 mm at Pico Veleta as well as other sites. If anomalous refraction is important, the apparent positions of radio sources appear to move away from their actual positions by up to $40''$ for time periods of 30 s. This effect occurs more frequently in summer afternoons than during the night. Anomalous refraction is caused by small variations in the H_2O content, perhaps by single cells of moist air. In the mm and sub-mm range, there are measurements of rapidly time variable noise contributions, the so-called *sky noise*. This is probably produced by variations in the water vapor content in the telescope beam, and it does not depend in an obvious way on the transmission of the atmosphere. But experience has shown that times of high atmospheric attenuation are often times of low sky noise. The fluctuations of the water content along the line-of-sight probably cancel reasonably well when the water content is high, but then the transmission is small. As expected, sky noise increases with increasing telescope beam separation, being larger for small telescopes ($D < 3$ m) than for large telescopes ($D > 10$ m). This behavior is expected if the effects arise within a few km above the telescope and the cells have limited sizes.

8.2 Calibration Procedures

8.2.1 General

In Sect. 7.1, a set of characteristic parameters was given that describes the basic properties of an antenna as a measurement device. These parameters have to be determined for a specific antenna. The efficiencies defined in Chap. 6 are more difficult to estimate. For smaller antennas used in communications, calibrations are usually

carried out using an antenna test stand using transmitters with known power output. Such a transmitter must be situated in the far field, at a distance d,

$$d > 2D^2/\lambda \,, \tag{8.12}$$

if easily interpretable results are to be obtained. The required distance d is of the order of 2×10^3 km for a telescope with $D = 100$ m at $\lambda = 1$ cm. Such calibration set up is not possible to arrange on a routine basis and therefore radio telescopes and larger communications antennas are best calibrated using astronomical sources as standards.

Thus, in radio astronomy, one must follow a two step procedure characteristic for many astronomical applications. First, the measurements must be calibrated using a set of celestial primary calibration sources, but to establish these primary calibration sources is a complicated task. Second, once the primary flux density calibrators are available, relative calibrations can be made, using secondary standards.

The primary calibration standards are measured using horn antennas. The antenna parameters can either be computed theoretically, or can be measured in an antenna test range. The same sources are then measured again with the larger telescope to be calibrated. The scale measured with the larger telescope is then adjusted to produce the values obtained with the calibrators in catalogs. A list of radio sources with known flux densities for a wide range of frequencies is available; for convenience a sample is given in Appendix F.

For a uniformly illuminated circular aperture, as in an optical telescope, the far field is given by the Airy pattern (the first entry in Table 6.1). The illumination of radio telescopes differs from that used in classical optical telescopes in two ways: (1) because of the low focal ratios, the illumination pattern differs significantly from the Airy pattern, and (2) the waveguide feeds give a nearly Gaussian variation of electric field strength across the aperture. Such a situation is better represented by the second entry in Table 6.1. For a Gaussian shaped main beam, the solid angle is given by

$$\boxed{\Omega = 1.133 \; \theta_b^2} \quad , \tag{8.13}$$

where θ_b is the full width to half power (FWHP). From direct comparisons, this relation has an accuracy of 5%. An accuracy of 1% can be obtained by using the 0.1 power point of the beam:

$$\Omega = 0.3411 \; \theta_{0.1\,\mathrm{power}}^2 \,. \tag{8.14}$$

8.2.2 Compact Sources

Usually the beam of radio telescopes are well approximated by Gaussians. As mentioned previously, Gaussians have the great advantage that the convolution of two Gaussians is another Gaussian. For Gaussians, the relation between the observed source size, θ_o, the beam size θ_b, and actual source size, θ_s, is given by:

$$\theta_{\mathrm{o}}^2 = \theta_{\mathrm{s}}^2 + \theta_{\mathrm{b}}^2 . \qquad (8.15)$$

This is a completely general relation, and is widely used to deconvolve source from beam sizes. Even when the source shapes are not well represented by Gaussians these are usually approximated by sums of Gaussians in order to have a convenient representation of the data. The accuracy of any determination of source size is limited by (8.15). A source whose size is less than 0.5 of the beam is so uncertain that one can only give as an upper limit of $0.5\,\theta_{\mathrm{b}}$.

If the (lossless) antenna (outside the earth's atmosphere) is pointed at a source of known flux density S_ν with an angular diameter that is small compared to the telescope beam, a power at the receiver input terminals

$$W_\nu \, \mathrm{d}\nu = \tfrac{1}{2} A_{\mathrm{e}} \, S_\nu \, \mathrm{d}\nu = k T_{\mathrm{A}} \, \mathrm{d}\nu$$

is available. Thus

$$\boxed{T_{\mathrm{A}}' = \Gamma S_\nu} \qquad (8.16)$$

where Γ is the *sensitivity* of the telescope measured in K Jy^{-1}. Introducing the aperture efficiency η_{A} according to (7.9) we find

$$\boxed{\Gamma = \eta_{\mathrm{A}} \frac{\pi D^2}{8k}} \qquad . \qquad (8.17)$$

Thus Γ or η_{A} can be measured with the help of a calibrating source provided that the diameter D and the noise power scale in the receiving system are known. In practical work the inverse of relation (8.16) is often used. Inserting numerical values we find

$$S_\nu = 3520 \, \frac{T_{\mathrm{A}}'[\mathrm{K}]}{\eta_{\mathrm{A}}[D/\mathrm{m}]^2} . \qquad (8.18)$$

The *brightness temperature* is defined as the Rayleigh-Jeans temperature of an equivalent black body which will give the same power per unit area per unit frequency interval per unit solid angle as the celestial source. Both T_{A} and T_{b} are defined in the classical limit, and *not* through the Planck relation, even when $h\nu \approx kT$. In this sense, both T_{A} and T_{b} may be idealizations. However the brightness temperature scale has been corrected for antenna efficiency. Usually this scale, the main beam brightness temperature, is determined by measurements of the planets, or measurements of calibration sources. The conversion from source flux density to source brightness temperature for sources with sizes small compared to the telescope beam is given by the Rayleigh-Jeans relation:

$$S = \frac{2k T_{\mathrm{MB}} \, \Omega}{\lambda^2} . \qquad (8.19)$$

For Gaussian source and beam shapes, this relation is:

$$S = 2.65 \frac{T_{MB} \, \theta_o^2}{\lambda^2} , \tag{8.20}$$

where the wavelength is in centimeters, and the observed source size is taken to be the beam size, given in arc minutes. Then the flux density is in Jy, and the brightness temperature is in Kelvin.

The expression T_{MB} is still antenna dependent, in the sense that the temperature is a weighted average over the telescope beam, but this relation does take into account corrections for imperfections in the antenna surface and the efficiency of feed coupling. For sources small compared to the beam, the antenna and main beam brightness temperatures are related by the main beam efficiency, η_B:

$$\eta_B = \frac{T_A'}{T_{MB}} . \tag{8.21}$$

The actual source brightness temperature, T_s is related to the main beam brightness temperature by:

$$T_s = T_{MB} \frac{(\theta_s^2 + \theta_b^2)}{\theta_s^2} . \tag{8.22}$$

Here, we have made the assumption that source and beam are Gaussian shaped. The actual brightness temperature is a property of the source. To obtain this, one must determine the actual source size. This is a science driver for high angular resolution (i.e. interferometry) measurements. Although the source may not be Gaussian shaped, one normally fits multiple Gaussians to obtain the effective source size.

8.2.3   Extended Sources

For sources extended with respect to the beam, the process is vastly more complex, because the antenna side lobes also receive power from the celestial source, and a simple relation using beam efficiency is not possible without detailed measurements of the antenna pattern. As shown in Sect. 8.2.5 the error beam may be a very significant source of errors, if the measurements are carried out near the limit of telescope surface accuracy. In principle η_{MB} could be computed by numerical integration of $P_n(\vartheta, \varphi)$ [cf. (7.3) and (7.4)], provided that $P_n(\vartheta, \varphi)$ could be measured for large range of ϑ and φ. Unfortunately this is not possible since nearly all astronomical sources are too weak; measurements of bright astronomical objects with known diameters can be useful.

If we assume a source has a uniform brightness temperature over a certain solid angle Ω_s, then the telescope measures an antenna temperature given by (7.16) which, for a constant brightness temperature across the source, simplifies to

$$T'_A = \frac{\displaystyle\int_{\text{source}} P_n(\vartheta,\varphi)\,d\Omega}{\displaystyle\int_{4\pi} P_n(\vartheta,\varphi)\,d\Omega}\, T_b$$

or, introducing (7.3, 7.4 and 7.5),

$$T'_A = \eta_B \frac{\displaystyle\int_{\text{source}} P_n(\vartheta,\varphi)\,d\Omega}{\displaystyle\int_{\substack{\text{main}\\\text{lobe}}} P_n(\vartheta,\varphi)\,d\Omega}\, T_b = \eta_B f_{\text{BEAM}} T_B\,, \tag{8.23}$$

where f_{BEAM} is the beam filling factor. If the source diameter is of the same order of magnitude as the main beam the correction factor in (8.23) can be determined with high precision from measurements of the normalized power pattern and thus (8.23) gives a direct determination of η_B, the beam efficiency. A convenient source with constant surface brightness in the long cm wavelength range is the moon whose diameter of $\cong 30'$ is of the same order of magnitude as the beams of most large radio telescopes and whose brightness temperature

$$T_{b\,\text{moon}} \cong 225\,\text{K} \tag{8.24}$$

is of convenient magnitude. In the mm and submillimeter range the observed Moon temperature changes with Moon phase. The planets form convenient thermal sources with known diameters that can be used for calibration purposes (Altenhoff 1985).

8.2.4 Calibration of cm Wavelength Telescopes

In the centimeter wavelength range, the noise from the atmosphere is small (Fig. 8.1), so $T_{rx} \gg T_{atm}$. Then a small amount of noise from a broadband calibration source, whose value is known in Kelvin, is added to the system noise. The cycle consists of two parts: system plus calibration noise on, system plus calibration noise off. If there is no zero offset these two measurements give the noise of the system (= receiver + atmosphere). There are two methods to establish the calibration scale: (1) if the temperature of the calibration noise is determined by a comparison with astronomical sources, then noise source intensity can be given in Jy or T_{MB}, or (2) if the noise source is calibrated using the response to a hot and cold absorber in front of the prime feed, the calibration noise source intensity is given in Kelvin, antenna temperature, T_A. This T_A value is usually determined at one frequency since the response of the calibration varies with frequency because of mismatches in the calibration system. To convert from a T_A to a T_{MB} scale, one can use the beam efficiency but

must also correct for atmospheric extinction. More accurate is the measurement of an astronomical source of known flux density (or equivalently T_{MB}). If at a different elevation, the variation caused by the atmosphere and telescope gain must be taken into account.

8.2.5 Calibration of mm and sub-mm Wavelength Telescopes for Heterodyne Systems

In the millimeter and submillimeter wavelength range, the atmosphere has a larger influence and can change rapidly, so we must make accurate corrections to obtain well calibrated data. In addition, in the mm range, most large telescopes are close to the limits caused by their surface accuracy, so that the power received in the error beam may be comparable to that received in the main beam. Thus, one must use a relevant value of beam efficiency. We give an analysis of the calibration procedure which is standard in spectral line mm astronomy following the presentations of Downes (1989) and Kutner and Ulich (1981). This calibration reference is referred to as the *chopper wheel* method. The procedure consists of: (1) the measurement of the receiver output when an ambient (room temperature) load is placed before the feed horn, and (2) the measurement of the receiver output, when the feed horn is directed toward cold sky at a certain elevation. For (1), the output of the receiver while measuring an ambient load, T_{amb}, is V_{amb}:

$$V_{amb} = G(T_{amb} + T_{rx}).   \tag{8.25}$$

For step (2), the load is removed; then the response to empty sky noise, T_{sky} and receiver cabin (or ground), T_{gr}, is

$$V_{sky} = G[F_{eff}T_{sky} + (1 - F_{eff})T_{gr} + T_{rx}].   \tag{8.26}$$

F_{eff} is referred to as the *forward efficiency*. This is basically the fraction of power in the forward beam of the feed. If we take the difference of V_{amb} and V_{sky}, we have

$$V_{cal} = V_{amb} - V_{sky} = G F_{eff} T_{amb} e^{-\tau_\nu},   \tag{8.27}$$

where τ_ν is the atmospheric absorption at the frequency of interest. The response to the signal received from the radio source, T_A, through the earth's atmosphere, is

$$\Delta V_{sig} = G T_A' e^{-\tau_\nu}$$

or

$$T_A' = \frac{\Delta V_{sig}}{\Delta V_{cal}} F_{eff} T_{amb}$$

where T'_A is the antenna temperature of the source outside the earth's atmosphere. We define

$$T^*_A = \frac{T'_A}{F_{\text{eff}}} = \frac{\Delta V_{\text{sig}}}{\Delta V_{\text{cal}}} T_{\text{amb}}. \qquad (8.28)$$

The quantity T^*_A is commonly referred to as the *corrected antenna temperature*, but it is really a *forward beam brightness temperature*. This is the T_{MB} of a source filling a large part of the sky, certainly more than $30'$.

For sources (small compared to $30'$), one must still correct for the telescope beam efficiency, which is commonly referred to as B_{eff}. Then

$$T_{\text{MB}} = \frac{F_{\text{eff}}}{B_{\text{eff}}} T^*_A$$

for the IRAM 30 m telescope, $F_{\text{eff}} \cong 0.9$ down to 1 mm wavelength, but B_{eff} varies with the wavelength. So at $\lambda = 3$ mm, $B_{\text{eff}} = 0.65$, at 2 mm $B_{\text{eff}} = 0.6$ and at 1.3 mm $B_{\text{eff}} = 0.45$, for sources of diameter $< 2'$. For an object of size $30'$, B_{eff} at all these wavelength is 0.65. As usual T_{MB} can be considered a black body with the temperature T_{MB}, which just fills the beam. This analysis is the one used at IRAM; an earlier analysis by Kutner and Ulich (1981) is common in the USA. This uses a somewhat different notation, but the physics is basically the same. We give a comparison between these systems.

Kutner & Ulich	CLASS/IRAM
η_l	F_{eff}
η_s	B_{eff}
η_c	–
η_{fss}	$B_{\text{eff}}/F_{\text{eff}}$
η_f	–

$\eta_s = \eta_l \cdot \eta_{\text{fss}}, \; \eta_f = \eta_{\text{fss}} \cdot \eta_c$

In the notation of Kutner and Ulich (1981), $T^*_R = T_{\text{MB}}$, in terms of our notation in Chap. 7

$$\eta_{\text{MB}} = \frac{\Omega_{\text{MB}}}{\Omega_F} = \frac{B_{\text{eff}}}{F_{\text{eff}}},$$

while in that of Kutner and Ulich

$$\eta_{\text{fss}} = \frac{\Omega_D}{\Omega_F},$$

where Ω_D is the solid angle of the diffraction pattern and Ω_F is the forward beam solid angle.

An antenna pointing at an elevation z to a position of empty sky will produce an antenna temperature

$$T_A(z) = T_{rx} + T_{atm}\, \eta_l \left(1 - e^{-\tau_0 X(z)}\right) + T_{amb}(1 - \eta_l), \qquad (8.29)$$

where

T_{rx}: system noise temperature,
T_{atm}: effective temperature of the atmosphere,
T_{amb}: ambient temperature,
η_l: feed efficiency (typically $\eta_l = 0.9$),
τ_0: zenith optical depth,
$X(z)$: air mass at zenith distance z.

These parameters can be determined by a series of calibration measurements. The efficiency η_l and the other parameters can be determined by a least squares fit of (8.29), that is a *skydip* giving T_A as a function of $X(z)$. Depending on the weather conditions these measurements have to be repeated at time intervals from 15 min to hours or so, to be able to detect variations in the atmospheric conditions. At some observatories a small separate instrument, a *taumeter* is available to sample the opacity τ at 10 min intervals.

For small telescopes used for dedicated projects, such as the Harvard CfA 1.2 m dish, or its identical Chilean counterpart, operated by the University of Chile, one has an ideal situation for accurate calibrations, so one can carry out tipping measurements often, and determine all the parameters in (8.29). In addition, the surface accuracy is quite high, so that the error beam contribution is small. Using these, one can accurately correct the data. The CO in and near the galactic plane has been mapped, and these results are on the T_{MB} scale.

For larger mm wavelength telescopes one cannot perform tipping measurement often. If a taumeter is not available one must use a more elaborate procedure. By measuring the response to a cold load, one can determine the receiver noise, and can obtain a good estimate of the noise from the atmosphere. Then, assuming a value of T_{atm} and $\eta_l = F_{eff}$, one can then determine $\tau = \tau_0 X(z)$, and can use this to correct for atmospheric absorption.

At present many millimeter and submillimeter front ends are still double sideband mixers. This can cause additional uncertainties for line measurements. One is that the sensitivity for line radiation is lowered compared to wide band continuum signals. The reason is that the spectral lines will be present only in one of the side bands, but the calibration signal and the noise, from both the atmosphere and receiver, will enter both side bands. Additional complications may arise if the atmospheric absorption is noticeably different in the two side bands. Generalizing (8.29) for a double side-band system with a gain of 1 in the signal band and a relative gain g_i in the image band, with optical depths in the image band, τ_i and signal band, τ_s, we find

$$T_{cal} = (T_{amb} - T_{atm})(1 + g_i)\, e^{-\tau_s} + T_{atm}(1 + g_i\, e^{\tau_s - \tau_i}). \qquad (8.30)$$

To calibrate spectral lines, one frequently measures sources for which one has single sideband spectra. Finally observations often have to be corrected for yet another effect: the telescope efficiency usually depends on elevation. Usually the

telescope surface is set optimally for some intermediate zenith distance $z \approx 40°$. Both for $z \approx 0°$ and $70°$ the efficiency usually decreases by about 25%.

8.2.6 Bolometer Calibrations

Since most bolometers are AC coupled, the D. C. response to "hot–cold" or "chopper wheel" calibration methods are not used. Instead astronomical data are calibrated in two steps: (1) measurements of atmospheric emission to determine the opacities at the azimuth of the target source, and (2) the measurement of the response of a nearby source with a known flux density; immediately after this, a measurement of the target source is carried out.

8.3 Continuum Observing Strategies

8.3.1 Point Sources

Even the most carefully designed astronomical receivers are affected by random noise and instabilities. Prolonging the observing time will diminish the influence of the first kind of error, but after a certain time the growth of instabilities will worsen the result. This is the physical content of the Allan plots described in Chap. 4. Thus one must minimize the effect of instabilities. The solution to this problem involves the use of differential techniques as far as possible; this has been standard practice both in astronomy and physics since the days of Wheatstone. The various schemes differ, depending on the dominant source of instability. In the early days of radio astronomical systems the amplifiers were the main source of instabilities and therefore sophisticated compensation schemes with Dicke switches etc. as described in Chap. 4 have been implemented. The technical advances in the art of receiver construction resulted in an ever improving stability of these systems so that rapid switching is not needed to cancel receiver instabilities. In the cm range it is now possible to have total power systems with separate feed horns mounted side-by-side. The observing method consists of switching between the outputs in software. This "software switching" is used for sensitive continuum measurements at the 100–m telescope where variations in the atmospheric properties are the dominant source of instabilities.

In the submillimeter range the earth's atmosphere is a large source of radiation. In average weather conditions at the 2.4 km high Pico Veleta site in Spain, at 1.3 mm the atmospheric emission contributes 100 K (≈ 680 Jy in the $12''$ beam of the IRAM 30 m telescope) at $30°$ elevation. If flux densities in the mJy range are to be measured with multi-beam cameras, sophisticated compensation schemes are used. These involve both rapid beam switching and the subtraction of "off-source" from "on-source" measurements.

From (Eq. 8.5) there are two effects: absorption and emission. The emission always raises the system noise. For narrow band spectral lines the effects are somewhat less serious since usually the emission effects can be neglected because of the narrow band and only absorption has to be taken into account. For continuum measurements, the emission can be significant and strongly affects the continuum data. For antennas with single beams, various on–off schemes are used when "point" sources are to measured. Whether the position switching is made in azimuth-elevation, right ascension and elevation or any other coordinate system depends on the telescope or on the problem to be investigated. Usually the on–off measurements are arranged to be symmetrical, to balance the atmospheric effects.

A much better compensation of transmission variations in the atmosphere is possible if double beam systems can be used. In the simplest system the individual telescope beams should be spaced by a distance of at least 3 FWHP beam widths, and the receiver should be switched between them. The separate beams can be implemented in different ways depending on the frequency and the technical facilities at the telescope. At fairly low frequencies, such as the 10 GHz system at the Effelsberg 100 m telescope, separate feed horns and receivers are installed in the secondary focus. After detection the receiver outputs are differenced in a computer. At 1.3 cm, on this telescope, the direction of the single feed horn is mechanically moved rapidly by a few mm, changing the illumination of the dish by this. This changes the direction of the beam direction by a few beamwidths.

At higher frequencies, in the mm and submillimeter range, the rapid movement of the telescope beam over small angles, so-called "wobbling" is used to produce two beams on the sky from a single pixel. This is used at all large millimeter facilities.

Multi-beam bolometer systems are now the rule. With these, one can measure a fairly large region simultaneously. This allows a higher mapping speed, and also provides a method to better cancel sky noise due to weather. Such weather effects are sometimes referred to as "coherent noise". Some details of more recent data methods are given in e.g. Motte et al. (2007). Usually, a wobbler system is needed for such arrays, since the bolometer output is AC–coupled.

Observing procedures for a double beam system are usually as follows: the source is first centered on beam one, and the difference of the two beams is measured, optimally by wobbling the sub reflector. Then the source is centered on beam two, and again the difference is measured. This on–off method (better called on–on, because the source is always in one of the beams) is often arranged in a time symmetric fashion so that time variations of the sky noise and other instrumental effects can be eliminated.

8.3.2 Imaging of Extended Continuum Sources

If extended areas are to be mapped, some kind of raster scan is employed: there must be reference positions at the beginning and the end of the scan. Usually the area is measured at least twice in orthogonal directions. After gridding, the differences of

the images are least squares minimized to produce the best result. This procedure is called "basket weaving".

Extended emission regions can also be mapped using a double beam system, with the receiver input periodically switched between the first and second beam. In this procedure, there is some suppression of very extended emission. A simple summation along the scan direction has been used to reconstruct infrared images. A more sophisticated scheme, the so-called "EKH" method (Emerson et al. 1979) is given here. Let $T_A(s)$ be the distribution of the antenna temperature as measured by a single beam telescope. A double beam system consisting of two identical beams that point at positions differing by the (constant) B then gives a response

$$\Delta_T(s,B) = \int \left[\delta(t-s) - \delta(t-(s+B))\right] T_A(t)\,dt,\tag{8.31}$$

$$\boxed{\Delta_T(s,B) = \int \text{II}_a(t-s)T_A(t)\,dt}\quad,\tag{8.32}$$

where

$$\boxed{\text{II}_a(t) = \delta(t) - \delta(t+B)}\tag{8.33}$$

is the antisymmetric impulse pair and $\delta(t)$ the Dirac delta function. The result of the observations with the double beam system therefore can be described as the convolution of the antenna temperature distribution with the antisymmetric impulse pair (8.33). The aim of the reduction software is to reconstruct the antenna temperature distribution $T_A(s)$ from the measured $\Delta_T(s)$.

This reconstruction can again be written as a convolution equation

$$\boxed{T_A(s) = \int \text{III}_a(t-s)\,\Delta_T(t,B)\,dt}\tag{8.34}$$

if a solving kernel $\text{III}_a(t)$ can be constructed. Designating the Fourier transform of $F(t)$ by $\overline{F}(s)$, i.e.

$$\overline{F}(s) = \int F(t)\,e^{-2\pi i s t}\,dt$$

then according to the convolution theorem of Fourier transforms (see Appendix B)

$$\overline{\Delta_T}(t) = \overline{\text{II}_a}(t) \cdot \overline{T_A}(t)\tag{8.35}$$

so that

$$\overline{\text{III}_a}(t) = \frac{1}{\overline{\text{II}_a}(t)}\,.\tag{8.36}$$

Since

$$\overline{\amalg_a}(z, B) = 1 - e^{-2\pi i Bz}, \tag{8.37}$$

we find formally

$$\overline{\text{III}_a}(z, B) = \frac{1}{1 - e^{-2\pi i Bz}}. \tag{8.38}$$

Using the identity

$$(1 - e^{-2\pi i Bz}) \left[1 + \sum_{n=1}^{\infty} \left(e^{-2\pi i nBz} - e^{2\pi i nBz} \right) \right] = 2 \tag{8.39}$$

we find

$$\overline{\text{III}_a}(z, B) = \frac{1}{2} \left[1 + \sum_{n=1}^{\infty} \left(e^{-2\pi i nBz} - e^{2\pi i nBz} \right) \right] \tag{8.40}$$

so that

$$\boxed{\text{III}_a(t, B) = \frac{1}{2} \left[\delta(t) + \sum_{n=1}^{\infty} \left(\delta(t - nB) - \delta(t + nB) \right) \right]} \quad . \tag{8.41}$$

This antisymmetric replicating function is the solving kernel for reconstructing the distribution function of the antenna temperature. With this deconvolution algorithm, one can recover most, but not all of the information. Most telescopes therefore have wobbler switching in azimuth to cancel ground radiation. By measuring a source using scans in azimuth at different hour angles, and then combining the maps (see Johnstone et al. 2000) one can recover more information. Another account of data processing for multi-beam bolometers is contained in Motte et al. (2007).

8.4 Additional Requirements for Spectral Line Observations

In addition to the requirements placed on continuum receivers, there are three requirements specific to spectral line receiver systems.

8.4.1 Radial Velocity Settings

If the observed frequency of a line is compared to the known rest frequency, the relative radial velocity of the line emitting (or absorbing) source and the receiving system can be determined. But this velocity contains the motion of the source as well as that of the receiving system. Both are measured relative to some standard of rest. However, usually only the motion of the source is of interest. Thus the velocity

of the receiving system must be determined. This velocity can be separated into several independent components.

1) **Earth Rotation.** Due to the rotation of the earth, the receiving system moves with a velocity $v = 0.46510\cos\varphi$ km s^{-1} in the direction due east in the horizontal coordinate system. Here φ is the geographic latitude of the observing station. If the contribution of this velocity is subtracted the resulting radial velocity is said to refer to the *geocentric* system.

2) **The Motion of the Center of the Earth Relative to the Barycenter of the Solar System.** If this contribution is eliminated the radial velocity is said to be reduced to the *heliocentric* system. This velocity of the earth could be computed from the annually published *Astronomical Ephemeris*, but due to the many effects that must be taken into account, this is a complicated procedure.

Today there are convenient computer algorithms that can be run on any personal computer and which correct the observations for the motion of the earth relative to center of mass of the solar system. For high-precision radial velocity or pulsar timing data, even relativistic corrections must be included. The resulting radial velocities are then as close to an inertial system as we can hope to come, so there is no physical reason to transform the observed radial velocities further. Stellar radial velocity observations and practically all extragalactic work are therefore usually published in this, the Heliocentric velocity system. In galactic work it is, however, convenient to obtain the radial velocities in a system such that gas in the solar neighborhood is at rest. Therefore the motion of the center of mass of the solar system relative to the local gas has to be determined. For this, neutral hydrogen gas as given by the 21 cm line both in emission and in absorption is best suited, but stellar data can be used also. All results obtained by many independent investigations [for a summary see Crovisier (1978)] show that the solar system moves with a velocity given by the *standard solar motion* ($v_0 = 20$ km s^{-1} towards $\alpha_{1900} = 18$ h, $\delta_{1900} = +30°$). This is the solar motion relative to the mode of the velocity of the stars in the solar neighborhood. In practice this is the velocity relative to stars most commonly listed in general catalogs of radial velocity and proper motion; these stars are mostly of spectral types A to G. The reason why the gas velocity is the mode of the velocity of the stars is probably that interstellar gas is collision dominated and therefore not sensitive to outlying extreme velocities, while the moments of the collisionless stellar velocity distribution will be affected. Data from which the standard solar motion has been eliminated are said to refer to the *local standard of rest* (LSR). This is a point coinciding with the position of the sun and moving with the local circular velocity around the galactic center. Sometimes it is advantageous to refer velocities to a system in which the galactic center is at rest. The required correction obviously depends on the adopted galactic circular velocity of the LSR. For many years the value $\Theta_0 = 250$ km s^{-1} as proposed by IAU convention has been used; recently slightly smaller values of $\Theta_0 = 220\text{--}230$ km s^{-1} are preferred, although values as low as $\Theta_0 = 185$ km s^{-1} have been proposed. In all cases the velocity vector is directed towards $l = 90°, b = 0°$.

8.4.2   Stability of the Frequency Bandpass

In addition to the stability of the total power of the receiver, one must also have a
stable shape of the receiver bandpass. At millimeter and sub-mm wavelengths, it is
possible that changes in the weather conditions between on-source and reference
measurements may lead to serious baseline instabilities. If so, the time between on-
source and reference measurements must be shortened until stable conditions are
reached. Such stability is easier to obtain if the bandwidth of the spectrometer is
narrow compared to the bandwidth of those parts of the receiver in front of the
spectrometer.

8.4.3   Instrumental Frequency Baselines

The result of any observing procedure should result in a spectrum in which $T_A(v) \rightarrow$
0 for v outside the frequency range of the line. However, quite often this is not so
because the signal response was not completely compensated for by a reference
measurement, even if receiver stability is ideal. For larger bandwidths, there is an
instrumental spectrum and a "baseline" must be subtracted from the difference spec-
trum. Often a linear function of frequency is sufficient, but sometimes some curva-
ture is found, so that polynomials of second or higher order must be subtracted. This
should be done with great care because high-order polynomials can easily introduce
spurious effects when fitted to disjointed sections of the line spectrum. On many oc-
casions a sinusoidal or quasi-periodic baseline ripple is present. This ripple appears
because quite often a small fraction of the signal is reflected off obstructions in the
aperture plane of the telescope. In axially symmetric telescopes this reflected signal
can form a standing wave pattern. A phase change of 2π radians will occur if either
the distance, d, over which the signals are interfering is changed by $\lambda/2$ (where λ
is the wavelength) or if the frequency is changed by

$$\Delta v = \frac{c}{2d}. \tag{8.42}$$

For the 100 m telescope at Effelsberg with $d = f \approx 30$ m where f is the focal
length of the telescope, so $\Delta v \approx 5$ MHz; the 43 m Green Bank telescope with its
smaller dimensions has $\Delta v = 10.4$ MHz. Attempts to eliminate this ripple usually
employ defocusing of the telescope along the axis by $\pm\lambda/8$, thereby shifting the
phase of the ripple by π radians. The sum of the two baselines then usually shows
considerably less ripple; for the 43 m Green Bank telescope it is decreased by about
one order of magnitude; experience with the Effelsberg 100 m telescope shows that
this procedure is less effective, since the decrease is only a factor of ≈ 3. Very
probably, for the 43 m telescope the largest reflection that produces the standing
wave pattern occurs along the axis, with a single path. Thus the reflected waves
can be canceled by appropriate defocusing. For the 100 m dish a larger part of the
power must be reflected by off-axis structures, perhaps in the prime focus support

structure. Thus, there is no adequate defocusing procedure. In any case the 17 %
aperture blocking of the 100 m telescope certainly gives rise to larger instrumental
baseline effects.

There are several possible sources of reflected radiation: (1) the front end of
the receiver that injects some noise power into the antenna, part of which is then
reflected back; or (2) strong continuum radiation from cosmic sources. In both cases
the partial reflection of the radiation in the horn aperture is the main cause of the
instrumental baseline ripple (Figs. 8.2 and 8.3). Both changes in the position of the
telescope and small changes in the receiving equipment can cause large changes in
the amplitude of the observed ripple. Sometimes the amplitude of baseline ripple
can be reduced considerably by installing a cone at the apex of the telescope that
scatters the radiation forming the standing wave pattern.

The different receiving methods outlined in Sect. 8.4.5 are susceptible to baseline
ripple by quite different amounts. If the background emission is position indepen-
dent the method of position switching using total power mode or "on the fly" map-
ping should result in the smallest ripple, while frequency switching will produce the
largest. Clearly instruments such as the GBT with off-axis receivers and very small
aperture blocking have a vastly lower amount of baseline ripples and also have much
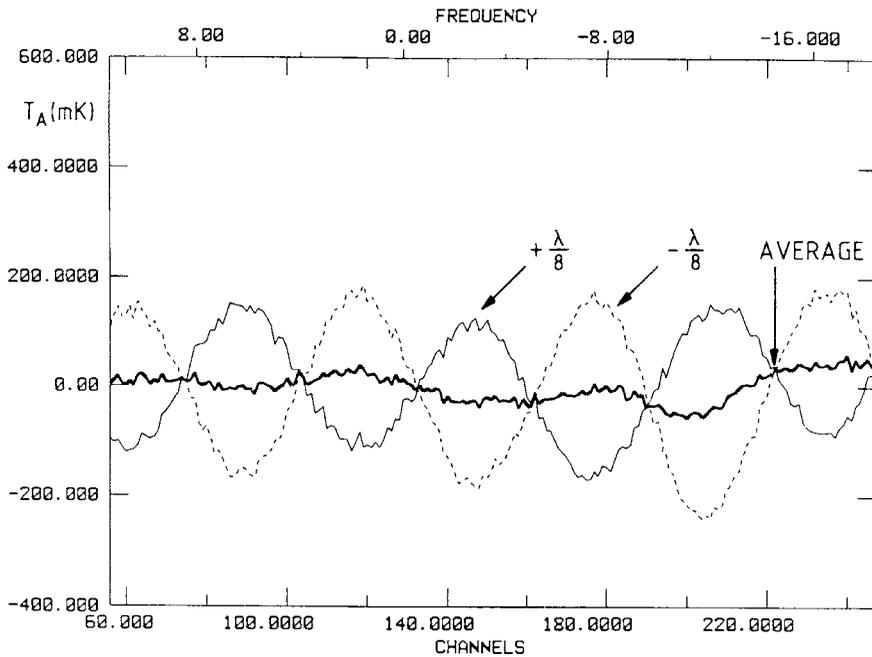lower sidelobe levels.

Fig. 8.2 Instrumental baseline ripples for the Effelsberg 100 m telescope measured at $\pm\lambda/8$ axial
offset from the best focus and the average. The data were taken toward Virgo A at 3.5 cm (Bania
et al. 1994)

Fig. 8.3 Formation of
instrumental baseline ripple
by reflections in the feed
horn: (**a**) noise emitted by the
first stages of the receiver and
radiated into the telescope;
(**b**) noise from a strong source
reflected at the feed horn and
after reflection in the
telescope accepted by the
feed horn. In both cases, only
a part of the power is
accepted by the feed horn
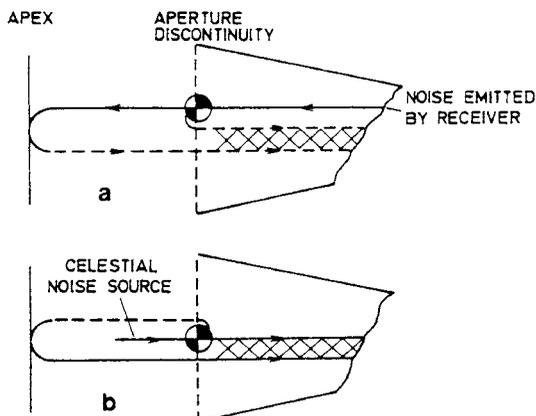
8.4.4   The Effect of Stray Radiation

Accurate measurements are much more difficult to obtain for extended sources, especially for regions of low *brightness temperature*. This is because, according to (7.16), the observed temperature T_A is related to the actual source temperature, T_b, by the power pattern $P(x,y)$

$$T_A(x,y) = \frac{\int P(x-x',y-y')\,T_b(x',y')\,dx'\,dy'}{\int P(x',y')\,dx'\,dy'} \quad . \tag{8.43}$$

Here we have set the radiation efficiency η_R of the antenna equal to 1. In reality spherical coordinates have to be used, but this merely introduces some minor practical and mathematical complications.

Our goal is to invert (8.43), to obtain T_b in terms of the measured T_A. To derive an approximate expression, the radiation received by the antenna is separated into one part received by the *main beam* (mb) and another by the *stray pattern* (sp). The integral (8.43) can then be separated into

$$T_A(x,y) = \frac{1}{\Omega_A}\left[\int_{(mb)} P(x-x',y-y')\,T_b(x',y')\,dx'\,dy' \right.$$

$$\left. + \int_{(sp)} P(x-x',y-y')\,T_b(x',y')\,dx'\,dy' \right] . \tag{8.44}$$

If we now suppose that the position dependence of T_b varies very little over angular scales comparable to the beam width, T_b can be extracted from the integral for the main beam, yielding that part of T_A received by the main beam, T_{AM}:

$$T_{AM}(x,y) = \frac{1}{\Omega_A} \int\limits_{(mb)} P(x-x',y-y')\,T_b(x',y')\,dx'\,dy'$$

$$= \frac{\overline{T_b}(x,y)}{\Omega_A} \int\limits_{(mb)} P(x,y)\,dx\,dy,$$

$$T_{AM} = \eta_B\,\overline{T_b}. \tag{8.45}$$

Here we have introduced the *main beam efficiency* as defined in (7.5). Substituting this into (8.44) and solving for $\overline{T_b}$ which we will call the corrected brightness temperature, we obtain

$$\overline{T_b}(x,y) = \frac{1}{\eta_B}\left[T_A(x,y) - \frac{1}{\Omega_A} \int\limits_{(sp)} P(x-x',y-y')\,T_b(x',y')\,dx'\,dy' \right].$$
$$\tag{8.46}$$

The integral involves the unknown $T_b(x,y)$, but since this is an average over large angles, there will not be a large change if T_b is replaced by $\overline{T_b}$ from (8.45). The computation of T_b thus results in an iterative procedure leading to a Neumann series. As shown by Kalberla et al. (1980) this sequence can be solved by use of

$$\overline{T_b}(x,y) = \frac{1}{\eta_B}\left[T_A(x,y) - \int\limits_{(sp)} R(x-x',y-y')\,T_A(x',y')\,dx'\,dy' \right],$$
$$\tag{8.47}$$

where R is the so-called resolving kernel which can be derived from P by successive approximations. For practical applications R usually differs very little from P. The correction procedure is very complicated and requires knowledge of the radiation T_A across the full sky!

These arguments apply to both wide band continuum radiation and to spectral line radiation. For continuum measurements of regions of low T_{MB}, within a few degrees of an intense source there may be serious errors. The largest side lobes in $P(x,y)$ are caused by the feed support legs; for alt-azimuth telescopes, the position of these side lobes varies with the hour angle, so that the error is a function of time. However, the T_{MB} of these astronomical sources is assumed to remain constant. This observation allows a practical solution: one can employ a beam correction technique in which one first determines the beam shape over a limited angle to $\approx 0.1\%$ of the maximum intensity by mapping an intense point source out to the limits of the proposed image. This is carried out for different hour angles; the source of interest is

also mapped at the same hour angles, and then the beam is iteratively subtracted from the source image. After subtractions, the images are compared to eliminate artifacts, so that a set of intensities as a function of position remains. This is convolved with a Gaussian beam having the telescope resolution. The continuum image of Orion at 6 cm (Fig. 11.5) was produced using such a process.

For spectral line radiation we meet with more complexity, since there are different features at different frequencies; also T_b has a wide distribution for H I. For CO line emission at $\lambda = 2.6$ mm the problems are less since: (1) CO is less widely distributed than H I, and (2) an 8′ FWHP beam requires a 100 m dish at 21 cm, while at 2.6 mm, only a 1.2 m antenna with a cassegrain focus is needed. Use of this focus allows low feed leg blockage. All other molecular and recombination line measurements are probably free from this effect because these lines are emitted by discrete sources that cover only a tiny fraction of the sky.

The contribution to T_A from the stray pattern puts rather stringent limits on the calibration schemes used extensively in galactic 21 cm line work. The calibration of H I data makes use of the following procedure: A number of regions have been measured with well-calibrated, but lower angular resolution horn antennas so that accurate line profiles $T_b(v)$ can be given. If these regions are now mapped with larger parabolic antennas, scaling factors for transforming the measured output of the profiles into T_b values can be determined directly. Unfortunately both the output of the larger antenna and the published reference profiles are contaminated by time-dependent stray radiation. Kalberla et al. (1980, 1982) have shown how to correct the reference spectra but stray radiation contributions are still present in the corrected data. It will thus be very difficult to achieve a T_b scale with a precision of better than 10%. Higher accuracies are needed, however to estimate the column density of galactic H I at high galactic latitudes. This is critical for a comparison with data, such as X ray emission, measured toward extragalactic sources. This situation is considerably better for measurements done with an off axis paraboloid such as the GBT. Clearly, H I data taken with the GBT are to be preferred to data taken with e.g. the Effelsberg 100 m telescope.

8.4.5 Spectral Line Observing Strategies

In radio astronomy line radiation is almost always only a small fraction of the total power received; the signal sits on a large pedestal of wide band noise signals contributed by different sources: the system noise, spillover from the antenna and in some cases, a true background noise. To avoid the stability problems encountered in total power systems (see Chap. 4) the signal of interest must be compared with another signal that contains the same total power and differs from the first only in that it contains no line radiation. To achieve this aim modern spectral line receivers usually permit four different observing modes that differ only in the way the comparison signal is produced.

1) **Switching Against an Absorber.** If receiver fluctuation time scales are too short for switching in the total power mode, the receiver can be connected alternately to the antenna and to a matched resistive load. By noise injection the output power in both switch positions can be equalized, and then the difference of the signals is the line radiation. A good balance for both switch positions is essential if good results are to be achieved (see also Sect. 4.2.2 for a discussion of the receiver stability achieved using this method). Particularly serious residual instrumental baseline ripples may be present in spectra obtained in this observing mode. This method is used only in exceptional circumstances today.

2) **Frequency Switching**. For many sources, spectral line radiation is a narrow-band feature, that is, the emission is centered at v_0, present over a small frequency interval, Δv, with $v_0/\Delta v \approx 10^{-6}$. If all other effects vary very little over Δv, then changing the frequency of a receiver by perhaps $10\Delta v$ produces a comparison signal with the line shifted. It is assumed that other contributions hardly differ. The final spectrum is proportional to the difference of these two measurements. Such "frequency switched" measurements can be done with almost any speed, and produces a particularly good compensation for wide-band atmospheric instabilities. Such observations can be made for mm wave radiation even in poor weather conditions but functions best for lines having widths of less than a few MHz. If the spectral line is included in the analyzing band in both the signal and the reference phases, the integration time is doubled. Early measurements of the 21 cm line of neutral hydrogen from the plane of our galaxy were made using this method.

3) **Position Switching and Wobbler Switching.** The received signal "on source" is compared with another signal obtained at a nearby position in the sky. If the emission is rather extended and the atmospheric effects are large (for example in the case of galactic Carbon Monoxide emission), one may use two reference measurements, one at a higher, and the other at a lower elevation. A number of conditions must be fulfilled: (1) the receiver is stable so that any gain and band-pass changes occur only over time scales which are long compared to the time needed for position change, and (2) there is little line radiation at the comparison region. If so, this method is efficient and produces excellent line profiles. This method is especially advantageous if baseline ripples are a problem, since these can be cancelled quite well if this method is used, provided that the broadband emission from sky, ground or continuum sources are similar at both positions. A variant of this method is wobbler switching. This is very useful for compact sources, especially in the mm and sub–mm range.

4) **On the Fly Mapping.** This very important observing method is an extension of method (3). In this procedure, one takes spectral line data at a rate of perhaps one spectrum or more per second. As with total power observing, usually one first takes a reference spectrum, and then takes data along a given direction. Then one changes the position of the telescope in the perpendicular direction, and repeats the procedure until the entire region is sampled. Because of the short integration times an entire image of perhaps $15' \times 15'$ taken with a $30''$ beam could be finished in roughly 20 min. At each position, the signal to noise ratio may be

low, but the procedure can be repeated. With each data transfer, the telescope position is read out. Even if there are absolute pointing errors, over this short time and small angle the relative positions where spectra were taken are accurate. The accuracy of the result is improved because the spectra are oversampled and weather conditions are uniform over the region mapped. To produce the final image the individual spectra are placed on a grid and then averaged.

8.5 The Confusion Problem

8.5.1 Introduction

The classical approach to the topic of discrete "source confusion" as presented in the 1950s and 1960s was done to provide completeness for source surveys, in that one could count sources to limits that were not possible with instruments then available. That is, the noise in excess of that from the receivers and earth's atmosphere was assumed to be caused by the sum of the sources that are too weak or numerous to be detected individually.

At first, the analysis used was in terms of obtaining a result for a given limit of source flux density, rather than a given level of instrumental response. This was noted by Condon (1974); we will follow the approach of Condon.

We define the effective beam, Ω_e, as

$$\Omega_e = \int [P_{\mathrm{n}}(\boldsymbol{n})]^{\gamma-1}\,\mathrm{d}\Omega. \tag{8.48}$$

where $[P_{\mathrm{n}}(\boldsymbol{n})]$ is the normalized antenna power pattern.

The deflection of the instrument is

$$x = fS$$

where $f(\theta,\phi)$ is the antenna response. Then the average differential number of sources with flux densities between S and $S + \mathrm{d}S$, $\mathrm{d}\bar{n}$, is

$$d\bar{n} = \int n(S)\,\mathrm{d}\Omega dx \tag{8.49}$$

A standard expression for $n(S)$ is

$$n(S) = kS^{-\gamma} \tag{8.50}$$

Using these expressions, we have

$$d\bar{n} = k \int f^{-\gamma-1}\,S^{-\gamma}\,\mathrm{d}\Omega\,\mathrm{d}x \tag{8.51}$$

or

$$d\bar{n} = k \, f^{-\gamma-1} \, S^{-\gamma} \, \Omega_e \, dx \tag{8.52}$$

The distribution of $d\bar{n}$ is Poisson, so the mean value is the variance. The sum of the variances of the responses from 0 to the cut off D_c is

$$\sigma^2 = \int_0^{D_c} x^2 \, d\bar{n} \tag{8.53}$$

When Eq. (8.52) is substituted into Eq. (8.53), one obtains the result

$$\sigma = \left(\frac{k \, \Omega_e}{3 - \gamma}\right) D_c^{3-\gamma} \tag{8.54}$$

For $2, \gamma < 3$ Eliminating D_c by a factor q times σ, we have

$$\sigma = \left(\frac{q^{3-\gamma}}{3 - \gamma}\right)^{\frac{1}{\gamma-1}} (k \, \Omega_e)^{\frac{1}{\gamma-1}} \tag{8.55}$$

The first factor in Eq. (8.55) depends only on the number count exponent and the choice of q. The second factor is related to the angular resolution as given by Eq. (8.48).

Therefore, the higher the angular resolution, the deeper the survey before confusion occurs. Usually there is a uniform flux density cutoff limit S_c This flux density is the *confusion limit* for the telescope at the chosen frequency. For q a value of 5 is usually considered to be acceptable; this corresponds to a probability of 10^{-6} for erroneous source identifications.

The problem of source confusion remains, but the interest in radio source counts as a cosmological tool has declined in the last 30 years, since radio sources evolve strongly with time. Thus, the interpretation of source counts in terms of a universe consisting of a collection of sources with the same characteristics, but with different distances, is no longer accepted. It is clear that progress will be made only by studying statistically significant collections of sources at different redshifts. Thus accurate flux densities and positions are needed. For this reason, the effect of confusion must be reduced by making beamsizes and reducing the effect of sidelobes. Such concepts are *not* limited to radio astronomy, but must be considered in the interpretation of deep surveys carried out in the infrared.

Problems

1. Investigate the effect of the earth's atmosphere on radio observations by using a single layer atmosphere (Eq. 1.37). Suppose we know that the atmospheric optical depth, τ, is 0.1, and the temperature is 250 K.

(a) What is the excess noise from the atmosphere, and what is the reduction in the intensity of a celestial source?

(b) Repeat for $\tau = 0.5, 0.7, 1.0, 1.5$.

(c) If τ is related to the optical depth in the zenith by $\tau = \tau_z/\sin(\text{elv})$, determine the increase in τ between $30°$ and $20°$ elevation. (Elevation is measured relative to the horizon.)

(d) Repeat this calculation for the increase between $20°$ and $19°$, then $20°$ and $15°$.

(e) For spectral line measurements, one is interested in a comparison of the responses of the receiver system over a (relatively) small frequency interval. Consider the measurement of a 10 mK spectral line through an atmosphere with $\tau = 0.2$, if the receiver noise is 100 K. Repeat this calculation for a receiver noise of 20 K.

2. A standard method to determine atmospheric τ values employs a receiver to determine the emission of the earth's atmosphere at 225 GHz. Suppose this emission is found to be 15 K at elevation $90°$, 18 K at $60°$, 30 K at $30°$, and 42 K at $20°$. If the temperature of the atmosphere is 250 K, what is the zenith τ? Is the curve in Fig. 8.1 consistent with ratios of zenith τ to that at 225 GHz are 3.4 (at 340 GHz), 6.7 (at 410 GHz), 9.9 (at 460 GHz) and 19.0 (at 490 GHz).

3. Suppose you are observing at 1 cm wavelength with a filled aperture telescope. When pointed toward cold sky, in the zenith, your system noise temperature is twice what you expect. Normally the receiver noise temperature is 70 K and system noise temperature is 100 K. Your partner notices that the radio telescope is filled with wet snow. Assuming that the snow has a temperature of 260 K, and is a perfect absorber at 1 cm, how much of the telescope surface is covered with snow?

4. A group observe sources at 1.3 cm at elevations between $8°$ and $11°$. If the zenith optical depth is $\tau_z = 0.1$, use an assumed dependence of $\tau = \tau_z/\sin(\text{elv})$ to determine τ at the lowest and highest elevations. These astronomers see at most a 30% change in τ over this range of elevations. Is this reasonable? If the *receiver* noise is 40 K, what is the *system* noise, including the atmospheric contribution, for a 200 K atmosphere, at these elevations? The observations are mostly of spectral lines; how much is the attenuation? The temperature scale is calibrated using a nearby source with peak main beam brightness temperature 16 K. What is the RMS error for each continuum data point, from noise only, if the bandwidth used is 40 MHz and the integration time is 1 s?

5. Use the Rayleigh–Jeans approximation to calculate the numerical relation between flux density, S_ν and brightness temperature, T_B, if the source and beam have Gaussian shapes. S_ν must be in units of janskys ($= 10^{-26}$ W m^{-2} Hz^{-1}), wavelength must be in cm, and the observed angle θ_0 in arc min.

6. For a Gaussian-shaped source of actual angular size θ_{source} and observed size θ_{observed}, find the relation between the apparent or main beam brightness temperature, T_{MB}, and the actual brightness temperature, T_B. (Use the fact that the flux density of a discrete source must not depend on the telescope.) Show that $T_B > T_{MB}$. Show that the observed or apparent, actual and telescope beam sizes, θ_{observed}, θ_{source} and θ_{beam}, are related by $\theta_{\text{observed}}^2 = \theta_{\text{actual}}^2 + \theta_{\text{beam}}^2$.

7. An outburst of an H_2O maser (at 22.235 GHz) in the Orion region (distance from the Sun 500 pc) gave a peak flux density of 10^6 Jy over a 1 MHz band. If this maser radiation were measured with the 100 m telescope, which has a collecting area of 7800 m^2, and antenna efficiency 0.4, what is the peak power? If the safety level for microwave radiation for humans is $10\,mW\,cm^{-2}$, at what distance would the Orion maser be a threat for humans?

8. Use the Rayleigh–Jeans relation to calculate the flux density of the Sun at 30 GHz if the disk has a diameter of $30'$ at a uniform surface temperature 5800 K? Suppose we had a 40 m radio telescope with effective collecting area 1000 m^2. What is the value of T_{MB}? If $\eta_A = 0.5$ and $\eta_{MB} = 0.65$, what is T_A?

9. Use Eq. (8.20) to determine the peak main beam brightness temperature of the planetary nebula NGC7027 at 1.3 cm with the 100 m telescope ($S(\text{Jy}) = 5.4$ Jy, $\theta_0 = 43''$).
(a) If the actual source size is $\theta_s = 10''$, use Eq. (8.22) to determine the actual source brightness temperature T_s. Then use Eq. (1.37), with $T_0 = 0$, and $T_\nu = 14000$ K to determine the peak optical depth of this region at 1.3 cm.

10. A celestial source has a flux density of 1 Jy at 100 MHz. If the angular size is $10''$, and source and telescope beams are Gaussians, estimate the source brightness temperature in the Rayleigh–Jeans limit. Repeat this for an observing frequency of 1 GHz.

11. The planet Venus is observed at the distance of closest approach, a distance of 0.277 AU. The radius of Venus is 6100 km. What is the full angular width of Venus in arc seconds? Suppose the measured brightness temperature of Venus at 3.5 cm wavelength in a telescope beam of $8.7'$ is 8.5 K. What is the actual surface brightness temperature of Venus?

12. In the sub-millimeter range, sky noise dominates, but one wants to have the most sensitive receivers possible. Is this a contradiction? If not, why not?

13. The APEX submillimeter telescope on the ALMA site has a diameter of 12 m, an estimated beam efficiency of 0.5 at a wavelength of 350 μm. At 350 μm the atmospheric transmission is 5%.
(a) Show that this is equivalent to a τ of 3.
(b) What is the sky noise for this situation if the physical temperature of the sky is 200 K?
(c) If the *receiver* noise is 50 K, what is the total *system* noise?
(d) Suppose you plan to measure a small diameter source with a flux density of 0.1 Jy. After what length of time will you have a signal-to-noise ratio of unity if the receiver bandwidth is 2 GHz?

14. Spectral line observations are carried out using position switching, that is the "on–off" observing mode. Thus effects of ground radiation should cancel in the difference spectrum. However, there is usually a residual instrumental baseline found in the case of centimeter wavelength observations. The amplitude of this residual

instrumental baseline is found (with the 100 m telescope) to be $\sim 10^{-3}$ of the contin-
uum intensity of the source being observed. This effect is caused by the correlation
of signal voltage E_i, with that reflected by the primary feed horn, E_r. How much
power flux, E_r^2, (in W m^{-2}) relative to E_i, is reflected from the feed?

15. A search for dense molecular gas in the Orion cloud shows the presence of 125
sources, each with a FWHP of $1'$. The region searched is $15'$ by $120'$. If the beam
size is $20''$, what is the mean number of sources per angular area? Now use Poisson
statistics $P = e^{-m} m^n / n!$ where n is the number of expected sources, and m is the
mean, to find the probability of finding a dense clump of gas in this region if one
uses a $20''$ beam. What is the chance of finding two such sources?

16. (a) In an extragalactic survey, the average number of sources per beam is 0.04.
Use Poisson statistics to find the chance of finding 2 or 3 sources in the same beam?
(b) Use these results to estimate the number of beam areas per source needed to
insure that source confusion is a small effect.

17. (a) Derive the result in (8.55) showing all steps.
(b) For radio telescopes, the one dimensional power pattern is $y(x) =$
$A \exp \left(-\frac{4\ln 2 x^2}{\theta_{1/2}^2} \right)$ Use this expression to evaluate (8.48).
(c) Calculate $k = \gamma N_c S_c^{-\gamma}$ for $\gamma = 1.5$, $q = 5$, $\Omega = 80 \times 100$ arc sec, $N_c = 10^5$ per
steradian, and $S_c = 10^{-28}$ W m^{-2} Hz^{-1}.

Chapter 9
Interferometers and Aperture Synthesis

9.1 The Quest for Angular Resolution

In Chap. 6, we have shown that from diffraction theory, the angular resolution of a radio telescope is $\theta = k\lambda/D$, where θ is the angular resolution, λ is the wavelength of the radiation received, D is the diameter of the instrument and k is a factor of order unity that depends on details of antenna illumination. For a given wavelength, to improve this angular resolution, the diameter D must be increased. Materials limit the size of a single telescope to ~ 300 m. As shown by Michelson (see, e.g. Jenkins and White 2001), a resolving power $\theta \approx \lambda/D$ can be obtained by coherently combining the output of two reflectors of diameter $d \ll D$ separated by a distance D. We will show that this is the case in Sect. 9.2.

A more complex topic is aperture synthesis, that is, producing high quality images of a source by combining a number of independent measurements in which the antenna spacings cover an equivalent aperture. In Sect. 9.3 of this chapter, we give an introduction to the principles of aperture synthesis. More detailed accounts are to be found in Thompson, Moran and Swenson (2001), Dutrey (2000) or Taylor, Carilli and Perley (1999). Techniques similar to those used in aperture synthesis have been applied to radar as "Synthetic Aperture Radar" or SAR (see e.g. Mensa 1991)

9.1.1 The Two Element Interferometer

The basic principle governing angular resolution can be understood from a consideration of Fig. 9.1. In panel (a) is the response of a single uniformly illuminated aperture of diameter D. In panels (b) and (c) we show the response of a two element interferometer consisting of two small antennas (diameter d) separated by a distance D and $2D$, where $d \ll D$. The interferometer response is obtained from the multiplication of the outputs of the two antennas. The uniformly illuminated aperture has a dominant main beam of width $\theta = k\lambda/D$, accompanied by smaller secondary maxima, or sidelobes. There are two differences between the case of a single dish response compared to the case of an interferometer. First, for an interferometer,

Fig. 9.1 Power patterns for different antenna configurations. The horizontal axis in this figure is angle. Panel (**a**) shows that of a uniformly illuminated full aperture with a diameter D. This full width to half power (FWHP) is $k\lambda/D$, with $k \approx 1$. In panel (**b**) we show the power pattern of a two element multiplying interferometer consisting of two antennas of diameter d spaced by a distance D where $d \ll D$. In panel (**c**) we show the power pattern of the interferometer system described in (**b**) but now with a spacing $2D$
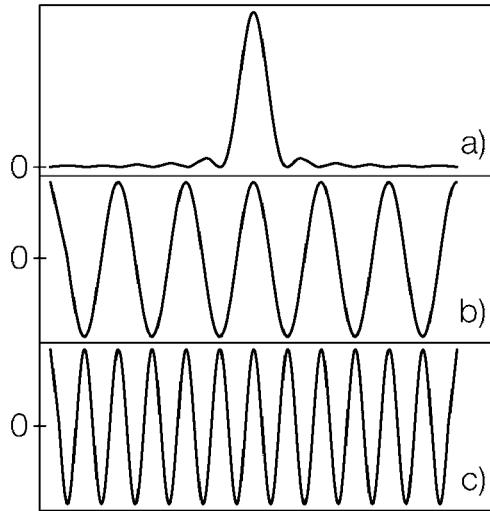
the nomenclature is different. Instead of ′main beam and sidelobes′ one speaks of ′fringes′. There is a central fringe (or ′white light′ fringe in the analogy with Young's Two Slit experiment) and adjacent fringes. Second, as we will show in Sect. 9.2.3.2, for the correlation of the outputs of two antennas, the fringes are centered about zero; this procedure improves the dynamic range of the measurements since the large total power output of each antenna is supressed (also the signal-to-noise ratio is better; see problem 9). This comes with a cost: some of the information (i.e. total power) is not available, so for a given spacing only source structure comparable to (or smaller than) a fringe is recorded fully. We compare responses of two systems in Fig. 9.1. In panel (a) is a single dish of diameter D. In (b) we show the case of an interferometer composed of two small dishes (with dish diameter d much smaller than the separation D) there is no prominent main beam and the sidelobe level does not decrease with increasing angular offset from the axes of the antennas. In panel (c) the separation of the two dishes is $2D$. Comparing the width of the fringes in panels (b) and (c) one finds that by doubling the separation D of the small antennas, the fringe width is halved. For the interferometer spacing (usually referred to as *the baseline*) D, in panel (b) the resolving power of the filled aperture is not greatly different from the single dish in panel (a), but the collecting area of this two element interferometer is smaller. For larger spacings, the interferometer angular resolution is greater.

By increasing D, finer and finer source structure can be measured. Combining the outputs of independent data sets for spacings of D and $2D$ shows that these select different structural components of the source. Finer source structure can be recorded if in addition, nD antenna spacings are measured. Such a series of measurements can be made by increasing the separation of two antennas whose outputs are coherently combined.

A general procedure, *aperture synthesis*, is now the standard method to obtain high quality, high angular resolution images. The first practical demonstration of Aperture Synthesis in radio astronomy was made by Ryle and his associates. Aperture synthesis allows us to reproduce the imaging properties of a large aperture by sampling the radiation field at individual positions within the aperture. Superficially, this appears to be similar to the Kirchhoff diffraction theory of Sect. 6.4, which leads to (6.54). However in the case of Kirchhoff diffraction theory, the individual samples would have to be measurements of the E or B fields including phase, at each position. Thus, sampling would be a complex process. In analogy with the approach used by Michelson in the optical wavelength range, the advance in radio astronomy was to measure the *mutual coherence function* and to show that these results were sufficient to produce images. This approach involved inverting the van Cittert-Zernike relation (see Appendix G), which allows one to derive the intensity distribution. The mutual coherence function describes the cross–correlation of the radiation field at two given points. This has the dimension of power, and can be easily measured using two element interferometers. This can be done by sequential measurements with different antenna spacings. Using this approach, a remarkable improvement in radio astronomical imaging was possible.

In the following, both the hardware and the software tools for radio interferometry will be described. In Appendix G we outline some concepts, mainly the van Cittert-Zernike theorem. These are needed to proceed from the E field emitted by an extended source to the production image. The first step is the description of a phase stable two-element interferometer. This is presented in the next section. We follow with an introduction to aperture synthesis in Sect. 9.3. The sensitivity of such a device is discussed in Sect. 9.5. Thus far, we have described interferometers whose individual antennas are connected in real time by cables, optical fibers or microwave links. An extension of this situation is *Very Long Baseline Interferometer* (VLBI). In VLBI, the antennas have such large separations that real time links are difficult. Currently, the link is achieved by precise timing, with data and time recorded at each antenna. Producing fringes requires aligning the data recorded at each antenna using the time signals. The production of images with VLBI requires additional software techniques. There is progress in using near real-time links via internet connections, however. This technique is known as e-VLBI.

9.2   Two-Element Interferometers

The coherence function $\Gamma(\boldsymbol{u}, \tau)$ is measured by correlating the outputs of two antenna systems. The simplest example of this process is a two-element interferometer. Let us assume that the interferometer consists of two antennas A_1 and A_2 separated by the distance $\boldsymbol{B}$ (directed from A_2 to A_1), and that both antennas are sensitive only to radiation of the same state of polarization (Fig. 9.2).

A plane electromagnetic wave (from a very distant source) of amplitude E induces the voltage U_1 at the output of antenna A_1

Fig. 9.2 A schematic
diagram of a two-element
correlation interferometer.
The antenna output voltages
are V_1 and V_2; the
instrumental delay is τ_i and
the geometric delay is τ_g
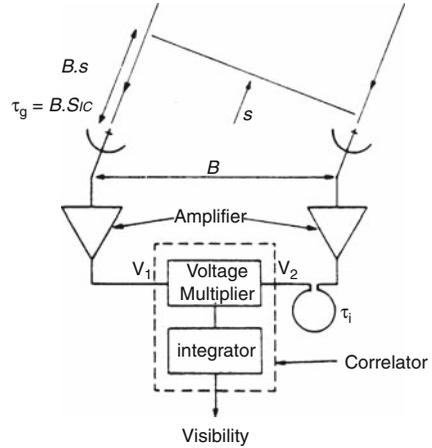
$$U_1 \propto E\,\mathrm{e}^{\,\mathrm{i}\omega t}\,, \tag{9.1}$$

while at A_2 we obtain

$$U_2 \propto E\,\mathrm{e}^{\,\mathrm{i}\omega(t-\tau)}\,, \tag{9.2}$$

where τ is the geometric delay caused by the orientation of the interferometer base-line $\boldsymbol{B}$ relative to the direction of the wave propagation. For simplicity, in (9.1) and (9.2) we have neglected receiver noise and instrumental effects. The outputs will be correlated. In a *correlation* the signals are input to a multiplying device followed by an integrator. The output is proportional to

$$R(\tau) \propto \frac{E^2}{T} \int_0^T \mathrm{e}^{\,\mathrm{i}\omega t}\,\mathrm{e}^{\,-\mathrm{i}\omega(t-\tau)}\,\mathrm{d}t\,.$$

If T is a time much longer than the time of a single full oscillation, i.e., $T \gg 2\pi/\omega$ then the average over time T will not differ much from the average over a single full period; that is

$$R(\tau) \propto \frac{\omega}{2\pi} E^2 \int_0^{2\pi/\omega} \mathrm{e}^{\,\mathrm{i}\omega\tau}\,\mathrm{d}t$$

$$\propto \frac{\omega}{2\pi} E^2 \mathrm{e}^{\,\mathrm{i}\omega\tau} \int_0^{2\pi/\omega} \mathrm{d}t\,,$$

resulting in

$$\boxed{R(\tau) \propto \tfrac{1}{2}E^2\,\mathrm{e}^{\,\mathrm{i}\omega\tau}} \qquad . \tag{9.3}$$

The output of the correlator + integrator thus varies periodically with τ, the delay time; this output is the mutual coherence function (Appendix H, Eq. G.3) of the received wave. If the relative orientation of interferometer baseline $\boldsymbol{B}$ and wave propagation direction $\boldsymbol{s}$ remain invariable, τ remains constant, so does $R(\tau)$. But since $\boldsymbol{s}$ is slowly changing due to the rotation of the earth, τ will vary, and we will measure *interference fringes* as a function of time.

In order to understand the response of interferometers in terms of measurable quantities, we consider a two-element system. The basic constituents are shown in Fig. 9.2. If the radio brightness distribution is given by $I_\nu(\boldsymbol{s})$, the power received per bandwidth $d\nu$ from the source element $d\Omega$ is $A(\boldsymbol{s})I_\nu(\boldsymbol{s})\,d\Omega\,d\nu$, where $A(\boldsymbol{s})$ is the effective collecting area in the direction $\boldsymbol{s}$; we will assume the same $A(\boldsymbol{s})$ for each of the antennas. The amplifiers are assumed to have constant gain and phase factors which we neglect for simplicity.

The output of the correlator for radiation from the direction $\boldsymbol{s}$ (Fig. 9.2) is

$$r_{12} = A(\boldsymbol{s})\,I_\nu(\boldsymbol{s})\,e^{i\omega\tau}\,d\Omega\,d\nu \tag{9.4}$$

where τ is the difference between the geometrical and instrumental delays τ_g and τ_i. If $\boldsymbol{B}$ is the baseline vector for the two antennas

$$\tau = \tau_g - \tau_i = \frac{1}{c}\boldsymbol{B}\cdot\boldsymbol{s} - \tau_i \tag{9.5}$$

and the total response is obtained by integrating over the source S

$$R(\boldsymbol{B}) = \iint_\Omega A(\boldsymbol{s})I_\nu(\boldsymbol{s})\exp\left[i2\pi\nu\left(\frac{1}{c}\boldsymbol{B}\cdot\boldsymbol{s} - \tau_i\right)\right]d\Omega\,d\nu \qquad . \tag{9.6}$$

This function $R(\boldsymbol{B})$, the *Visibility Function* is closely related to the mutual coherence function of the source (G.17) but, due to the power pattern $A(\boldsymbol{s})$ of the individual antennas, it is not identical to $\Gamma(\boldsymbol{B},\tau)$ as given in Appendix G. For parabolic antennas it is usually assumed that $A(\boldsymbol{s}) = 0$ outside the main beam area so that (9.6) is integrated only over this region.

A one dimensional version of (9.6), with a baseline B, $\nu = \nu_0$ and $\tau_i = 0$, is

$$R(B) = \int A(\theta)\,I_\nu(\theta)\exp\left[i2\pi\nu_0\left(\frac{1}{c}\boldsymbol{B}\cdot\boldsymbol{\theta}\right)\right]d\theta \tag{9.7}$$

9.2.1 Hardware Requirements

In an interferometer, the $\boldsymbol{E}$ fields received by each antenna (which can be separated by distances of up to hundreds of kilometers for the MERLIN system) must be coherently combined. This involves the use of heterodyne receiver systems. One shifts the frequency of the input from the sky frequency to a frequency that allows

transmission to a central point where the outputs of the different antennas are combined pairwise. This frequency shifting process must preserve the phase of the signal. For interferometers operating at frequencies of a few GHz and above, the limits to the performance are set by the stability of the local oscillator (LO). For such systems a small part of the LO signal is reflected at each antenna. This reflected signal is compared to the input LO signal. Any known, calculated, geometric shift in phase is corrected by changing the cable lengths or digital delays in a so-called "line length corrector". However, there may be unwanted, time variable delays in the signal phase that corrupt the geometrical delay. For example, at low frequencies, there are fluctuations in the ionosphere, while at high frequencies, there are fluctuations in the water vapor.

Presently, the antenna outputs are phase switched by shifting the first LO at the antenna by 180° synchronously with an equal change of phase in a second LO at the central point where the outputs are combined. In this way, the signal to be combined is not affected, but the instrumental effects are greatly reduced. Compared to single dish measurements this procedure has the advantage that the signals are correlated so receiver gains and offsets are less important.

In addition to the phase of the LO phase, the interfometer baselines must be constant over the time needed to carry out a set of measurements of calibrator and source. Such "baseline corrections" can be made in the post-data taking phase, but it is very useful to have a fairly accurate estimate of the baseline (to a fraction of a wavelength if possible) before starting the measurements. This is carried out by measurement of a number of calibration sources with known positions over a range of hour angles.

As will be discussed in the next section, the use of a very large bandwidth will reduce the field of view of an interferometer. This can be avoided by dividing the continuum band into smaller wavelength intervals, and using the second LO to place the output on the central, or "white light" fringe.

9.2.2   Calibration

Two quantities that must be calibrated for continuum measurements are amplitude and phase. In addition, for spectral line measurements the instrument passband must also be calibrated.

The amplitude scale is calibrated using methods that are similar to those used for single dish measurements. This consists of using the response of each antenna to determine the system noise of the receiver being used. In the centimeter range, the atmosphere plays a small role while in the millimeter and sub-mm wavelength ranges, the atmospheric effects must be accounted for. For phase measurements, a suitable point-like source with an accurately known position is required to determine the instrumental phase $2\pi\nu\tau_i$ in Eq. 9.6. For interferometers, the calibration sources are usually unresolved or point-like sources. Most often these are extra-galactic sources (of necessity these are compact sources; sometimes these may be

time variable). To calibrate the response in units of flux density or brightness temperature, these measurements must be referenced to a thermal calibrator.

The calibration of the instrument passband is carried out by an integration of an intense source to determine the channel-to-channel gains and offsets. The amplitude, phase and passband calibrations are carried out (at least) before the source measurements. The passband calibration is usually carried out once per observing session. The amplitude and phase calibrations are made more often. The frequency depends on the stability of the electronics and weather. At millimeter wavelengths, the calibrations are usually made every few minutes, but may have to be made more often in bad weather or at shorter wavelengths. If weather demands that frequent measurements of calibrators are required, this is referred to as *fast switching*.

9.2.3 Responses of Interferometers

9.2.3.1 Finite Bandwidth

So far, in (9.6), we have assumed that the radiation is monochromatic. This is certainly not the case in most applications. Equation (9.6) can be used to estimate the effect of a finite bandwidth Δv. The geometric delay $\tau_g = \frac{1}{c}\boldsymbol{B}\cdot\boldsymbol{s}$ is by definition independent of frequency, but the instrumental delay τ_i may not be. Adjusting τ_i the sum $\tau = \tau_g - \tau_i$ can be made equal to zero for the center of the band. Introducing the relative phase of a wave by

$$\varphi = \left[\frac{c\tau}{\lambda}\right]_{\text{fractional part}}$$

we obtain

$$\varphi = \frac{1}{\lambda}\boldsymbol{B}\cdot\boldsymbol{s} + \varphi_i \,, \tag{9.8}$$

where φ_i is the instrumental phase corresponding to the instrumental delay. This phase difference varies across the band of the interferometer Δv by

$$\Delta\varphi = \frac{1}{\lambda}\boldsymbol{B}\cdot\boldsymbol{s}\,\frac{\Delta v}{v}. \tag{9.9}$$

The fringes will disappear when $\Delta\varphi \simeq 1$ radian. As can be seen the response is reduced if the frequency range, that is, the bandwidth, is large compared to the time delay caused by the separation of the antennas. For large bandwidths, the loss of visibility can be minimized by adjusting the delay τ_i between antennas so that the time delay between the antennas (see Fig. 9.2) is negligible. In effect, this is only possible if the exponential term in Eq. (9.6) is kept small. In practice, this is done by inserting a delay between the antennas so that $1/c\boldsymbol{B}\cdot\boldsymbol{s}$ equals τ_i. In the first interferometric systems this was done by switching lengths of cable into the system; currently this is accomplished by first digitizing the signal after conversion to an

intermediate frequency, and then using digital shift registers. In analogy with the optical wavelength range, this adjustment of cable length is equivalent to centering the response on the central, or *white light fringe* in Young's two-slit experiment.

The reduction of the response caused by finite bandwidth can be estimated by an integration of Eq. (9.6) over frequency. Taking $A(s)$ and $I_v(s)$ to be constants, and integrating over a range of frequencies $\Delta v = v_1 - v_2$. Then the result is an additional factor, $\sin(\Delta v \tau)/\Delta v \tau$ in Eq. (9.6). This will reduce the interferometer response if $\Delta \varphi \gtrsim 1$. For typical bandwidths of $100\,\text{MHz}$, the offset from the zero delay must be $\ll 10^{-8}\,\text{s}$. This adjustment of delays is referred to as *fringe stopping*. This causes the response of (9.6) to lose a component. To recover this input, an extra delay of a quarter wavelength relative to the input of the correlator is inserted, so that the sine and cosine response in (9.6) can be measured. In digital cross-correlators, (see Sect. 5.4.3.2), the sine and cosine components are obtained from the positive and negative delays. The component with even symmetry is the cosine component, while that with odd symmetry is the sine component.

9.2.3.2   Source Size and Minimum Spacing

We now consider an idealized (square) source, of shape $I(v_0) = I_0$ for $\theta < \theta_0$ and $I(v_0) = 0$ for $\theta > \theta_0$. In addition, we take the primary beamsize of each antenna to be much larger than the source size, so the beam size of each antenna can be neglected. We define θ_b the fringe spacing of the interferometer as $\frac{\lambda}{B}$, The result is

$$R(B) = A I_0 \cdot \theta_0 \exp\left[i\pi \frac{\theta_0}{\theta_b}\right] \left[\frac{\sin(\pi\theta_0/\theta_b)}{(\pi\theta_0/\theta_b)}\right] \tag{9.10}$$

The first terms are normalization and phase factors. The important term, in the second set of brackets, is a $\sin x/x$ function. If $\theta_0 \gtrsim \theta_b$, the interferometer response is reduced. This is sometimes referred to as the problem of "missing short spacings".

9.2.3.3   Bandwidth and Beam Narrowing

In 9.2.3.1, we noted that on the *white light fringe* the compensation must reach a certain accuracy to prevent a reduction in the interferometer response. However for a finite primary antenna beamwidth, A, this cannot be the case over the entire beam. In Fig. 9.4 we show the geometry for a two element interferometer used to measure a source at an offset angle θ_{offset}. For two different wavelengths λ_l and λ_s, there will be a phase difference

$$\Delta\phi = 2\pi d \left[\frac{\sin(\theta_{\text{offset}})}{\lambda_s} - \frac{\sin(\theta_{\text{offset}})}{\lambda_l}\right]$$

converting the wavelengths to frequencies, and using $\sin\theta = \theta$, we have

$$\Delta \phi = 2\pi \, \theta_{\text{offset}} \frac{d}{c} \Delta \nu$$

With use of the relation $d = \frac{\lambda}{\theta_b}$, we have

$$\Delta \phi = 2\pi \frac{\theta_{\text{offset}}}{\theta_b} \frac{\Delta \nu}{\nu} \qquad (9.11)$$

The effect in Eq. (9.11) is most important for continuum measurements made with large bandwidths. This effect can be reduced if the continuum measurements are carried out using a series of contiguous Intermediate Frequency (IF) sections, where for each of these IF sections, an extra delay is introduced to center the response at the value which is appropriate for that wavelength. This arrangement introduces extra complexity. However this compensation is performed after digitization. The cost of such digital components has decreased over the last decades, so that these systems have now become standard.

9.2.3.4 Multibeam Interferometer Systems

The concepts presented in the previous sections can be used to understand the operation of interferometer arrays such as LOFAR, the Low Frequency Array and the Allen Telescope Array or ATA. Both arrays are in the construction phase. These consists of many individual fairly small antennas spread over a large region, connected by optical fibers. For LOFAR the antennas are sets of dipoles (see Sect. 6.3.1). The antennas of the ATA are small paraboloids. LOFAR will operate at rather low frequencies, where the combination of small physical size and low frequency results in a rather large primary antenna beam. Since the contribution of receiver noise can be quite small even at $\lambda = 1.3$ cm (see Eq. 5.20), the signal from each antenna can be amplified almost without degradation and then digitized. The digitized signals can be processed using different sets of digital delay lines to produce a number of fringes with zero delay. Some of these can be directed far from the axis of the individual small antennas. In this way, large portions of the sky can be measured simultaneously. A more ambitious version using similar design concepts is the SKA, or the Square Kilometer Array (see, e.g. Hall 2005). The concept of the SKA is not yet fully determined, but presently the plan is an array with the collecting area equal to 100 times that of a 100 m radio telescope. The SKA is more ambitious than either ATA or LOFAR in the size of collecting area and in the short wavelength limit, 1.3 cm. Since SKA is also planned to operate at long wavelengths, the design of the individual elements including feeds and the receivers may have to include a variety of different concepts.

Another type of multibeam interferometer system has been proposed for use at millimeter wavelengths. This involves the use of a number of cooled receiver systems located at the Cassegrain focus of each individual parabolic reflectors in the interferometer array. In this concept, there would be three rows of three receiver systems, each with an individual output. Thus, each antenna has nine simultaneous

main beams; this multibeam system can map extended sources at nine times the speed of a single beam interferometer. The complexity is nine times higher, but in contrast to LOFAR, ATA and SKA this complexity includes the quite expensive cooled receivers. In all of these examples, the data rates are much greater than the simplest interferometer systems and vastly greater than any single dish. Thus there will be a much higher demand for data processing facilities.

9.3 Aperture Synthesis

Aperture Synthesis is a designation for methods used to derive the intensity distribution $I_v(s)$ of a part of the radio sky from the measured visibility function $R(\textbf{B})$. To accomplish this we must invert the integral equation (9.6). This involves Fourier transforms. For even simple images, a large number of computations are needed. Thus Aperture Synthesis and digital computing are intimately connected. In addition, a large number of approximations have to be applied. We will outline the most important steps of this development without, however, raising any claims to completeness.

9.3.1 An Appropriate Coordinate System

To solve equation (9.6) for $I_v(s)$ by measuring $R(\textbf{B})$ a convenient coordinate system must be introduced for the two vectorial quantities s and $\textbf{B}$. The image center can be chosen as the position of zero phase. This geometry can be introduced using a unit vector s pointing towards origin chosen (Fig. 9.5)

$$s = s_0 + \boldsymbol{\sigma}\,, \qquad |\boldsymbol{\sigma}| = 1\,,$$

where s_0 is a conveniently chosen position close to the center of the region investigated. Substituting this, (9.6) can be written as

$$R(\textbf{B}) = \exp\left[\mathrm{i}\,\omega\left(\frac{1}{c}\textbf{B}\cdot s_0 - \tau_\mathrm{i}\right)\right]\mathrm{d}v \iint_S A(\boldsymbol{\sigma})I(\boldsymbol{\sigma})\exp\left(\mathrm{i}\frac{\omega}{c}\textbf{B}\cdot\boldsymbol{\sigma}\right)\mathrm{d}\boldsymbol{\sigma}\,. \tag{9.12}$$

The exponential factor extracted from the integral describes a plane wave which defines the phase of $R(\textbf{B})$ for the image center. The integral V of the intensity distribution $I(\boldsymbol{\sigma})$,

$$V(\textbf{B}) = \iint_S A(\boldsymbol{\sigma})I(\boldsymbol{\sigma})\exp\left(\mathrm{i}\frac{\omega}{c}\textbf{B}\cdot\boldsymbol{\sigma}\right)\mathrm{d}\boldsymbol{\sigma} \tag{9.13}$$

Since the phases of the correlated signals are adjusted to produce a zero delay at the image center, the visibility is referred to this position.

If the coordinate systems are chosen such that

$$\frac{\omega}{2\pi c}\mathbf{B} = (u, v, w), \qquad \frac{\omega \pm \delta\omega}{2\pi c} = \frac{f}{c}\left(1 \pm \frac{\Delta f}{f}\right),$$

where u, v, w are measured in units of the wavelength $\lambda = 2\pi c/\omega$. The direction $(0, 0, 1)$ is parallel to $\mathbf{s}_0$, u points in the local east direction while v points north; the vector $\boldsymbol{\sigma} = (x, y, z)$ is defined such that x and y are the direction cosines with respect to the u and v axes. Then the xy plane represents a projection of the celestial sphere onto a tangent plane with the tangent point (and origin) at $\mathbf{s}_0$ (Fig. 9.6). In these coordinates (9.13) becomes

$$V(u, v, w) = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} A(x, y) I(x, y)$$
$$\times \exp[i 2\pi(ux + vy + w\sqrt{1 - x^2 - y^2})] \frac{dx\,dy}{\sqrt{1 - x^2 - y^2}}.$$
$$(9.14)$$

The integration limits have been formally extended to $\pm\infty$ by demanding that $A(x, y) = 0$ for $x^2 + y^2 > l^2$; where l is the full width of the primary telescope beams. Equation 9.14 closely resembles a two dimensional Fourier integral; these would be identical if the term $w\sqrt{1 - x^2 - y^2}$ could be extracted from the integral. If only a small region of the sky is to be mapped then $\sqrt{1 - x^2 - y^2} \cong \text{const} \cong 1$ and (9.14) becomes

$$V(u, v, w)\,e^{-i 2\pi w} = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} A(x, y) I(x, y)\,e^{i 2\pi(ux + vy)}\,dx\,dy \qquad . \qquad (9.15)$$

The factor $e^{-i 2\pi w}$ is the approximate conversion required to change the observed phase of V to the value that would be measured with antennas in the uv plane:

$$V(u, v, w)\,e^{-i 2\pi w} \cong V(u, v, 0). \qquad (9.16)$$

Substituting this into (9.15) and performing the inverse Fourier transform we obtain

$$I'(x, y) = A(x, y) I(x, y) = \int\limits_{-\infty}^{\infty} V(u, v, 0)\,e^{-i 2\pi(ux + vy)}\,du\,dv \qquad , \qquad (9.17)$$

where $I'(x, y)$ is the intensity $I(x, y)$ modified by the primary beam shape $A(x, y)$. One can easily correct $I'(x, y)$ by dividing by $A(x, y)$.
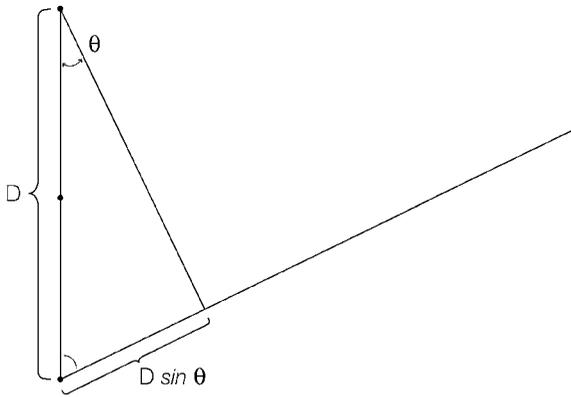
Fig. 9.3 A sketch to illustrate the angles used to derive the expression for the narrowing of the field caused by bandwidth

If only simple geometries are to be determined with a small number of measurements, models can be fitted to the distribution of V as a function of hour angle. Such distributions are shown in Fig. 9.3. This approach is limited to situations where only a few antennas are available (this has been the case in sub-mm interferometry).

In more detailed investigations, comparisons can be made between the correlated flux densities and the flux densities obtained with a single telescope. If the source structures are small compared to a single fringe, the single dish and interferometer flux densities should agree. If the flux densities remain constant as a function of the separation between the interferometer antennas, the source is said to be unresolved (i.e. point-like). If however, the angular size of the sources, θ_s, are extended over more than one fringe, the fringe visibilities (for a correlation interferometer) are both positive and negative, and the correlated flux densities are reduced by a factor $\sin x / x$, where $x = 2\pi \theta_s / B_m$ with B_m the baseline spacing. In this case, the flux density measured with correlation interferometers will be smaller than that measured with a single telescope, or with a smaller spacing between the antennas. If the flux densities are plotted against antenna spacings, a source size can be deduced. Usually a Gaussian fit is used, and the source size is referred to as an equivalent Gaussian size. An example is shown in Fig. 9.3b.

As the sensitivity of interferometer arrays improved and the number of antennas increased, it became clear that the simple representations of source structure in Fig. 9.3 were only a first approximation. The simple *Model Fits* illustrated in Fig. 9.4 have been replaced by images using the technique of *Aperture Synthesis* which involves solutions of (9.17). An example of this process is shown in Fig. 11.6. The detail in this image is vastly greater than the examples in Fig. 9.4. This improvement is due to:

 i the increased number of independent measurements of the visibilities of this
    source,
 ii better receiver sensitivity and
iii more sophisticated data processing, especially the increased speed and storage
    capacity of digital computers and to new image processing algorithms.

(a)   Visibility function   Brightness distribution

(b)

(c)

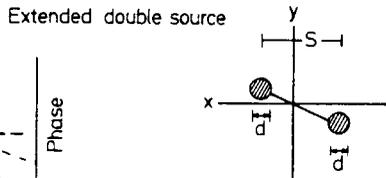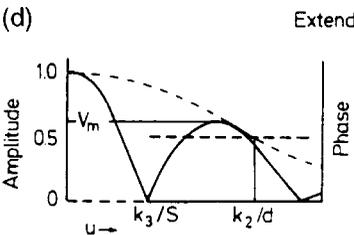$$V_m = \frac{R - 1}{R + 1} \quad , \quad \Delta\Phi = \frac{1}{1 + R}$$

(d)
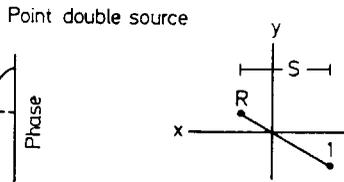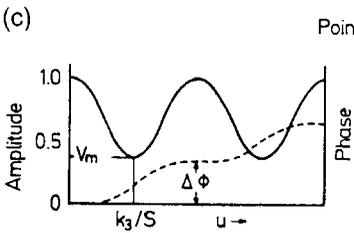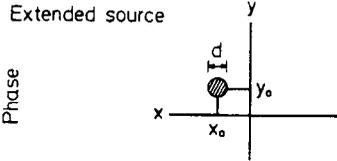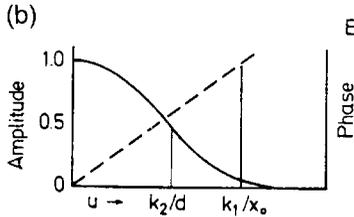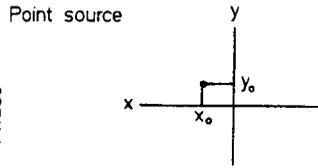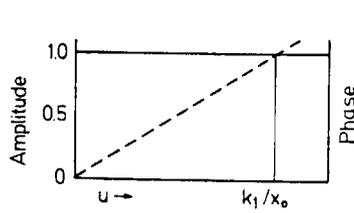
Fig. 9.4 The visibility function for various brightness distribution models. The solid lines are amplitudes, the dashed lines are phases. (**a**) A point source displaced from the phase center; a displacement of $x_0 = 1''$ shifts phase by one fringe for a $k_1 = 206265$ wavelength baseline. (**b**) A displaced Gaussian shaped extended source of FWHP $1''$; the amplitude reaches a value of 0.5 at $k_2 = 91000$ wavelengths. (**c**) Two point sources with an intensity ratio R; the period of amplitude and phase depends on the separation. If the centroid of the double is the phase center, the sign of phase gives the direction of the more intense components, with positive to the east. (**d**) Two extended double sources; this has been obtained from the response to a pair of point sources by multiplying the visibility amplitude by the envelope shown in (**b**). The numerical values are $k_3 = 103000$ if $s = 1''$ and $k_2 = 91000$ if $d = 1''$ [after Fomalont and Wright (1974)]
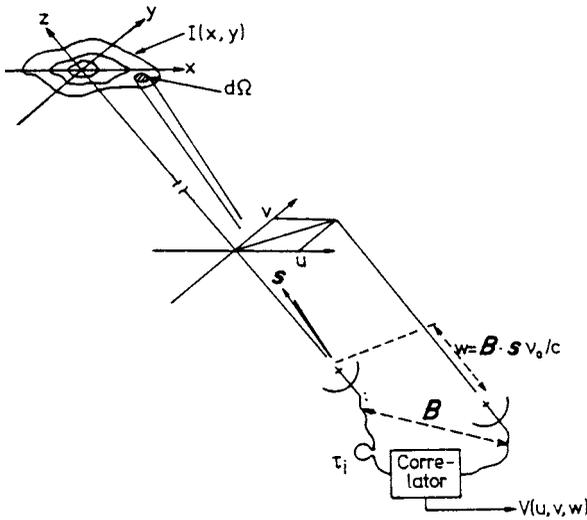
Fig. 9.5 Geometry and coordinates for a detailed discussion of interferometry, leading to aperture synthesis. Here the coordinates u and v are shown [after Thompson et al. (1982)]
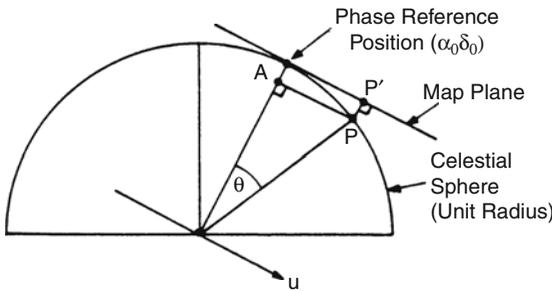
Fig. 9.6 Mapping of points on the celestial sphere onto a plane surface [after Thompson et al. (1986)]

In solving (9.17) the main question is: How many values of $V(u, v, 0)$ are sufficient to produce a good image? We will attempt to answer this question after a brief historical survey.

9.3.2 Historical Development

The theory of producing images from interferometry was published in the late 1940s. The development of practical synthesis radio interferometers began at the Mullard Radio Astronomy Observatory at Cambridge University UK. This culminated in the 5 km Ryle Telescope (RT). The RT consists of an east-west (EW) configuration of eight 13-m antennas, 4 of which are movable.

Classical designs using east-west baselines were continued with the Westerbork Synthesis Radio Telescope (WSRT) which came into operation in the late 1960s (see Fig. 9.7). WSRT consists of fourteen 25-m antennas on an EW line; ten of these are fixed. WSRT can be used for both spectral line and continuum measurements. The spacings of the fixed antennas are equal so that one has many measurements of the same Fourier component. This redundancy increases the accuracy of the (u, v) components that are measured, but limits the number of independent (u, v) samples.

The Fourier transform of such a data set is again a system of elliptical rings with semi-axes $kc/\nu\Delta L$ and $kc/\nu\Delta L \sin\delta_0$ where ΔL is the interval in the baseline and k is an integer. These are referred to as *grating lobes*. If the interferometer consists of telescopes of diameter D, spaced by some multiple of D, the presence of such grating rings usually cannot be avoided. If the array contains antennas at equidistant positions such as the Westerbork array, identical functions V_{ik} can be measured simultaneously by many antenna pairs, and this can then be used by the so-called *redundant* calibration method (Hamaker et al. 1977). However the (u, v) plane is not as efficiently filled. At the Jodrell Bank observatory, the instrumentational development was carried out in a somewhat different way. A major goal was to achieve very
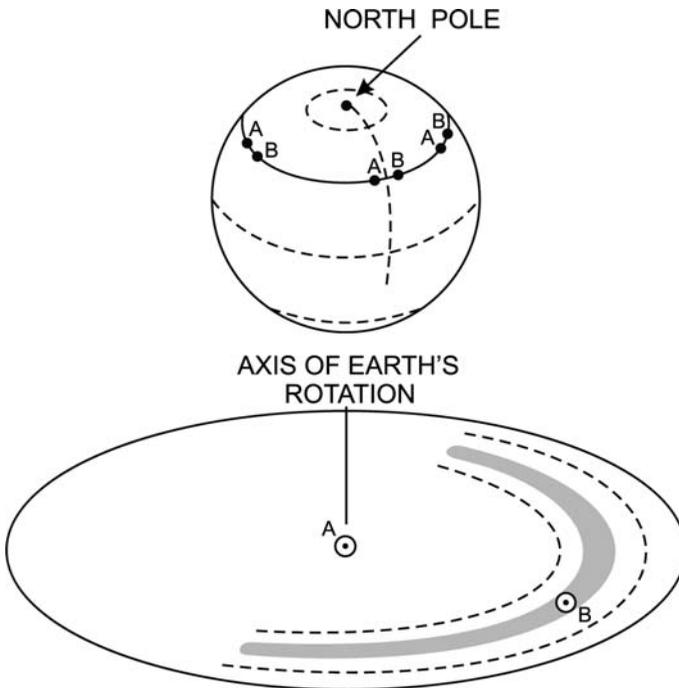
Fig. 9.7 A sketch of the locations in the *uv* plane which are filled with correlated data from the outputs of pairs of antennas located along an east-west (EW) baseline. The *uv* data are collected while tracking a source from rising to setting. The data form a system of elliptical rings with the small axis in the *v* direction. The Fourier transform of such a system produces an elliptical beam with the Declination elongated by $1/\sin\delta_0$, where δ_0 is the source Declination

high angular resolution; this was carried out by transmitting the information from each antenna by means of radio links. With this method, sub-arcsecond angular resolutions were obtained. At the CSIRO Radiophysics Laboratory the developments led to the Mills Cross and grating interferometers; these produced images using analog methods. At the Caltech Owens Valley Observatory the major efforts were devoted to model fitting using results obtained with a two element system.

Starting with the Very Large Array (VLA), technical advances allowed one to use *minimum redundancy arrays* which have geometries that are spiral-like distributions. These produce (u, v) pairs faster and more efficiently, so that, for a given number of antennas, many more (u, v) components could be measured simultaneously. The most recent design using this concept is ALMA (see Fig. 9.10). Even though modern arrays have many antennas and make use of minimum redundancy concepts, there are still gaps in the (u, v) plane. These gaps produce beam structures, or *point spread functions* that are in many cases inferior to those produced by filled apertures.

Both WSRT and VLA are designed to measure spectral lines as well as continuum radiation. As with single telescope observations, spectral line interferometry places some additional requirements on the receiver systems. For example, the frequency stability of the systems must be higher, of the order of 0.1 of a linewidth. However, the delay setting error, Δt_i, can be larger, since Δf, the resolution needed to resolve the lines, is usually much smaller than the bandwidth used for continuum observations. In the early days of spectral line interferometry, one carried out the measurements by having a bank of contiguous phase matched filters at each telescope. With large systems such as WSRT and VLA, this is not practical and one uses instead a cross correlation system (see Sect. 5.4.3.2). The data rates for spectral line measurements are factors of $\sim$100–1000 larger than those from continuum measurements. In the last few decades, the number of interferometer arrays has increased. The largest are MERLIN, operated by Manchester University, the Australia Telescope, operated by CSIRO Radiophysics Division and the GMRT, Pune, India, operated by a part of the Tata Institute. The most recent developments include the merging of the Caltech Owens Valley and Berkley-Illinois-Maryland-Association arrays into CARMA, which consists of six 10.4 m and nine 6.1 m antennas (see Fig. 9.8). CARMA operates to a wavelength of 0.8 mm, with a total geometric collecting area of 772 m^2. The Harvard-Smithsonian/ASIAA SMA consists of eight 6 m sub-mm antennas (Fig. 9.9).

For the following discussions, important definitions are:

(1) *Dynamic Range*: The ratio of the maximum to the minimum intensity in an image. In images made with an interferometer array, it should be assumed that the corrections for primary beam taper have been applied. If the minimum intensity is determined by the random noise in an image, the dynamic range is defined by the signal to noise ratio of the maximum feature in the image. The dynamic range is an indication of the ability to recognize low intensity features in the presence of intense features. If the minimum noise is determined by artefacts, i.e. noise in excess of the theoretical noise, the image can be improved by ′image improvement techniques′.

Fig. 9.8 The Combined Array for Research in Millimeter-wave Astronomy (CARMA) is located in northern California, at an elevation of 2.2 km. CARMA is the merger of the Owens Valley Radio Observatory (OVRO) millimeter array (consisting of six 10-m dishes) and the Berkeley-Illnois-Maryland Association (BIMA) array (consisting of nine 6-m dishes). Funding for CARMA is provided by the U. S. National Science Foundation and the consortium universities (photo courtesy of J. Miller and D. Bock)

Fig. 9.9 The Sub Millimeter Array (SMA) consists of eight 6-m antennas that are designed to operate to 0.3 mm, with a total geometric collecting area is $226\,\mathrm{m}^2$. The SMA is located on Mauna Kea, at an altitude of 4.1 km. It was built and is operated by the Harvard-Smithsonian Center for Astrophysics (CfA) and the Academia Sinica Institute for Astronomy & Astrophysics (ASIAA), Taiwan (photo courtesy of N. Patel, CfA.)

(2) *Image Fidelity*: This is defined by the agreement between the measured results and the actual (so-called "true") source structure. A quantitative comparison would be

$$F = |(S-R)|/R$$

where F is the fidelity, R is the resulting image obtained from the measurement, and S is the actual source structure. Of course one cannot have a priori
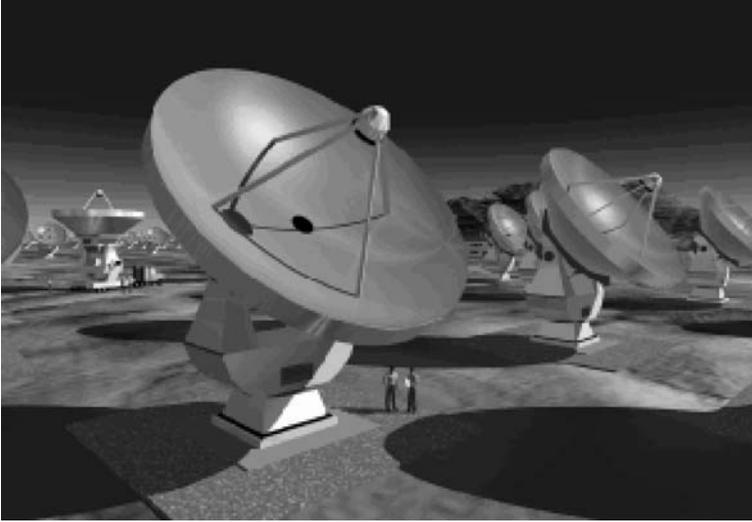
Fig. 9.10 The most ambitious construction project in radio astronomy is the Atacama Large Millimeter Array (ALMA), a joint project of North America, Europe and East Asia. ALMA will be built in north Chile on a 5 km high site. It will consist of fifty-four 12-m and twelve 7-m antennas, operating in 10 bands between wavelength 1 cm and 0.3 mm. The collecting area of ALMA is 55% that of the VLA, but is vastly more ambitious because of the more complex receivers, the need for highly accurate antennas, and the high altitude site. In addition, the data rates will be orders of magnitude higher than any presently operating interferometer. At the longest antenna spacing, and shortest wavelength, the angular resolution will be ≈5 milli arcseconds (sketch courtesy European Southern Observatory)

knowledge of the correct source structure. In the case of simulations, S is a source model, R is the result of processing S through R.

9.3.3 Interferometric Observations

Usually measurements are carried out in 1 of 3 ways.

- In the first procedure, measurements of the source of interest and a calibrator are made. This is as in the case of single telescope position switching. One significant difference with single dish measurements is that the interferometer measurement extends over a wide range of hour angles to provide a better coverage of the *uv* plane, if the baseline is EW (see 9.7). One first measures a calibration source or reference source, which has a known position and size, to remove the effect of instrumental phases in the instrument and atmosphere and to calibrate the amplitudes of the sources in question. Sources and calibrators are usually observed alternately. The time variations caused by instrumental and weather effects must be slower than the time between measurements of source and calibrator. If, as is the case for millimeter and sub-mm wavelength measurements,

weather is an important influence, one must switch frequently between on and off. In *fast switching* one might spend 10 s on a nearby calibrator, then a few minutes on-source. This method will reduce the amount of phase fluctuations, but also the amount time available for source measurements. For more rapid changes in the earth's atmosphere, one can correct the phase using measurements of atmospheric water vapor, or changes in the system noise temperature of the individual receivers caused by atmospheric effects. The corrections for instrumental amplitudes and phases are assumed to be constant over the times when the source is observed. The ratio of amplitudes of source and calibrator are taken to be the normalized source amplitudes. Since the calibrators have known flux densities and positions, the flux densities and positions of the sources can be determined. The reference source should be as close to the on-source as possible, but must have a large enough intensity to guarantee a good signal-to-noise ratio after a short time. Frequently nearby calibrators are time variable over months, so a more distant calibrator with a known or fixed flux density is measured at the beginning or end of the session. This source is usually rather intense, so may also serve as a bandpass calibrator for spectral line measurements. The length of time spent on the off-source measurement is usually no more than few minutes.

- In the next procedure, the so-called *snapshots*, one makes a series of short observations (at different hour angles) of a large number of sources. For sensitivity reasons these are usually made in the radio continuum or intense maser lines. As in the first observing method, one intersperses measurements of a calibration source which has a known position and size to remove the effect of instrumental phases in the instrument and atmosphere and to calibrate the amplitudes of the sources in question. The images are affected by the shape of the synthesized beam of the interferometer system. If the size of the source to be imaged is comparable to the primary beam of the individual telescopes, the power pattern of the primary beams will have a large effect. This effect can be corrected easily.

- In the third procedure, one aims to produce a high-resolution image of a source where the goal is either high dynamic range or high sensitivity. The *dynamic range* is the ratio of the highest to the lowest brightness level of reliable detail in the image. This may depend on the signal-to-noise ratio for the data, but for centimeter aperture synthesis observations, spurious features in the image caused by the incomplete sampling of the (u, v) plane are usually more important than the noise. Frequently one measures the source in a number of different interferometer configurations to better fill the uv plane. These measurements are taken at different times and after calibration, the visibilities are entered into a common data set.

  In order to eliminate the loss of source flux density due to missing short spacings, one could supplement the interferometer data with single dish measurements. The diameter of the single dish telescope should be larger than the shortest spacing between interferometer dishes. This single dish image must extend to the FWHP of the smallest of the interferometer antennas. When Fourier transformed and appropriately combined with the interferometer response, this data set has

no missing flux density. Such "missing spacings" are frequently a problem with interferometer images. Usually, interferometer images have shortcomings. Improvements to such images will be surveyed in the next sections.

An extension of this procedure may involve the measurement of adjacent regions of the sky. This is *mosaicing*. In a mosaic, the primary beams of the telescopes should overlap, ideally this would be at the half power point. In the simplest case, the images are formed separately and then combined to produce an image of the larger region.

9.3.4 Improving Visibility Functions

Ideally the relation between the measured $\widetilde{V_{ik}}$ and *actual* V_{ik} functions can be considered as linear:

$$\widetilde{V_{ik}}(t) = g_i(t)\, g_k^*(t)\, V_{ik} + \varepsilon_{ik}(t)\ . \tag{9.18}$$

Average values for the antenna gain factors g_k and the noise term $\varepsilon_{ik}(t)$ are determined by measuring calibration sources as frequently as possible. Actual values for g_k are then computed by linear interpolation. These methods make full use of the fact that the (complex) gain of the array is obtained by multiplication the gains of the individual antennas. If the array consists of n such antennas, $n(n-1)/2$ visibilities can be measured simultaneously, but only $(n-1)$ independent gains g_k are needed (for one antenna, one can arbitrarily set $g_k = 1$ as a reference). In an array with many antennas, the number of antenna pairs greatly exceeds the number of antennas. For phase, one must determine n phases.

Often these conditions can be introduced into the solution in the form of *closure errors*. Introducing the phases φ, θ and ψ by

$$\begin{aligned}
\widetilde{V_{ik}} &= |\widetilde{V_{ik}}| \exp\{\mathrm{i}\,\varphi_{ik}\}\,, \\
G_{ik} &= |g_i|\,|g_k| \exp\{\mathrm{i}\,\theta_i\} \exp\{-\mathrm{i}\,\theta_k\}\,, \\
V_{ik} &= |V_{ik}| \exp\{\mathrm{i}\,\psi_{ik}\}\,.
\end{aligned} \tag{9.19}$$

From (9.18) the phase ψ_{ik} on the baseline ik will be related to the observed phase φ_{ik} by

$$\varphi_{ik} = \psi_{ik} + \theta_i - \theta_k + \varepsilon_{ik}\,, \tag{9.20}$$

where ε_{ik} is the phase noise. Then the *closure phase* Ψ_{ikl} around a closed triangle of baseline ik, kl, li,

$$\Psi_{ikl} = \varphi_{ik} + \varphi_{kl} + \varphi_{li} = \psi_{ik} + \psi_{kl} + \psi_{li} + \varepsilon_{ik} + \varepsilon_{kl} + \varepsilon_{li}\,, \tag{9.21}$$

will be independent of the phase shifts θ introduced by the individual antennas and the time variations. With this procedure, on can minimize phase errors.

Closure amplitudes can also be formed. If four or more antennas are used simultaneously, then ratios, the so-called *closure amplitudes*, can be formed. These are independent of the antenna gain factors:

$$A_{klmn} = \frac{|V_{kl}||V_{mn}|}{|V_{km}||V_{ln}|} = \frac{|\Gamma_{kl}||\Gamma_{mn}|}{|\Gamma_{km}||\Gamma_{ln}|} \; . \tag{9.22}$$

Both phase and closure amplitudes can be used to improve the quality of this complex function.

If each antenna introduces an unknown complex gain factor g with amplitude and phase, the total number of unknown factors in the array can be reduced significantly by measuring closure phases and amplitudes. If four antennas are available, 50% of the phase information and 33% of the amplitude information can thus be recovered; in a 10 antenna configuration, these ratios are 80% and 78% respectively. A more extensive discussion is to be found in the review articles by Pearson and Readhead (1984) and Cornwell and Fomalont (1989) where references are given.

9.3.5   Multi-Antenna Array Calibrations

For two antenna interferometers, phase calibration can only be made pair-wise. This is referred to as "baseline based" solutions for the calibration. For a multi-antenna system, there are other and better methods. One can use sets of three antennas to determine the best phase solutions and then combine these to optimize the solution for each antenna. For amplitudes, one can combine sets of four antennas to determine the best amplitude solutions and then optimize this solution to determine the best solution. This process leads to 'antenna based' solutions are used. Antenna based calibrations are used in most cases. These are determined by applying phase and amplitude closure for subsets of antennas and then making the best fit for a given antenna.

9.3.6   Data Processing

9.3.6.1   Gridding *uv* Data

Before we consider specialized techniques, we must arrange the data in a mathematically and computationally useful way. The computation of (9.28) even for a modest data set is quite time consuming. Therefore methods for inverting (9.15) that are based on the Cooley-Tukey fast Fourier transform algorithm (FFT) are generally used. In order to use the FFT in its simplest version, the visibility function must be placed on a regular grid with total sizes that are powers of two of the sampling interval. Since the observed data seldom lie on such regular grids, an interpolation scheme must be used. If the measured points are randomly distributed, this interpolation is best carried out using a convolution procedure.

The *gridded visibility function* may be represented by

$$V''(u,v) = \mathrm{III}\,(u,v)\left\{G(u,v) \otimes V'(u,v)\right\}, \tag{9.23}$$

where $V'(u,v)$ is the measured function sampled on the irregular u_i, v_i grid, $G(u,v)$ is a convolving function, and $\otimes$ is the convolution operator by which a value for an interpolated function $V''(u,v)$ is defined for every u, v. Finally, the Sha function

$$\text{III}(u,v) = \Delta u \Delta v \sum_{j,k=-\infty}^{\infty} \delta(u - j\Delta u)\, \delta(v - k\Delta v) \qquad (9.24)$$

defines the regular grid on the (uv) plane with Δu and Δv being the cell sizes.

The intensity distribution $I'(x,y)$ is now obtained by substituting (9.23) into (9.17). The Fourier transform (9.17) can be computed by the FFT because $V''(u,v)$ is now arranged on a regular grid. However, (9.17) with (9.23) have certain drawbacks that are most easily seen by rewriting these equations using some of the fundamental properties of Fourier transforms. Recalling that the Fourier transform of the Sha function is another Sha function with a grid spacing $1/\Delta u$ and $1/\Delta v$

$$\text{III}(x,y) = \sum_{i,j=-\infty}^{\infty} \delta(x - i/\Delta u)\, \delta(y - i/\Delta v) \qquad , \qquad (9.25)$$

and applying the convolution theorem to (9.23), we obtain

$$I(x,y) = \text{III}(x,y) \otimes \left[g(x,y) I'(x,y) \right] \qquad , \qquad (9.26)$$

where $g(x,y)$ is the Fourier transform of $G(u,v)$, that is, the grading that gives rise to the beam $G(u,v)$. The important property of (9.26) is the fact that gI' is convolved with a Sha function that extends over the full x, y space. If the grading $g(x,y)$ does not remain equal to zero outside the image area, radiation from outside this region may be aliased into the image. This will happen for most practical convolving functions. Adopting a *pill box beam*,

$$G(u,v) = \begin{cases} 1 \text{ for } u^2 + v^2 \leqq u_{\text{max}}^2 \\ 0 \text{ otherwise} \end{cases} \qquad , \qquad (9.27)$$

the corresponding grading will be a $\sin x / x$ function which has nonzero values extending over the full xy plane. Other convolving functions will produce slightly less aliasing, but this effect can never be completely avoided. By a proper choice of the convolution function, aliasing can be suppressed by factors of 10^2 to 10^3 at two to three image radii, and this usually is sufficient. Image structures caused by aliasing can, however, be recognized by regridding the function $V(u,v)$ to another grid size Δu, Δv, since the aliased data will then be shifted in position in the resulting image.

9.3.6.2 Principal Solution, Dirty Map and Dirty Beam

If some of the spatial frequencies present in the intensity distribution are not present in the (u,v) plane data, then changing the amplitude or phase of the corresponding visibilities will not have any effect on the reconstructed intensity distribution – these have been filtered out by the dirty beam.

Expressed mathematically, if Z is an intensity distribution containing only the unmeasured spatial frequencies, and P_{D} is the dirty beam, then

$$P_{\mathrm{D}} \otimes Z = 0.$$

Hence, if I is a solution of the convolution equation (9.29) then so is $I + \alpha Z$ where α is any number. This obviously shows that there is no unique solution to the problem.

The solution with visibilities $V = 0$ for the unsampled spatial frequencies is usually called the *principal solution*, and it differs from the true intensity distribution by some unknown *invisible* or *ghost distribution*. It is the aim of image reconstruction to obtain reasonable approximations to these *ghosts* by using additional knowledge or plausible extrapolations, but there is no direct way to select the "best" or "correct" image from all possible images. The familiar linear deconvolution algorithms are not adequate and nonlinear techniques must be used. For quick look or snapshot images, hybrid techniques are used that improve the appearance of the images by employing heuristic recipes. These methods often form the first steps in a more sophisticated image restoration procedure.

The result obtained from the gridded uv data can be Fourier transformed to obtain an image with a resolution corresponding to the size of the array. However, this may still contain artifacts caused by the details of the observing procedure, especially restricted coverage of the (uv) plane. Therefore the dynamic range of such so-called *dirty* maps is rather small. This can be improved by further data analysis, as will be described next.

If the calibrated function $V(u,v)$ is known for the full (u,v) plane both in amplitude and in phase, this can be used to determine the (modified) intensity distribution $I'(x,y)$ by performing the Fourier transformation (9.17). However, in a realistic situation $V(u,v)$ is only sampled at discrete points within a radius $\cong u_{\max}$ along elliptical tracks, and in some regions of the (u,v) plane, $V(u,v)$ is not measured due to missing short spacings, antenna shadowing or perhaps a missing angular wedge in the orientation of the baseline vector $\boldsymbol{B}$. We show a typical (u,v) plane distributions for a low Declination source as measured by the VLA in Fig. 9.11.

In direct analogy with single telescope illumination patterns, we can weight the visibilities by a grading function, g. Then for a discrete number of visibilities, we have a version of (9.17) involving a summation, not an integral, to obtain an image via a discrete Fourier transform (DFT):

$$I_{\mathrm{D}}(x,y) = \sum_{k} g(u_k, v_k) V(u_k, v_k) \, \mathrm{e}^{-\mathrm{i}2\pi(u_k x + v_k y)}, \qquad (9.28)$$
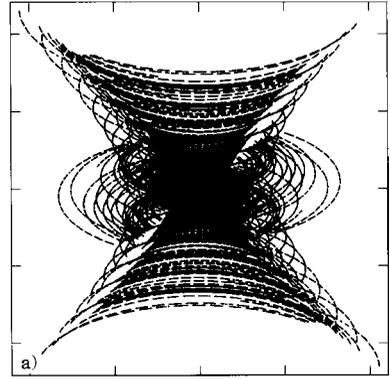
where $g(u,v)$ is a weighting function called the grading or apodisation. To a large
extent $g(u,v)$ is arbitrary and can be used to change the effective beam shape and
side lobe level. There are two widely used weighting functions: uniform and natu-
ral. Natural weighting uses $g(u_k, g_k) = 1/N_s(k)$, where $N_s(k)$ is the number of data
points within a symmetric region of the (u,v) plane. In contrast, uniform weighting
uses $g(u_k, v_k) = 1$. In a simple case $N_s(k)$ would be a square centered on point k.
Data which are naturally weighted result in lower angular resolution but give a bet-
ter signal-to-noise ratio than uniform weighting. But these are only extreme cases.
One can choose intermediate weighting schemes. These are often referred to as *ro-
bust* weighting (in the nomenclature of the AIPS data reduction package). Often the
reconstructed image I_D may not be a particularly good representation of I', but these
are related. In another form, (9.28) is

$$I_D(x,y) = P_D(x,y) \otimes I'(x,y),   \tag{9.29}$$

where

$$P_D = \sum_k g(u_k, v_k) e^{-i2\pi(u_k x + v_k y)}   \tag{9.30}$$

is the response to a point source. This is the *point spread function* PSF for the dirty
beam. Thus the *dirty beam* can be understood as a transfer function that distorts
the image. (The dirty beam, P_D, is produced by the Fourier transform of a point
source in the regions sampled; this is the response of the interferometer system to a
point source). That is, the *dirty map* $I_D(x,y)$ contains only those spatial frequencies
(u_k, v_k) where the function has been measured. The sum in (9.30) extends over the
same positions (u_k, v_k) as in (9.28), and the side lobe structure of the beam depends
on the distribution of these points.

In Fig. 9.12 we show the (u,v) plane distribution for an extremely dense sampling
of a low-declination source. In Fig. 9.12c, the resulting point spread function using
all the independent data points is shown, and in Fig. 9.12b the PSF resulting from
uniform weighting is given. Clearly, the uniformly weighted PSF will emphasize
larger spacings in the (u,v) plane. Since less than all of the information is used, the
signal-to-noise ratio will be lower.

Fig. 9.12 A block diagram
showing the steps needed to
convert aperture synthesis
measurements into images.
Each of the steps is described
in Sect. 9.4. For schemes in-
volving self-calibration, the
final image from the proce-
dure above can be used as
the input model. Usually one
CLEANs the image produced
and then self-calibrates the re-
sulting image. In modern data
reduction, considerable effort
is expended in self-calibration
and image restoration sections
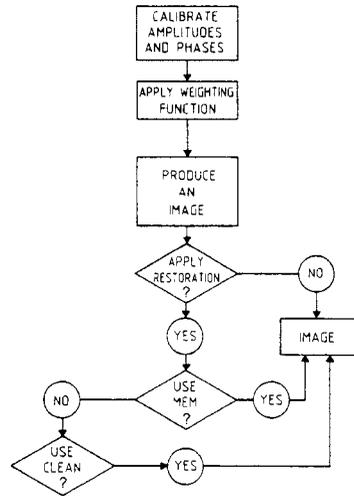
9.4   Advanced Image Improvement Methods

The image produced from visibilities measured by an interferometer is first formed
in the digital computer. Digital computing therefore is clearly an integral part of
synthesis array data processing, and a large part of the advances in radio synthesis
imaging during the last 15–20 years relies on the progress made in the field of image
restoration. A schematic of this improvement process is shown in Fig. 9.12.

9.4.1   Self-Calibration

Amplitude and phase errors scatter power across the image, giving the appearance
of enhanced noise. Quite often this problem can be alleviated to an impressive ex-
tent by the method of *self-calibration*. This process can be applied if there is a
sufficiently intense source in the field contained within the primary beam of the in-
terferometer system. Basically, self-calibration is the equivalent of focusing on the
source, analogous to using the focus of a camera to sharpen up an object in the field
of view. One can restrict the self-calibration to an improvement of phase alone or to
both phase and amplitude. However, self-calibration is carried in the (u, v) plane. If
properly used, this method leads to a great improvement in interferometer images
of compact intense sources such as those of masering spectral lines. If this method
is used on objects with low signal-to-noise ratios, this method may give very wrong
results by concentrating random noise into one part of the interferometer image.

In measurements of weak spectral lines, the self-calibration is carried out with
a continuum source in the field. The corrections are then applied to the spectral
line data. In the case of intense lines, one of the frequency channels containing the
emission is used. If self-calibration is applied, the source position information is
usually lost.

9.4.2 Applying CLEAN to the Dirty Map

CLEANing is the most commonly used technique to improve single radio interferometer images. The *dirty map* is a representation of the principal solution, but with shortcomings. In addition to its inherent low dynamic range, the dirty map often contains features such as negative intensity artifacts. These cannot be real. Another unsatisfactory aspect is that the principal solution is quite often rather unstable, in that it can change drastically when more visibility data are added. Instead of a principle solution that assumes $V = 0$ for all unmeasured visibilities, values for V should be adopted at these positions in the (u, v) plane. These are obtained from some plausible model for the intensity distribution.

The CLEAN method approximates the actual but unknown intensity distribution $I(x, y)$ by the superposition of a finite number of point sources with positive intensity A_i placed at positions (x_i, y_i). It is the aim of CLEAN to determine the $A_i(x_i, y_i)$ such that

$$I''(x, y) = \sum_i A_i P_D(x - x_i, y - y_i) + I_\varepsilon(x, y) \qquad (9.31)$$

where I'' is the dirty map obtained from the inversion of the visibility function and P_D is the dirty beam (9.30). $I_\varepsilon(x, y)$ is the residual brightness distribution after decomposition. Approximation (9.31) is deemed successful if I_ε is of the order of the noise in the measured intensities. This decomposition cannot be done analytically, rather an iterative technique has to be used.

The concept of CLEAN was first devised by J. Högbom (1974). The algorithm is most commonly applied in the image plane. This is an iterative method which functions in the following fashion: First find the peak intensity of the dirty image, then subtract a fraction γ with the shape of the dirty beam from the image. Then repeat this n times. This *loop gain* $0 < \gamma < 1$ helps the iteration converge, and it is usually continued until the intensities of the remaining peaks are below some limit. Usually the resulting point source model is convolved with a *clean beam*, usually of Gaussian shape with a FWHP similar to that of the dirty beam. Whether this algorithm produces a realistic image, and how trustworthy the resulting point source model really is, are unanswered questions, although Schwarz (1978) has been able to prove its validity under some conditions.

There are several modifications of the CLEAN algorithm that can reduce the computing time under certain conditions. Clark (1980) shifted part of the computation from the image plane to the (u, v) plane. Cotton and Schwab (see Schwab 1984) also shifted some of these computations to the visibility data. CLEAN is used widely in radio interferometry, even if the problems of non uniqueness and instabilities of solution are not solved.

9.4.3 Maximum Entropy Deconvolution Method (MEM)

The Maximum Entropy Deconvolution Method (MEM) is commonly used to produce a single optimal image from a set of separate but contiguous images. The problem of how to select the "best" image from many possible images which all

agree with the measured visibilities is solved by MEM. Using MEM, those values of the interpolated visibilities are selected, so that the resulting image is consistent with all previous relevant data. In addition, the MEM image has maximum smoothness. This is obtained by maximizing the *entropy* of the image. One possible definition of entropy is given by

$$\mathscr{H} = -\sum_i I_i \left[\ln\left(\frac{I_i}{M_i}\right) - 1 \right], \tag{9.32}$$

where I_i is the deconvolved intensity and M_i is a reference image incorporating all "a priori" knowledge. In the simplest case M_i is the empty field $M_i = \text{const} > 0$, or perhaps a lower angular resolution image.

Additional constraints might require that all measured visibilities should be reproduced exactly, but in the presence of noise such constraints are often incompatible with $I_i > 0$ everywhere. Therefore the MEM image is usually constrained to fit the data such that

$$\chi^2 = \sum \frac{|V_i - V_i'|^2}{\sigma_i^2} \tag{9.33}$$

has the expected value, where V_i is the measurement while V_i' corresponds to the MEM image and σ_i is the error of the measurement.

Algorithms for solving this maximization problem have been given by various authors (Wernecke and d'Addario 1976, Gull and Daniell 1978, Cornwell and Evans 1985). Programs for MEM are available in different astronomical data reduction packages. Both CLEAN and MEM are used to produce radio aperture synthesis images, although CLEAN is used perhaps ~ 100 times more frequently. The use of MEM has been widely adopted in the fields of image processing since MEM is better suited to handling images of sources which are larger than a telescope primary beam, that is, multi-field images. Point sources are best handled by CLEAN, while MEM is better for extended sources of low surface brightness. Computationally CLEAN is usually faster than MEM provided the image is not too big. The break-even point is around one million pixels.

9.5 Interferometer Sensitivity

The random noise limit to an interferometer system is calculated following the method used for a single telescope. The RMS fluctuations in antenna temperature are

$$\Delta T_A = \frac{M\,T_{\text{sys}}}{\sqrt{t\,\Delta v}}, \tag{9.34}$$

where M is a factor of order unity used to account for extra noise from analog to digital conversions, digital clipping etc. If we next apply the definition of flux density, S_v in terms of antenna temperature for a two-element system, we find:

$$\Delta S_v = 2\,k\,\frac{T_{\text{sys}}\,e^\tau}{A_e\,\sqrt{2t\,\Delta v}}, \tag{9.35}$$

where τ is the atmospheric optical depth and A_e is the effective collecting area of a single telescope of diameter D. There is additional in this expression since a multiplying interferometer does not process all of the information (i.e. the total power) that the antennas receive. In this case, there is an additional factor of $\sqrt{2}$ compared to the noise in a single dish with an equivalent collecting area since there is information not collected by a multiplying interferometer. We denote the system noise corrected for atmospheric absorption by $T'_{\mathrm{sys}} = T_{\mathrm{sys}} \exp \tau$, in order to simplify the following equations. For an array of n identical telescopes, there are $N = n(n-1)/2$ simultaneous pair-wise correlations. Then the RMS variation in flux density is

$$\Delta S_\nu = \frac{2 M k T'_{\mathrm{sys}}}{A_e \sqrt{2 N t \Delta \nu}} . \qquad (9.36)$$

This relation can be recast in the form of brightness temperature fluctuations using the Rayleigh-Jeans relation;

$$S = 2 k \frac{T_b \Omega_b}{\lambda^2} . \qquad (9.37)$$

Then the RMS brightness temperature, due to random noise, in aperture synthesis images is

$$\Delta T_b = \frac{2 M k \lambda^2 T'_{\mathrm{sys}}}{A_e \Omega_b \sqrt{2 N t \Delta \nu}} . \qquad (9.38)$$

For a Gaussian beam, $\Omega_{\mathrm{mb}} = 1.133 \, \theta^2$ (see (8.13)), we can then relate the RMS temperature fluctuations to observed properties of a synthesis image.

A few qualitative comments concerning (9.38) are in order. First, with shorter wavelengths, the RMS temperature fluctuations are predicted to be lower. Thus, for the same collecting area and system noise, if weather changes are unimportant, a millimeter image should be more sensitive than an image made at centimeter wavelengths. Second, if the effective collecting area remains the same and for a larger main beam solid angle, temperature fluctuations will decrease. For this reason, smoothing an image will result in a lower RMS noise in an image. However, if smoothing is too extreme, this process effectively leads to a decrease in collecting area; then there will be no further improvement in sensitivity. Finally, it is frequently noted that multi-element interferometers are capable of producing images faster than single dishes. This is due to the fact that there are n receivers in an interferometer system. If a single dish of equivalent collecting area and angular resolution also had n receivers, it would produce an image of the same area in the same amount of time.

Inserting numerical values in (9.36) and (9.38), we have

$$\Delta S_\nu = 1.02 \, \frac{M T'_{\mathrm{sys}}}{A_e \sqrt{N t \Delta \nu}} ,$$

$$\Delta T_b = 13.58 \, \frac{M \lambda^2 T'_{\mathrm{sys}}}{A_e \theta^2 \sqrt{N t \Delta \nu}} ,$$

where λ is expressed in mm, θ in arc sec, and $\Delta\nu$ in kHz. As an example, one can compare the performance of the VLA at 1.3 cm with the bilateral ALMA (i.e. fifty 12 m antennas) at 2.7 mm. We set $M = 1$ for the VLA. For the twenty-seven 25 m diameter antennas of the VLA at 1.3 cm, each with $A_e = 300\,m^2$, after a four hour integration on source with a 100 K system noise and 7.8 kHz $(= 0.1\,km\,s^{-1})$ frequency resolution, in a $3''$ synthesized beam, the flux density RMS sensitivity is 3.2 mJy and the brightness temperature RMS sensitivity is 0.8 K. For the fifty 12 m diameter antennas of ALMA, each with $A_e = 90\,m^2$ after a similar integration, with a 128 K system noise temperature at 2.6 mm and a 38 kHz frequency resolution $(= 0.1$, the RMS spectral line flux density sensitivity is 2.8 mJy, while the corresponding RMS temperature sensitivity is 0.03 K. For continuum measurements, the flux density sensitivities can be increased by using broader bandwidths. The bandwidth of ALMA is 8 GHz. This gives an RMS noise of 6.3 μJy. With the VLA one can have 50 MHz bandwidth, which gives an RMS noise of 40 μJy. The sensitivity may be limited by confusion. This was discussed in Sect. 8.5; if there are less than five beams per source, the angular resolution must be increased. For a $0.3''$ synthesized beam, the RMS temperature sensitivities are a factor of 100 larger.

The temperature sensitivity (in Kelvins) for higher angular resolution is worse than for a single telescope with an equal collecting area. From the Rayleigh-Jeans relation, the sensitivity in Jansky (Jy) is fixed by the antenna collecting area and the receiver noise, so only the wavelength and the angular resolution can be varied. Thus, the increase in angular resolution is made at the expense of temperature sensitivity. All other effects being equal, at shorter wavelengths one gains in temperature sensitivity (cf. The VLA versus ALMA in the previous example). This is not such a great problem for the high-brightness radio sources, which radiate by nonthermal processes, such as synchrotron radiation, but would be for thermal sources, for which the maximum brightness temperature is about 2×10^4 K for regions of ionized gas surrounding massive stars. Our sun has a corona with a temperature of 10^6 K. However, even this temperature is small compared to brightness temperatures of up to 10^{12} K, found for nonthermal sources.

Compared to single dishes, interferometers have the great advantage that uncertainties such as pointing and beam size depend fundamentally on timing. Such timing uncertainties can be made very small compared to all other uncertainties. In contrast, the single dish measurements are critically dependent on mechanical deformations of the telescope. In summary, the single dish results are easier to obtain, but source positions and sizes on arc second scales are difficult to estimate. The interferometer system has a much greater degree of complexity, but allows one to measure such fine details. The single dish system responds to the source irrespective of the relation of source to beam size; the correlation interferometer will not record source structures larger than a few fringes.

Aperture synthesis is based on sampling the function $V(u,v,0)$ with pairs of antennas to provide samples in the (u,v) plane. Many configurations are possible, but the goal is the densest possible coverage of the (u,v) plane. If one calculates the RMS noise in a synthesis image obtained by simply Fourier transforming the (u,v) data, one usually finds a noise level many times higher than that given by (9.38)

or (9.36). There are various reasons for this. One cause is phase fluctuations due to atmospheric or instrumental influences such as LO instabilities. Another cause is due to incomplete sampling of the (u, v) plane. This gives rise to instrumental features, such as stripe-like features in the final images. Yet another systematic effect is the presence of grating rings around more intense sources; these are analogous to *high side lobes* in single dish diffraction patterns. Over the past 20 years, it has been found that these effects can be substantially reduced by software techniques such as CLEAN and Maximum Entropy.

9.6   Very Long Baseline Interferometers

For a given wavelength the angular resolving power of an interferometer depends only on the length of the interferometer baseline $\boldsymbol{B}$. But the need to provide phase-stable links (optical fibers) between individual antennas and the correlator sets limits on $|\boldsymbol{B}|$ to $\sim$200 km as is the case for e-MERLIN. Over longer paths, it becomes more difficult to guarantee the phase stability, since transient irregularities in the transmission path will have detrimental effects, so systems such as MERLIN are limited to baselines of a few hundred kilometers.

The development of atomic clocks, together with extremely phase-stable oscillators opened up the possibility of completely avoiding the transmission of a phase-stable local oscillator signal. The measurements are made independently at the individual antennas of the interferometer. The data are recorded on storage media together with precise time marks. These data are correlated later. Currently the data are recorded on hard disks that are shipped to a central correlator location; in the near future these will be sent in real time or near real time over the internet. The antenna outputs contain accurate records of the time variation of the electrical field strength so that the appropriately time-averaged product obtained by multiplying the digitized signal gives the mutual correlation function directly.

Local oscillators with extreme phase stability are needed at each station for two reasons. The media which are normally used usually permit the recording of signals in a band reaching from zero to at most a hundred MHz; the signal therefore must be mixed down to this band, and for this a phase-stable local oscillator is needed, since all phase jumps of the oscillator affect the IF signal. The second use of the phase stability is to provide the extremely precise time marks needed to align the signals from two stations. Again phase jumps would destroy the coherence, that is the correlation of signals from the source. For the local oscillators, different systems have been used with varying success, ranging from rubidium clocks, free-running quartz oscillators and, most successfully, hydrogen masers. With present day hydrogen maser technologies, it is possible to have frequency and phase stability that allows measurements for many minutes. At longer centimeter wavelengths, the maximum time over which the visibility function can be coherently integrated, that is, where we can determine the amplitude and phase of the visibility function is *not* limited by

the best currently commercially available maser clocks. At wavelengths or 1 cm or shorter, the atmosphere is the limit.

Today in VLBI only digital data recording is used. The media from different observatories are processed on special purpose digital correlators that align the signals in time, account for local oscillator offsets and geometric delays, clock rate offsets and differential Doppler shifts arising from Earth rotation and then generate a correlation function for each pair of observation sites. Amplitudes and phases of these correlation functions are directly comparable to the complex visibilities of a conventional connected element interferometer. The delay time between the two independent telescopes can vary rapidly. In the past, one could not determine the instrumental phases from measurements of a calibration source, so one had to use *fringe fitting* to allow the accumulation of data over much longer times. (Fringe fitting is the process by which one determines the phase of the visibility and the rate of change of this phase.) The procedure is similar to that used, for example, at the VLA. The correlator delivers an amplitude as a function of time for a delay range larger than any uncertainty. In the Fourier transform domain the time axis becomes frequency (*residual fringe frequency* or *fringe rate*) and then the maximum should appear as a peak in this two-dimensional distribution. The coordinates of this *peak*, the *fringe rate* and the *lag* are the required parameters. For strong sources this maximum can be determined directly, but for weaker sources more sophisticated techniques are needed. Once these parameters are determined, further reduction procedures are basically identical to those used for the analysis of conventional synthesis array data. To remove the remaining errors, one solves and corrects for residual delays and fringe rate offsets. This is an additional reduction step beyond that used in VLA data reduction. This step, *fringe fitting*, is necessary for VLBI data reductions. There are several reasons for the phase variations:

(1) random delays in the atmospheric propagation properties at the individual sites and
(2) phase changes in the electronics and the independent clocks.

For these reasons, without fringe fitting the correlations will be only fairly short for a given source.

Observationally, VLBI can be used to carry out surveys of a large number of sources. Such surveys make use of procedures that are similar to those used in the *snapshot* observing mode of instruments such as the VLA. Each measurement has a total integration time of up to tens of minutes. For such surveys, the observing runs may extend over 1–2 days so that hundreds of sources are measured.

At present there are a number of instrumental limitations in VLBI. For example, VLBI surveys are biased toward bright sources because of the short integration times and limited bandwidths. Improvements are planned to provide more sensitivity and ease of observing. There are now the first successful efforts to transmit data over the internet. This exchange of data will occur more rapidly, and will facilitate the detection of fringes shortly after a run begins. This will help to eliminate instrumental problems early in an observing session. Improvements in sensitivity depend on collecting area, system noise and increased bandwidth; this in turn will require

increased storage/transmission capacity, faster correlators and computers. For example, the field of view can be expanded by recording the data with a larger number of contiguous channels each with a narrower bandwidth. From the discussion at the end of Sect. 9.2.3.3, with narrower bandwidths one can image sources over the entire field of view of the single dishes. The storage and sensitivity improvements allow for shorter integration times, and thus better compensation for fluctuations in phase. In addition, with software one can shift the phase center to positions far from center of the primary beam, and thus form additional images using the same data set. Usually, one applies self-calibration (Sect. 9.4.1) to increase the sensitivity and dynamic range. Since the computing needed varies as the square of the number of baselines. As an example, in going from 10 to 20 antennas requires 18 times more computing; one can hope that with *Moore's Law* computing will keep pace. It has become possible to use clusters of personal computers, PC's, for the correlation of the VLBI data. Although slower than hardware correlators, the PC software can easily be modified to apply new algorithms. The end result will be VLBI surveys extending to the μJy level over fields of many degrees in size.

9.7 Interferometers in Astrometry and Geodesy

The value of $V(\boldsymbol{B})$ for a brightness distribution $I_\nu(\boldsymbol{s})$ as measured by an interferometer provides a measure of the scalar product $\boldsymbol{B} \cdot \boldsymbol{s}$, and it therefore depends both on the orientation and size of the baseline $\boldsymbol{B}$ and on the source position $\boldsymbol{s}$. In deriving $I_\nu(\boldsymbol{s})$ from this we assumed that both $\boldsymbol{B}$ and $\boldsymbol{s}$ can be specified to an accuracy of a milli-arcsecond (mas). However this is not always a simple matter, and it may be necessary to include the variation of the earth's rotation vector and of the antenna positions. In favorable cases, when $I_\nu(\boldsymbol{s})$ is that of a high intensity unresolved simple source, this procedure may be reversed. With a sufficient number of measurements of a large number of sources, high accuracy source positions, information on the geodetic data of the antenna positions, variations in the rotation rate of the Earth and highly accurate time keeping data can be derived.

   This has become a highly sophisticated subject using specialized measurements. We will outline the basic principles here and will not give too much detail which can be found in the literature.

   Let us assume that we have observations of an unresolved point source at the position $\boldsymbol{s}$ made with an interferometer with the baseline vector $\boldsymbol{B}$. The fringe phase φ then is given by Eq. (9.8), and an error $\Delta\boldsymbol{B}$ of $\boldsymbol{B}$ and a position error $\Delta\boldsymbol{s}$ then will result in a variation of the fringe phase

$$\varphi + \Delta\varphi = \Delta\varphi = \frac{1}{\lambda}(\boldsymbol{B} - \Delta\boldsymbol{B}) \cdot (\boldsymbol{s} - \Delta\boldsymbol{s}) + \varphi_i$$

or

$$\Delta\varphi = \varphi_i - \frac{1}{\lambda}(\Delta\boldsymbol{B} \cdot \boldsymbol{s} + \boldsymbol{B} \cdot \Delta\boldsymbol{s}) \tag{9.39}$$

if second order terms are neglected.

Introducing hour angle – declination as our coordinate system (X, Y, Z) and expressing both $\mathbf{B}$ and $\mathbf{s}$ in these coordinates we obtain

$$\Delta \varphi = \frac{1}{\lambda} \left[E \cos H + F \sin H + G \right] + \varphi_i \tag{9.40}$$

where

$$
\begin{aligned}
E &= (\Delta B_x + \Delta \alpha B_y) \cos \delta - \Delta \delta B_x \sin \delta \\
F &= (-\Delta B_y + \Delta \alpha B_x) \cos \delta + \Delta \delta B_y \sin \delta \\
G &= \Delta B_z \sin \delta + \Delta \delta B_z \cos \delta
\end{aligned}
\tag{9.41}
$$

and $(\Delta B_x, \Delta B_y, \Delta B_z)$ is the (true - adopted) antenna baseline vector in (metric) units, and $(\Delta \alpha, \Delta \delta)$ is the (true - adopted) source position.

The computations leading from (9.39) to (9.40) and (9.41) are straightforward but tedious and are standard for typical astrometric derivations. They may be found in the books by Thompson et al. (2001 in Chap. 12), or in Green (1985, Chaps. 15 and 16). According to (9.40) we will therefore observe a variation of the fringe phase with the hour angle when there is either a $\Delta \mathbf{B}$ of the baseline vector or a position error $(\Delta \alpha, \Delta \delta)$ or both. Unfortunately there is an instrumental term φ_i in this variation. But for stable connected type interferometric arrays of kilometer–size this instrumental phase can be kept reasonably constant and can be determined by an observing program in which one intersperses measurements of target sources with measurements of sources with accurately known positions. It is usually even possible to determine the total number of fringes. For km–size two–dimensional arrays accuracies of 50 mas are generally obtained and the grid of calibration sources are known with an accuracy of 10 mas.

VLBI global arrays have baselines 100 times longer, so the variations of the fringe phase are correspondingly larger. Then it is almost impossible to analyse it using Eq. (9.40), because the instrumental phase φ_i is mainly determined by instabilities of the independent local oscillators, ionosphere and atmosphere. Differentiating (9.40) with respect to time we find for the fringe rate variation

$$\Delta v = \frac{1}{\lambda} \left[-E \sin H + F \cos H \right] + v_i . \tag{9.42}$$

Δv shows a diurnial variation, so E and H and v_i can be determined. In order to get and idea of the precision obtainable consider two antennas separated by an equatorial spacing of 1000 km, observing at a wavelength of 3 cm then the fringe frequency will be ~ 2 kHz for $\delta \simeq 0$. Assuming a coherence time for the local oscillators of 10 min then about 10^6 fringes can be counted resulting in a precision of about 10^{-7}. The corresponding error for $|\mathbf{B}|$ is 10 cm or 0.02 arcsec.

In the last few years there have been a number of advances in the art of extracting the astronomical or geodetic data from VLBI observations. These advances concern predominantly the software by applying self–calibration (for astronomy) and multiband (for geodesy) schemes. These have resulted in absolute source positions to tenths of milliarcseconds and station locations to a few millimeters (Walker 1999).

If the best possible positional information in VLA and VLBI observations are wanted, all source positions have to be analyzed with the full set of the Earth's precession and nutation expressions. These describe the effects of the spin vector of the earth that may be as large as 50 arcsec per year and show how their the effect on the coordinate system may be taken into account. This is described in textbooks on spherical astronomy or astrometry. A reference that explicitly treats radio interferometry is by R. M. Green (1985).

Precession and nutation affect the coordinate system in which the position vector pointing towards the source region is expressed. There is an additional effect, the *polar motion* that influences the baseline vector B. While the first two act on the spin vector the polar motion describes the shifting of the axis of the figure of the earth relative to the spin. This motion is partly periodic, partly irregular and not well understood. Over the last century the distance between geographic pole and the earth's figure has varied by 0.5 arcsec or 15 m on the earth's surface. This effect must be taken into account when the highest possible positional precision is needed.

Comparing position information on the base vector B of VLBI data obtained at different times possible time variations of B can be estimated and these may be interpreted as estimates for the shift of tectonic plates. For long times, the precision of such data of the order of a few cm were unrivalled, but recently the accuracy of global position systems (GPS) have given hope that these methods may rival those obtained by VLBI techniques.

Problems

1. In one dimension, one can make a simple interferometer from a paraboloid by masking off all of the reflecting surface except for two regions of dimension a, which are separated by a length b, where $b >> a$. Assume that the power incident on these two regions is reflected without loss, then coherently received at the prime focus. A receiver there amplifies and square law detects these signals. This system is used to measure the response of an isolated source.
(a) Write out a one-dimensional version of Eq. (6.56). Apply this equation and Eq. (6.57) to determine the far-field pattern of this instrument.
(b) Use Eq. (9.6) to analyze the response of such a one-dimensional 2 element interferometer consisting of 2 paraboloids of diameter a, separated by a distance b, measuring a star by a disk of size θ_s. Show how one can determine the size of the star from the response, R.

2. Show that the one-dimensional version of Eq. (9.6) is Eq. (9.7). Rearranging terms, show that one obtains

$$R(B) = \int A(\theta) I_\nu(\theta) \exp\left[i2\pi\nu_0 \left(\frac{B}{\lambda}\right) \cdot \theta\right] d\theta \qquad (9.43)$$

Interpret this relation that the FT pair is u and θ. In Fig. 9.3 one dimensional distributions of $u = D/\lambda$ are shown. Show that the "Image-plane distributions" (on the

UV Plane Distribution   **Image Plane Distribution**

a) Discontinuous Outer Edge — Near in Sidelobes

b) Missing Short Spacings — Negative Bowl

c) Regular Gaps — Grating Response
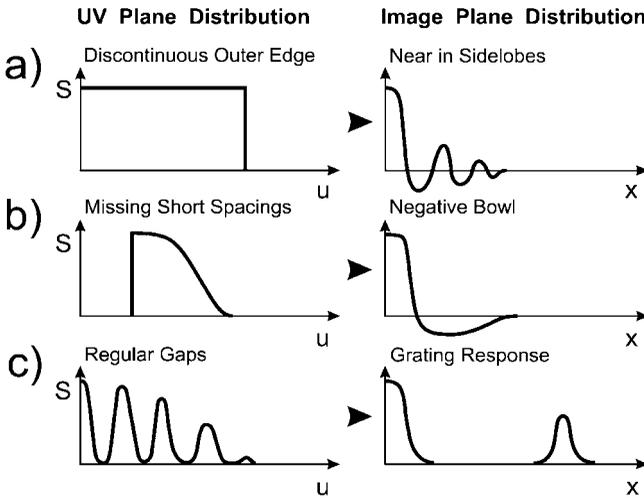
Fig. 9.13 One-dimensional (u, v) patterns on the *left*, and the resulting beams on the *right*

right) are related to the "one-dimensional" (u, v) plane distributions (on the left) using Eq. 9.43.

3. Use Eq. 9.43 to obtain the interferometer beam shapes (left side of Fig. 9.13 from the coverage of the (u, v) plane, shown on the right side of this figure.

4. The next three problems illustrate features of Fig. 9.3 and the use of Eq. 9.43. The Sun is assumed to be a uniformly bright disk of diameter $30'$. This source is measured using a multiplying interferometer at $10\,\text{GHz}$ which consists of two identical $1\,\text{m}$ diameter radio telescopes. Each of these dishes is uniformly illuminated. We assume that the instrumental phase is adjusted to zero, i.e. $\tau_i = 0$, the bandwidth of this system is small, and one measures the central fringe.
(a) What is the FWHP of each dish? Compare to the diameter of the Sun.
(b) Assume that the antenna efficiency and beam efficiency of each of the $1\,\text{m}$ telescopes are 0.5 and 0.7, respectively. What is the antenna temperature of the Sun, as measured with each? What is the main beam brightness temperature measured with each telescope?
(c) Now the outputs are connected as a multiplying interferometer, with a separation on an east–west baseline of $100\,\text{m}$. Suppose the Sun is observed when directly overhead. What is the fringe spacing? Express the response in terms of brightness temperature measured with each dish individually.
(d) Now consider the more general case of a source which is not directly overhead. Determine the response as a function of B, the *baseline*.
(e) What is the response when the two antennas are brought together as close as possible, namely $2\,\text{m}$?

5. Repeat Problem 4 for a simplified model of the radio galaxy Cygnus A. Take this source to be a one-dimensional double with centers separated by $2\theta_1 = 1.5'$.

Assume that each region have uniform intensity distributions, with FWHP sizes of $\theta_2 = 50''$. Each region has a total flux density of 50 Jy.

6. Repeat Problem 4 for the HII region Orion A, taking this as a one-dimensional Gaussian region with angular size FWHP 2.5′. Repeat for the supernova remnant Cassiopeia A, a ring-shaped source. In one dimension model this source is a region of outer diameter 5.5′ with a ring thickness of 1′.

7. Suppose the receivers of an interferometer are double-sideband mixers. In each mixer, power arrives from the *upper sideband* and from the *lower sideband*. Use Fig. 9.2 and Eq. 9.6 to show that the upper and lower sidebands can be separated since the geometric phase delays for the upper and lower sideband frequencies $\phi_g = 2\pi \nu \tau_g$, will differ.

8. The interferometer described in Problem 4 is used to measure the positions of intense water masers at 22.235 GHz. The individual masers are very compact sources, unresolved even with interferometer antenna spacings of hundreds to thousands of kilometers. These masers normally appear as clusters of individual sources, but usually do *not* have identical, radial velocities.
(a) Discuss using a set of contiguous narrow frequency filters as a spectrometer. Should these filters be placed *before* or *after* multiplication? How wide a frequency band can be analyzed without diminishing the response of this system? What must the phase and frequency characteristics of these filters be?
(b)* An alternative to filters is a *cross-correlation* spectrometer. Discuss how this system differs from the filter system. Analyze the response of such a cross correlator system if the instrumental phase differences between antennas can be eliminated before the signals enter the cross correlator.

9. Suppose we use an interferometer for which one: (1) added the voltage outputs of the two antennas, and then square-law detected this voltage and (2) inserted a phase difference of 180° into one of the inputs, (3) repeated this process and (4) then subtracted these outputs to obtain the correlated voltages. Compare the noise arising from this process with that from a direct multiplication of the voltages. Show that direct multiplication is more sensitive.

10. Derive Eq. 9.10 using Eq. 9.43. Derive Eq. 9.11. Show all steps in both derivations.

11. Use Eq. (9.38) with $M = 1$, to show that the following is an alternative form of this relation:

$$\Delta T_B \sim \frac{\lambda^{0.5} T_{sys} B_{max}^2}{n d^2 \sqrt{\tau \Delta V}},$$

where B_{max} is the maximum baseline of the interferometer system, d is the diameter of an individual antenna and ΔV is the velocity resolution. In addition, $N = n(n-1)/2 \approx n^2/2$ where n is the number of correlations.

12*. Suppose we have a filled aperture radio telescope with the same diameter and collecting area as an interferometer used to carry out a full synthesis.

(a) If the filled aperture diameter is $D(= B_{max}$ of Problem 11) and the diameter of each individual interferometer antenna is d, how many elements are needed to make up the interferometer? (This is the number of dishes of diameter d which fit into the area of the filled aperture D.)

(b) Calculate the times needed to map a region of a given size with the filled aperture (equipped with a single receiver) and the interferometer array.

(c) The following is related to "mosaicing", which is needed for interferometer imaging of a very extended source of size Θ which is *very extended* compared to the beamsize of each individual interferometer antenna, θ. Calculate how many pointings are needed to provide a complete image of the extended source. The RMS noise for a map made with a single pointing is

$$\Delta T_B \sim \frac{1}{nd^2\sqrt{2N\tau\Delta v}}$$

(see previous problem). If the total time available for the measurement of a region is T, show that the number of pointings is proportional to T/d^2. Then show that the RMS noise in a mosaiced map is $\Delta T_B \sim 1/d$ instead of $\Delta T_B \sim 1/d^2$.

13. A source with a FWHP of $\sim 30''$ and maximum intensity of 2.3 K, TMB is observed with ALMA interferometer. If a velocity resolution of $0.15\,\mathrm{km\,s^{-1}}$ is used to measure the $J = 1 - 0$ line of CO at 2.7 mm, with a $10''$ angular resolution, how long must one integrate to obtain a 5-to-1 peak signal-to-noise ratio?

14. The MERLIN interferometer system has a maximum baseline length of 227 km. At an observing frequency of 5 GHz, what is the angular resolution? Suppose that the RMS noise after a long integration is $50\,\mu$Jy, that is 5×10^{-5} Jy. Use the Rayleigh–Jeans relation to obtain the RMS noise in terms of main beam brightness temperature. If a thermal source has at most a peak temperature of 5×10^5 K, can one detect thermal emission?

Chapter 10
Emission Mechanisms
of Continuous Radiation

10.1 The Nature of Radio Sources

In the early days of radio astronomy the receiver sensitivities restricted measurements to the few hundred megahertz range. At such relatively low frequencies the resolving power of the available radio telescopes was low. Initially only very few of the discrete sources could be identified with objects known from the optical region of the spectrum. Further investigations showed an increase in the number of sources with decreasing source flux density and gave the distribution of sources in the sky. It was then concluded that there are two different families of sources: galactic sources, concentrated towards the galactic plane and extragalactic sources distributed more or less uniformly in space. The unresolved, spatially continuous radiation belongs to the galactic component. In addition, there is the 2.7 K thermal background radiation which is cosmological in origin.

The nature of the discrete sources was investigated by measurements at different frequencies to determine the spectral characteristics. Again two large families of sources appeared. While the flux density of one type of source is roughly constant with increasing frequency, the other type is more intense at lower frequencies (Fig. 10.1). Some of the most intense sources found were of the second type, for example the source Cassiopeia A which was later identified as the remnant of a supernova explosion in our Galaxy in the year 1667. Another prominent source is Cygnus A, an extragalactic radio source; Cygnus A is a radio galaxy, with a redshift of $z = 0.057$. Those sources which show an increasing flux density with increasing frequency could be identified with objects well known from the optical range of the spectrum. Both the moon and the sun are radio sources of this kind; for the sun this is true only if measurements are restricted to times when there are no disturbances, that is, to the quiet sun. Weaker emitters, such as the Orion nebula (Messier 42, NGC 1976) which is an H II region, and other similar sources such as M 17 which are optically thick at lower frequencies, belong to this category. The moon is an example of a black body and its spectrum is an almost exact representation of the Rayleigh-Jeans law for a temperature of $T \approx 225$ K. The spectrum of the Orion nebula also indicates an optically thick thermal origin for frequencies below 1 GHz. In that range the observations can be well represented by the Rayleigh-Jeans law, but
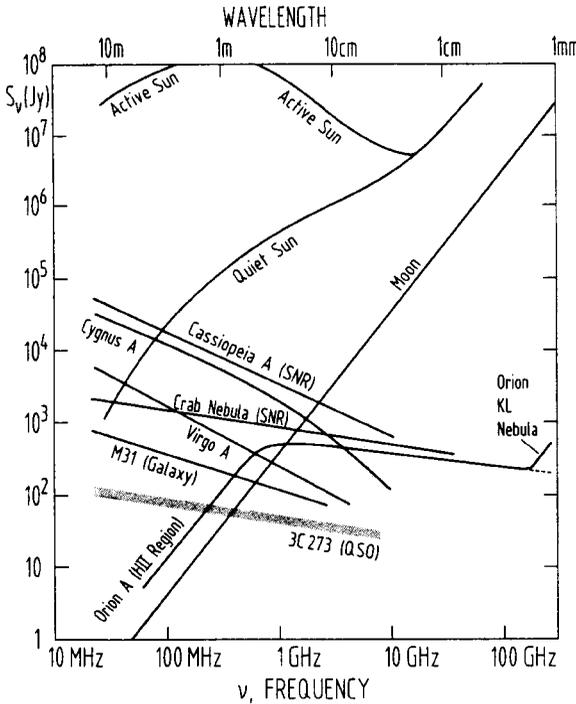
Fig. 10.1 The spectral distributions of various radio sources. The Moon, the quiet Sun and (at lower frequencies) the H II region Orion A are examples of Black Bodies. Close to 300 GHz there is additional emission from dust in the molecular cloud Orion KL. The active Sun, supernova remnants such as Cassiopeia A, the radio galaxies Cygnus A, Virgo A (Messier 87, 3C274) and the Quasi Stellar radio source (QSO) 3C273 are nonthermal emitters. The hatching around the spectrum of 3C273 is meant to indicate rapid time variability. (The 3C catalog is the fundamental list of intense sources at 178 MHz (Bennett 1962))

above this frequency the observed flux density falls below that extrapolated from lower frequencies.

For sources such as Orion A the explanation is fairly straightforward. If we consider the solution of the equation of radiation transfer (1.19) for an isothermal object without a background source

$$I_\nu = B_\nu(T) \left(1 - e^{-\tau_\nu}\right),$$

we see that $I_\nu < B_\nu$ if $\tau_\nu \lesssim 1$; the frequency variation of I_ν depends on τ_ν.

Radio sources can thus be classified into two categories: those which radiate by thermal mechanisms and the others, which radiate by nonthermal processes. In principle many different radiation mechanisms could be responsible for nonthermal emission, but in practice one single mechanism seems to dominate: synchrotron emission or magnetic bremsstrahlung.

The other division of the discrete radio sources into galactic and extragalactic ones, is in principle completely independent of this classification. However, we find

predominantly nonthermal sources among the extragalactic sources. This is simply a result of the fact that the most intense emitters are nonthermal in origin. Even if thermal sources are abundant in extragalactic objects, these will not easily be detected.

With the exception of thermal line emission of atoms and molecules and thermal emission from solid bodies, radio emission always arises from free electrons, and since free electrons can exchange energy by arbitrary amounts, no definite energy jumps – that result in sharp emission or absorption lines – will occur: thus we are dealing with a continuous spectrum. A free electron will only emit radiation if it is accelerated, and therefore the acceleration mechanisms for free electrons will have to be studied if the continuous radio spectrum of discrete sources is to be understood.

When astrophysicists had to explain the thermal radio emission of gaseous nebulae, it turned out to be unnecessary to make a completely new start when developing the physics of the processes involved. A starting point was the theory devised 40 years earlier to explain the continuous X-ray emission spectrum of ordinary X-ray tubes. Both in thermal radio emission and in X-ray tubes the electromagnetic radiation is emitted by free electrons that suffer sudden accelerations in the electric field of ions. The relative change of the kinetic energy of the electrons is of the same order of magnitude in both cases: radio radiation with $\nu = 1\,\text{GHz}$ and a quantum energy $h\nu \cong 4 \times 10^{-6}\,\text{eV}$ emitted by the thermal electrons in a hot plasma ($T \approx 10^4\,\text{K}$, $mv^2/2 = 1\,\text{eV}$). This corresponds to a relative change of the kinetic energy of $\Delta E/E \cong 10^{-5}$, while X-rays produced by a $100\,\text{keV}$ beam corresponds to changes $\Delta E/E \cong 10^{-2}$. These values are similar enough so that the same theory can be applied to both situations. We will present this theory not in its most general form but with simplifications that are valid for the radio range.

10.1.1   Black Body Radiation from Astronomical Objects

The first examples of black bodies found were solar system objects such as planets, and other solid bodies. For such objects, $\tau = \infty$, so from the equation of radiative transfer, (1.37) gives $I_\nu = B_\nu(T)$. A radio telescope measures the temperature at a depth corresponding to a wavelength or so. This temperature, referred to as the disk or surface temperature, is nearly constant with frequency. So the flux density increases with λ^{-2}, as shown in Fig. 10.1. In the last few years, the use of large millimeter radio telescopes and sensitive bolometers has given rise to the detection of asteroids, comets and moons of the outer planets. The broadband emission mechanism is the same as that for the Moon and planets, but as the newly found sources have small angular sizes, we have:

$$T_{\text{MB}} = f T_{\text{surface}} \tag{10.1}$$

where f is the beam filling factor. If the sizes are known, such observations can provide a reasonably good estimate of the surface temperature and albedo of such objects.

Another research area related to black body emission is the study of dust in molecular clouds (see Chap. 14). In the case of molecular clouds, the emission is from small dust particles. If we use the exact relation, we have

$$T_{\rm b}(\nu) = T_0 \left(\frac{1}{\exp\{T_0/T_{\rm dust}\} - 1} - \frac{1}{\exp\{T_0/2.7\} - 1} \right) (1 - {\rm e}^{-\tau_{\rm dust}}) , \qquad (10.2)$$

where $T_0 = h\nu/k$. This is completely general; if we neglect the 2.7 K background,

$$T_{\rm dust} = T_0 \left(\frac{1}{\exp\{T_0/T_{\rm dust}\} - 1} \right) (1 - {\rm e}^{-\tau_{\rm dust}}) . \qquad (10.3)$$

If $T_{\rm dust} \gg T_0$, we can simplify this expression, but the most important step is in making a quantitative connection between $\tau_{\rm dust}$ and $N_{\rm H_2}$. The relation between $\tau_{\rm dust}$ and the gas column density must be determined empirically. Unlike the planets, which have measured sizes, the radiation from dust grains depends on the surface area of the grains, which cannot be determined directly. If a relation between dust mass and τ can be determined, it is simple to convert to the total mass, since dust is generally accepted to be between 1/100 and 1/150th of the total mass. All astronomical determinations are based on the analysis of Hildebrand (1983). One typical parameterized version is given by Mezger et al. (1990) for $\lambda > 100\,\mu$m:

$$\tau_{\rm dust} = 7 \times 10^{-21} \frac{Z}{Z_\odot} b N_{\rm H} \lambda^{-2} \qquad (10.4)$$

where λ is the wavelength in μm, $N_{\rm H}$ is in units cm^{-2}, Z is the metalicity as a ratio of that of the sun $Z_\odot$. The parameter b is an adjustable factor used to take into account changes in grain sizes. Currently, it is believed that $b = 1.9$ is appropriate for moderate density gas and $b = 3.4$ for dense gas (but this is not certain). At long millimeter wavelengths, a number of observations have shown that the optical depth of such radiation is small. Then the observed temperature is

$$T = T_{\rm dust}\, \tau_{\rm dust} , \qquad (10.5)$$

where the quantities on the right side are the dust temperature and optical depth. Then the flux density is

$$S = \frac{2kT}{\lambda^2} = 2kT_{\rm dust}\lambda^{-2}\, \tau_{\rm dust}\Delta\Omega . \qquad (10.6)$$

If the dust radiation is expressed in Jy, the source in FWHP sizes, θ in arc seconds, and the wavelength, λ in μm, one has for the column density of hydrogen in all forms, $N_{\rm H}$, in the Rayleigh-Jeans approximation, the following relation:

If the dust radiation is expressed in mJy, the source FWHP size, θ, in arc seconds, and the wavelength, λ in mm, the column density of hydrogen in all forms, $N_{\rm H}$, in the Rayleigh-Jeans approximation, is:

$$N_{\mathrm{H}} = 1.93 \times 10^{24} \frac{S_\nu}{\theta^2} \frac{\lambda^4}{Z/Z_\odot \, b \, T_{\mathrm{dust}}} . \tag{10.7}$$

In the cm and mm wavelength range, the dust optical depth is small and increases with λ^{-2}; then flux density increases as λ^{-4}.

Observationally, it has been determined that, in most cases, the dust optical depth increases with λ^{-2}; then flux density increases as λ^{-4}. Thus, dust emission will become more important at millimeter wavelengths and in the infrared.

It appears that cold dust, with temperatures 10–30 K, makes up most of the mass of dust and by implication traces cold interstellar gas in our galaxy. For this temperature range, the mass can be estimated using either sub millimeter or far infrared data.

In addition to the total intensity of thermal dust emission, one can also measure the polarization properties of this emission. If the dust grains are non-spherical, and some process aligns the grains, one would expect the thermal emission to be polarized. In the simplest case, we expect the grains to have an electric charge, and a magnetic field causes the alignment; in this case, the largest dimension of the grains will be aligned perpendicular to the direction of the magnetic field and so the thermal emission will be more intense perpendicular to the B field. These measurements provide the direction of the B field, averaged along the line of sight, but cannot allow estimates of *B* field strengths. Since dust polarization at 0.87 mm is of order a few percent or less, and since the total power emission from dust is weak, polarization measurements require great care (see Chrysostomou et al. 2002; another method is the use of millimeter interferometry, see Marrone et al. 2007). Synchrotron radiation is also polarized; see Sects. 10.9.1 and 10.9.2.

Of course, the most famous example of black body radiation is the 2.73 K cosmic microwave background, CMB. This source of radiation is fit by a Planck curve to better than 0.1%. This radiation was discovered only in 1965. The difficulty in detecting it was not due to weak signals, but rather due to the fact that the radiation is present in all directions, so that scanning a telescope over the sky and taking differences will not lead to a detection. The actual discovery was due to absolute measurements of the noise temperature from the sky, the receiver and the ground, compared to the temperature of a helium cooled load using Dicke switching.

10.2 Radiation from Accelerated Electrons

For a general theory of the radiation from moving charges we must start with the appropriate electrodynamic potentials, the so-called Liénard-Wiechert potentials. These results can then be applied to all energy ranges, thermal as well as relativistic. The basic theories are presented in many textbooks on electrodynamics: Jackson (1975), Chaps. 14 and 15; Landau and Lifschitz (1967), Vol. II, §§ 68–71; Panofsky and Phillips (1962), Chaps. 19 and 20; Rybicki and Lightman (1979), Chaps. 3–6.tool-mar23. A delightful overview is given in Scheuer (1967).

We will not derive the full theory here from first principles but only cite those results that are useful for radio astronomy. The electric field induced at the position of the observer by a charge at the distance r moving with the velocity $v(t)$ is given by an expression that contains both the velocity $v(t)$ and the acceleration $\dot{v}$ of the charge e at the retarded time. While the terms depending on $v(t)$ vary like $1/r^2$ those depending on $\dot{v}$ vary only like $1/r$. If we observe the moving charge in a reference frame where its velocity is small compared to that of light, only field components depending on the acceleration are present. We orient the coordinate system such that $\vartheta = \pi/2$ points in the direction of v. Then we have

$$E_\vartheta = -\frac{e\,\dot{v}(t)}{c^2}\frac{\sin\vartheta}{r}\,. \tag{10.8}$$

The other components are zero. This is equivalent to the far field radiation of a Hertz dipole with

$$\frac{I\Delta l}{2\lambda} = \frac{e\,\dot{v}}{c^2}\,. \tag{10.9}$$

The Poynting flux, that is, the power per surface area and steradian emitted into the direction (ϑ, φ) is then according to (6.40)

$$|\mathbf{S}| = \frac{1}{4\pi}\frac{e^2\,\dot{v}^2}{c^3}\frac{\sin^2\vartheta}{r^2}\,. \tag{10.10}$$

Integrating this over the full sphere results in the total amount of power radiated when a charge e is accelerated by $\dot{v}$:

$$P(t) = \frac{2}{3}\frac{e^2\,\dot{v}^2(t)}{c^3}\,. \tag{10.11}$$

$P(t)$ is the power emitted at the moment t due to the acceleration $\dot{v}(t)$ of the electron. The total amount of energy emitted during the whole encounter is obtained by integrating (10.11); that is,

$$W = \int_{-\infty}^{\infty} P(t)\,\mathrm{d}t = \frac{2}{3}\frac{e^2}{c^3}\int_{-\infty}^{\infty} \dot{v}^2\,\mathrm{d}t\,. \tag{10.12}$$

In this development we have not specified the frequency at which this radiation will be emitted. This frequency is obviously governed by the speed with which E varies, that is, according to (10.8), the speed with which $\dot{v}(t)$ is varying. If $\dot{v}$ and E are different from zero only for a very short time interval then we can obtain the frequency dependence of E by a Fourier analysis of the pulse of E. But in order to do this we have to investigate the acceleration process in detail. In the next section we consider the radiation caused by the collision of the electron with an ion. This is necessary in order to compute the integral in (10.12).

10.3   The Frequency Distribution of Bremsstrahlung
     for an Individual Encounter

The parameter p/v alone determines the speed and strength of the electric field in-
tensity pulse. Here p is the *collision parameter* and v the *velocity* which the electron
attains when it is at closest approach, p, to the ion (Fig. 10.2). Now, the required rel-
ative change $\Delta E / E \cong 10^{-5}$ of the kinetic energy corresponds to such large values
of p that the electron's path, which in reality is a hyperbola, can be approximated
by a straight line. For the ion-electron distance l we then have

$$l = \frac{p}{\cos \psi} \, . \tag{10.13}$$

The acceleration of the electron is given by Coulomb's law

$$m\dot{v} = -\frac{Ze^2}{l^3} l \, .$$

For orbits with large p the acceleration is both parallel and perpendicular to the
orbit. The parallel component is nearly sinusoidal. We will simplify the problem by
only considering the perpendicular component,

$$\dot{v} = |\dot{v}| \cos \psi \, ,$$

so that

$$\dot{v} = -\frac{Ze^2}{m\,p^2} \cos^3 \psi \tag{10.14}$$

and

$$W = \frac{4}{3} \frac{Z^2 e^6}{c^3 m^2 p^4} \int_0^\infty \cos^6 \psi(t) \, dt \, . \tag{10.15}$$

The functional dependence $\psi(t)$ can be obtained from Kepler's law of areas, but
the energies of the electrons are low so that the electron velocites are nearly constant.
If we define
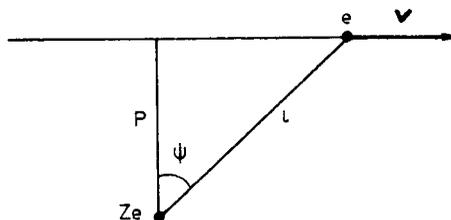
Fig. 10.2 An electron moving
past an ion of charge Ze

$$\mathrm{d}F = \tfrac{1}{2}\,l^2\,\mathrm{d}\psi,$$

then in any motion governed by a central force

$$\dot{F} = \frac{1}{2}\,l^2\,\frac{\mathrm{d}\psi}{\mathrm{d}t} = \mathrm{const}.$$

But at the time $t = 0$ the electron attains its closest approach p and has the velocity v, so

$$\dot{F} = \tfrac{1}{2}\,p\,v$$

and

$$\mathrm{d}t = \frac{l^2}{v\,p}\,\mathrm{d}\psi = \frac{p}{v\cos^2\psi}\,\mathrm{d}\psi. \tag{10.16}$$

Therefore we find

$$W = \frac{4}{3}\frac{Z^2\,e^6}{c^3\,m^2\,p^4}\frac{p}{v}\int\limits_{0}^{\frac{\pi}{2}}\cos^4\psi\,\mathrm{d}\psi$$

or, since

$$\int\limits_{0}^{\frac{\pi}{2}}\cos^4 x\,\mathrm{d}x = \frac{3}{16}\pi,$$

$$\boxed{W = \frac{\pi}{4}\frac{Z^2\,e^6}{c^3\,m^2\,p^3}\frac{1}{v}}. \tag{10.17}$$

This is the total energy radiated by the charge e if it moves in the field of an ion with the charge Ze. It is valid only for low-energy collisions for which the straight-line approximation is valid. For collisions with small p another approximation would have to be used.

The $\mathbf{E}$ field intensity induced by the accelerated electron during the encounter with the ion was given by (10.8). This again gives only a pulse $E(t)$ of the electric field intensity. Taking the Fourier transform of $E(t)$ we can represent this pulse as a wave packet formed by the superposition of harmonic waves of frequency ω; that is,

$$E(t) = \int\limits_{-\infty}^{\infty} A(\omega)\,\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega. \tag{10.18}$$

The wave amplitude $A(\omega)$ in (10.18) is given by

$$A(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} E(t)\,e^{i\omega t}\,dt\,, \qquad (10.19)$$

so that $A(\omega)$ is a real quantity if $E(t)$ is symmetric with respect to $t = 0$. According to (10.8) the Fourier analysis of $E(t)$ can be obtained by an analysis of $\dot{v}(t)$. We are therefore interested in

$$C(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \dot{v}(t)\cos\omega t\,dt\,. \qquad (10.20)$$

Substituting $\dot{v}(t)$ from (10.14) this is

$$C(\omega) = -\frac{Ze^2}{m\,p^2}\frac{1}{\pi} \int_{0}^{\infty} \cos\omega t \cos^3\psi(t)\,dt\,. \qquad (10.21)$$

Using (10.16) this can be written as an integral over ψ. The solution can be written in closed form using modified Bessel functions with an imaginary argument (Hankel functions), see e.g. Oster (1961), but the precise form of the spectrum $C(\omega)$ is of no great concern; crucial is that $C(\omega) > 0$ for only a finite range of ω (Fig. 10.3). We can estimate this limiting ω_g in the following way: The total amount of energy radiated in a single encounter as given by W in (10.17) must be equal to the sum of the energies radiated at the different frequencies. This can be stated as

$$\int_{0}^{\infty} |C(\omega)|^2\,d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\dot{v}(t)|^2\,dt\,. \qquad (10.22)$$
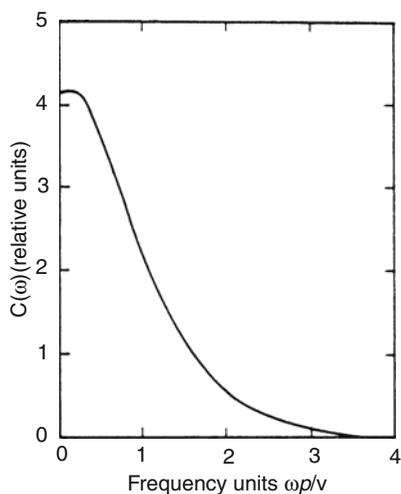
Fig. 10.3 The power spectrum of the radiation of a single encounter (Bekefi 1966). p is the impact parameter, v is the velocity, ω is the radiation frequency

This is nothing but the Rayleigh theorem in Fourier transformation theory. The left-hand side of (10.22) can be approximated by $C^2(0)\,\omega_g$, while (10.12) substituted into the right-hand side of (10.22) results in

$$C(0)\,\omega_g = \frac{3}{8\pi}\frac{c^3}{e^2}W\,.$$

From (10.21) and using (10.16) we obtain

$$C(0) = -\frac{Ze^2}{\pi m\,p\,v}\,, \tag{10.23}$$

so that finally, with (10.17), we find

$$\omega_g = \frac{3\pi^2}{64}\frac{v}{p} = 0.463\,\frac{v}{p}\,. \tag{10.24}$$

This is the limiting frequency below which the spectral density of the bremsstrahlung of an electron colliding with an ion can be considered to be flat.

10.4   The Radiation of an Ionized Gas Cloud

The radiation emitted by a single encounter of an electron and an ion thus depends in all its characteristic features – be this the total radiated energy W, its average spectral density or the limiting frequency v_g – on the collision parameters p and v. In a cloud of ionized gas these occur with a wide distribution of values, and thus the appropriate average will be the total radiation emitted by this cloud. In addition, the radiation of each collision will be polarized differently; then the resulting emission will be randomly polarized. If, therefore, a single polarization component is measured, this represents only $1/2$ of the total emitted power.

From the last section we adopt the result that a collision with the parameters p and v will emit bremsstrahlung with a flat spectrum, which, using (10.17) and (10.24), is

$$P_v(p,v) = \begin{cases} \dfrac{16}{3}\dfrac{Z^2 e^6}{c^3 m^2}\dfrac{1}{p^2 v^2} & \text{for } v < v_g = \dfrac{3\pi}{64}\dfrac{v}{p} \\[2mm] 0 & \text{for } v \gtrsim v_g \end{cases} \tag{10.25}$$

Since we consider only collisions with small relative energy changes $\Delta E\,/\,E$ for the electrons, their velocity is changed only very slightly during the collision, so that a Maxwell distribution can be adopted for the distribution function of v:

$$f(v) = \frac{4v^2}{\sqrt{\pi}}\left(\frac{m}{2kT}\right)^{3/2}\exp\left\{-\frac{mv^2}{2kT}\right\}\,. \tag{10.26}$$

The number of electrons in a unit volume of space with a given speed $v = \sqrt{v^2}$ that will pass by a given ion with a collision parameter between p and $p + \mathrm{d}p$ is (Fig. 10.4)

$$2\pi p \, \mathrm{d}p \, v \, N_e \, f(v) \, \mathrm{d}v.$$

But since there are N_i ions per unit volume, a total of

$$\mathrm{d}N(v, p) = 2\pi N_i N_e \, v \, p \, f(v) \, \mathrm{d}v \, \mathrm{d}p \qquad (10.27)$$

collisions with collision parameters between p and $p + \mathrm{d}p$ and velocities between v and $v + \mathrm{d}v$ will occur per second. As a result of these collisions a power of $4\pi\varepsilon_v \, \mathrm{d}v$ will be radiated in the frequency interval v to $v + \mathrm{d}v$ given by

$$4\pi\varepsilon_v \, \mathrm{d}v = P_v(v, p) \, \mathrm{d}N(v, p) \, \mathrm{d}v.$$

Substituting here (10.25) and (10.27) and integrating p from p_1 to p_2 and v from 0 to ∞ we find

$$4\pi\varepsilon_v = \int_{p_1}^{p_2} \int_0^\infty \frac{8}{3} \frac{Z^2 e^6}{c^3 m^2} \frac{1}{p^2 v^2} N_i N_e \, f(v) \, 2\pi \, p \, v \, \mathrm{d}p \, \mathrm{d}v$$

$$= \frac{32\pi}{3} \frac{Z^2 e^6}{c^3} \frac{N_i N_e}{m^2} \int_0^\infty \frac{1}{v} f(v) \, \mathrm{d}v \int_{p_1}^{p_2} \frac{\mathrm{d}p}{p}.$$

Using (10.26) the first integral becomes

$$\int_0^\infty \frac{1}{v} f(v) \, \mathrm{d}v = \sqrt{\frac{2m}{\pi k T}},$$

so that finally

$$\varepsilon_v = \frac{8}{3} \frac{Z^2 e^6}{c^3} \frac{N_i N_e}{m^2} \sqrt{\frac{2m}{\pi k T}} \ln \frac{p_2}{p_1}. \qquad (10.28)$$
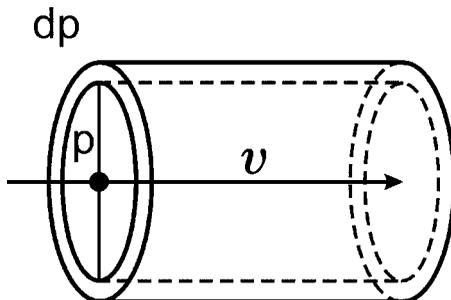
Fig. 10.4 A sketch illustrating the derivation of the probability of collisions given the impact parameter p

This is the coefficient for thermal emission of an ionized gas cloud. It contains the appropriate limits for the collision parameter p_1 and p_2. If either $p_1 \rightarrow 0$ or $p_2 \rightarrow \infty$, ε_ν diverges logarithmically; therefore appropriate values for these limits have to be estimated. For p_2, an upper limit should be the mean distance between the ions or the Debye length in the plasma. This is a very large quantity in the interstellar medium, so $p_2 = v/2\pi\nu$. The values of p_1 and p_2 are traditionally collected into the *Gaunt* factor.

Oster (1961) arrives at

$$\frac{p_{\max}}{p_{\min}} = \frac{p_2}{p_1} = \left(\frac{2kT}{\gamma m}\right)^{3/2} \frac{m}{\pi \gamma Z e^2 v}, \tag{10.29}$$

where $\gamma = e^C = 1.781$ and $C = 0.577$, Euler's constant, which is valid as long as $T > 20\,\mathrm{K}$ and $\nu_{\max} > 30\,\mathrm{GHz}$.

If the emission coefficient is known the absorption coefficient κ_ν can be computed using Kirchhoff's law (1.14):

$$\kappa_\nu = \frac{\varepsilon_\nu}{B_\nu(T)}$$

where $B_\nu(T)$ is the Planck function. Using the Rayleigh-Jeans approximation, we obtain

$$\kappa_\nu = \frac{4Z^2 e^6}{3c} \frac{N_{\mathrm{i}} N_{\mathrm{e}}}{v^2} \frac{1}{\sqrt{2\pi(mkT)^3}} \ln\frac{p_2}{p_1} \quad . \tag{10.30}$$

If we assume that the plasma is macroscopically neutral and that the chemical composition is given approximately by $N_{\mathrm{H}} : N_{\mathrm{He}} : N_{\mathrm{other}} \cong 10 : 1 : 10^{-3}$, then to high accuracy $N_{\mathrm{i}} = N_{\mathrm{e}}$. If further T_{e} is constant along the line of sight in an emission nebula, then it is useful to insert in the formula for the optical depth

$$\tau_\nu = -\int_0^s \kappa_\nu \, ds, \tag{10.31}$$

both (10.30) and the *emission measure* EM as given by

$$\frac{\mathrm{EM}}{\mathrm{pc\,cm^{-6}}} = \int_0^{s/\mathrm{pc}} \left(\frac{N_{\mathrm{e}}}{\mathrm{cm^{-3}}}\right)^2 d\left(\frac{s}{\mathrm{pc}}\right) \quad . \tag{10.32}$$

Substituting numerical values in (10.30), we have

$$\tau_\nu = 3.014 \times 10^{-2} \left(\frac{T_{\mathrm{e}}}{\mathrm{K}}\right)^{-3/2} \left(\frac{\nu}{\mathrm{GHz}}\right)^{-2} \left(\frac{\mathrm{EM}}{\mathrm{pc\,cm^{-6}}}\right) \langle g_{\mathrm{ff}} \rangle \quad , \tag{10.33}$$

where the Gaunt factor for free-free transitions is given by

$$\langle g_{\mathrm{ff}} \rangle = \begin{cases} \ln\left[4.955 \times 10^{-2} \left(\dfrac{\nu}{\mathrm{GHz}}\right)^{-1}\right] + 1.5 \ln\left(\dfrac{T_{\mathrm{e}}}{\mathrm{K}}\right) \\[2mm] 1 \quad \text{for} \quad \dfrac{\nu}{\mathrm{MHz}} \gg \left(\dfrac{T_{\mathrm{e}}}{\mathrm{K}}\right)^{3/2} \end{cases}. \tag{10.34}$$

Approximating $\langle g_{\mathrm{ff}} \rangle$ by $\alpha T^{\beta} \nu^{\gamma}$ and substituting this into (10.33), the simpler expression (Altenhoff et al. 1960)

$$\tau_{\nu} = 8.235 \times 10^{-2} \left(\frac{T_{\mathrm{e}}}{\mathrm{K}}\right)^{-1.35} \left(\frac{\nu}{\mathrm{GHz}}\right)^{-2.1} \left(\frac{\mathrm{EM}}{\mathrm{pc\,cm}^{-6}}\right) a(\nu, T) \tag{10.35}$$

can be derived. The correction $a(\nu, T)$ is usually $\cong 1$. Substituting (10.35) into (1.19) with the background set to zero, we find that the brightness of a gas cloud is as given in Fig. 10.5. Using (1.37) this can be converted into the frequency distribution $T_{\mathrm{b}}(\nu)$ shown in this diagram. For an optically thin H II region we have

$$T_{\mathrm{B}} = T_{\mathrm{e}} \tau_{\nu} = 8.235 \times 10^{-2} \left(\frac{T_{\mathrm{e}}}{\mathrm{K}}\right)^{-0.35} \left(\frac{\nu}{\mathrm{GHz}}\right)^{-2.1} \left(\frac{\mathrm{EM}}{\mathrm{pc\,cm}^{-6}}\right) a(\nu, T) \tag{10.36}$$
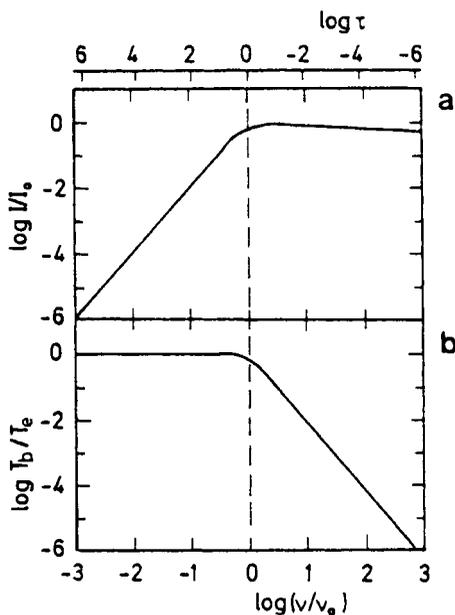
Fig. 10.5 Thermal radiation of a gas cloud. (a) Spectral distribution of the intensity. I_0 is the intensity, the electron temperature is T_{e}, and τ is the optical depth, ν_0 is the turn over frequency, where $\tau = 1$. (b) Spectral distribution of the brightness temperature

where once again, $a(v, T)$ is usually $\cong 1$.

When the frequency is given in units of v_0, where v_0 is that frequency at which the optical depth is unity, we have, from (10.35):

$$\frac{v_0}{\text{GHz}} = 0.3045 \left(\frac{T_e}{K}\right)^{-0.643} \left(\frac{a(v, T)\,\text{EM}}{\text{pc\,cm}^{-3}}\right)^{0.476} \qquad . \tag{10.37}$$

10.5 Nonthermal Radiation Mechanisms

Even though it had become clear that there are two kinds of radio sources, thermal and nonthermal, the radiation mechanism of the second type of source long remained an enigma since the proposed mechanisms were all unsatisfactory in one way or another. The solution to this problem was presented in two papers in 1950, one by Alfvén and Herlofson, the other by Kiepenheuer.

Alfvén and Herlofson proposed that the nonthermal radiation is emitted by relativistic electrons moving in intricately "tangled" magnetic fields of extended coronas believed to surround certain kinds of stars. Kiepenheuer modified this proposal, showing that the intensity of the nonthermal galactic radio emission can be understood as the radiation from relativistic cosmic ray electrons that move in the general interstellar magnetic field. Kiepenheuer deduced that a field of 10^{-6} Gauss and relativistic electrons of an energy of 10^9 eV would give about the observed intensity if the electron concentration is a small percentage of that of the heavier particles in cosmic rays.

This solution to the nonthermal radiation enigma proved to be tenable, and beginning in 1951 Ginzburg and collaborators developed this idea further in a series of papers. The radiation was given the name "synchrotron radiation" since it was observed to be emitted by electrons in synchrotrons, but much of the relevant theory had been developed earlier by Schott (1907, 1912) who used the name "magneto bremsstrahlung".

A fairly complete exposition of this theory which includes the polarization properties of this radiation is presented in the books of Pacholczyk (1970, 1977) where extensive bibliographical notes are also given. Very readable accounts of the theory are given by Rybicki and Lightman (1979) or by Tucker (1975). On a qualitative level, the article by Scheuer (1967) is to be recommended.

In radio astronomy we usually observe only the radiation from an ensemble of electrons with an energy distribution function that spans a wide range of electron energies. The electron distribution functions can quite often be expressed as a power law, and many of the details of the synchrotron radiation are thus lost in the averaging process. Therefore we will present here only a more qualitative discussion emphasizing, however, the underlying basic principles. In this we will follow the development by Rybicki and Lightman.

Since the electrons that emit the synchrotron radiation are highly relativistic, Lorentz transformations must be used to describe their motion. In the full theory starting with the Liénard-Wiechert potentials of moving charges, all this is taken into account, since Maxwell's equations are invariant under Lorentz transformations. Here we will start with a short review of these transformations.

10.6   Review of the Lorentz Transformation

The special theory of relativity is based on two postulates.

1) The laws of nature are the same in any two frames of reference that are in uniform relative motion.
2) The speed of light, c, is constant in all such frames.

If we now consider two frames K and K' that move with relative uniform velocity v along the x axis (Fig. 10.6), and if we assume space to be homogeneous and isotropic, then the coordinates in the two frames are related by the Lorentz transformations

$$\boxed{\begin{aligned} x &= \gamma(x' + vt') \\ y &= y' \\ z &= z' \\ t &= \gamma\left(t' + \beta\frac{x'}{c}\right) \end{aligned}} \qquad (10.38)$$

where

$$\beta = \frac{v}{c} \qquad (10.39)$$

and

$$\gamma = (1 - \beta^2)^{-1/2}. \qquad (10.40)$$
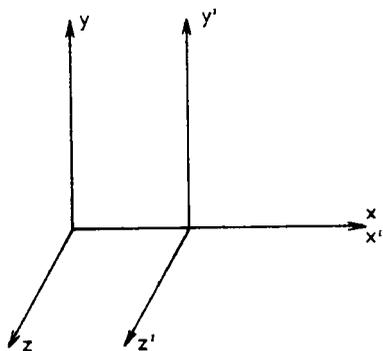
Fig. 10.6 Coordinate systems in relative motion. In the (x,y,z) system, (x',y',z') moves to the right with constant velocity, v

Define the (three) velocity $\boldsymbol{u}$ by

$$u_x = \frac{dx}{dt} = \frac{dx}{dt'}\frac{dt'}{dt} = \gamma(u'_x + v)\frac{dt'}{dt},$$

$$\frac{dt}{dt'} = \gamma\left(1 + \beta\frac{u'_x}{c}\right) = \gamma\sigma, \tag{10.41}$$

$$\sigma = 1 + \beta u'_x/c,$$

so that

$$u_x = \frac{u'_x + v}{1 + \beta\frac{u'_x}{c}} = \frac{1}{\sigma}(u'_x + v), \tag{10.42}$$

$$u_y = \frac{u'_y}{\gamma\left(1 + \beta\frac{u'_x}{c}\right)} = \frac{1}{\gamma\sigma}u'_y, \tag{10.43}$$

$$u_z = \frac{u'_z}{\gamma\left(1 + \beta\frac{u'_x}{c}\right)} = \frac{1}{\gamma\sigma}u'_z. \tag{10.44}$$

For an arbitrary relative velocity $\boldsymbol{v}$ that is not parallel to any of the coordinate axes, the transformation can be formulated in terms of the velocity components of u parallel and perpendicular to $\boldsymbol{v}$

$$\boxed{u_\parallel = \frac{u'_\parallel + v}{1 + \beta\frac{u'_\parallel}{c}}}, \tag{10.45}$$

$$\boxed{u_\perp = \frac{u'_\perp}{\gamma\left(1 + \beta\frac{u'_\parallel}{c}\right)}}. \tag{10.46}$$

Although special relativity theory is concerned with the transformation of physical quantities moving with uniform velocity, it is also possible to compute how an acceleration is transformed. Differentiating (10.42) to form

$$a_x = \frac{du_x}{dt} = \frac{du_x}{dt'}\frac{dt'}{dt}$$

and using (10.41) so that

$$\frac{d\sigma}{dt'} = \frac{\beta}{c}a'_x, \tag{10.47}$$

we obtain

$$a_x = \dot{u}_x = \gamma^{-3}\sigma^{-3}a'_x \tag{10.48}$$

and

$$a_y = \dot{u}_y = \gamma^{-2}\sigma^{-3}\left(\sigma a'_y - \beta\frac{u'_y}{c}a'_x\right). \tag{10.49}$$

For the special case when the particle is at rest in system K', that is, that initially $u'_x = u'_y = u'_z = 0$, we have $\sigma = 1$ and

$$a'_\| = \gamma^3 a_\|,$$
$$a'_\perp = \gamma^2 a_\perp. \tag{10.50}$$

The transformation equations (10.42–10.44) are fairly complicated and are different from the Lorentz transformations. This is so because neither the velocity u_x, u_y, u_z nor the accelerations a_x, a_y are components of four-vectors, and only four-vectors obey the simple transformation laws of the coordinates.

Equation (10.41) describes the time dilation, giving the different rate at which a clock at rest in the system K' is seen in the system K:

$$\Delta t = \gamma \Delta t'. \tag{10.51}$$

This should not be confused with the apparent time dilatation (contraction) known as the *Doppler effect*. Let the distance between clock and observer change with the speed v_r. During the intrinsic time interval $\Delta t'$ the clock has moved, so that the next pulse reaches the observer at

$$\Delta t = \left(1 + \frac{v_r}{c}\right)\gamma \Delta t'. \tag{10.52}$$

The factor γ is a relativistic effect, whereas $(1 + v_r/c)$ is due to nonrelativistic physics.

10.7   The Synchrotron Radiation of a Single Electron

The motion of a particle with the charge e and the mass m that moves with a (three) velocity v in a (homogeneous) magnetic field with the flux density B is governed by the relativistic Einstein-Planck equations

$$\frac{d}{dt}(\gamma m v) = \frac{e}{c}(v \times B). \tag{10.53}$$

If there is no electric field E, then energy conservation results in the additional equation

$$\frac{d}{dt}(\gamma m c^2) = 0.\tag{10.54}$$

But this implies that γ is a constant and therefore that $|v|$ is a constant. We project v into two components, $v_\parallel$ parallel to B and $v_\perp$ perpendicular to B. Then we find that

$$\frac{dv_\parallel}{dt} = 0\tag{10.55}$$

and

$$\frac{dv_\perp}{dt} = \frac{e}{\gamma m c}(v_\perp \times B).\tag{10.56}$$

Equation (10.55) has the solution $v_\parallel =$ constant so that, since $|v|$ is constant, $|v_\perp|$ must also be constant. The solution to (10.56) therefore must obviously be uniform circular motion with a constant orbital velocity $v_\perp = |v|$. The frequency of the gyration is

$$\omega_B = \frac{e B}{\gamma m c}, \qquad \omega_G = \frac{e B}{m c} = \gamma \omega_B.\tag{10.57}$$

Since the constant velocity $v_\parallel$ is superimposed on this circular motion, the path followed by the electron is a helix winding around B with the constant pitch angle

$$\tan \alpha = \frac{|v_\perp|}{|v_\parallel|}.\tag{10.58}$$

Inserting numerical values for e and m, we find

$$\frac{\omega_G}{\text{MHz}} = 17.6 \left(\frac{B}{\text{Gauss}}\right)$$

so that, for $B \cong 10^{-6}$ Gauss in interstellar space, $\omega_G = 18$ Hz. For a relativistic particle with $\gamma > 1$ ω_B will be even smaller! From (10.56) we see that the electron is accelerated in its orbit, this acceleration is directed $\perp$ to B and its magnitude is

$$a_\perp = \omega_B v_\perp.\tag{10.59}$$

Since the electron is accelerated, it will radiate. We will investigate this radiation in a three-step procedure for a single electron: (1) obtain the total power radiated, (2) derive the polar radiation pattern, and (3) calculate the frequency distribution of the emission. We will then consider an ensemble of electrons with a power law distribution.

10.7.1   The Total Power Radiated

Let us assume $v_{\parallel} = 0$ for simplicity. If we then select an inertial frame K' that moves with respect to rest frame K such that the electron is at rest at a certain time, then this particle will not remain at rest for long since its acceleration is not zero. However, for an infinitesimal time interval we can adopt this assumption. In frame K', the electron is nonrelativistic and radiates according to the Larmor formula (10.11)

$$P' = \frac{2e^2}{3c^3} a'^2_{\perp}, \tag{10.60}$$

where $a'_{\perp}$ is the acceleration of the electron in the rest frame K'.

We will now transform the emitted power P' into the rest frame K which moves relative to K' with the velocity $v_{\perp}$. The energy is one component of the four-component vector (momentum, energy), and it is transformed accordingly by

$$W = \gamma W'.$$

For a time interval (at constant space position) we have similarly

$$dt = \gamma dt'$$

so that for the power emitted

$$P = \frac{dW}{dt}, \qquad P' = \frac{dW'}{dt'}$$

we find that

$$P = P'. \tag{10.61}$$

Considering in addition the transformation of the acceleration (10.49) we find

$$P = \frac{2e^2}{3c^3} \gamma^4 a^2_{\perp} = \frac{2e^2}{3c^3} a'^2_{\perp} = P'. \tag{10.62}$$

Introducing the total energy E of the electron through

$$\gamma = \frac{E}{mc^2} \tag{10.63}$$

and using (10.59) we obtain as the power emitted by a relativistic ($\beta \cong 1$) electron

$$P = \frac{2e^4 v^2_{\perp} B^2}{3m^2 c^5} \left(\frac{E}{mc^2} \right)^2 = 2\sigma_T \gamma^2 c u_B \quad , \tag{10.64}$$

where $u_B = B^2/8\pi$ and $\sigma_T = 6.65 \times 10^{-25}$ cm^2 is the Thomson cross section. Note that the energy E introduced in (10.63) and (10.64) is a quantity entirely different

from the electric field intensity E of (10.8) or of Chaps. 2, 3, 4, 5, 6. This unfortunate double meaning for E cannot be avoided if we conform to the general usage. In the subsequent sections, E will always be used in the sense of (10.63).

10.7.2 The Angular Distribution of Radiation

The electron moving along the Larmor circle is radiating because it is accelerated. Because this acceleration is directed perpendicular to the direction of motion (in the system K'), the power pattern of this emission will be that of a dipole field with the dipole oriented along the direction of the acceleration. This power P' has an angular distribution given by

$$\frac{\mathrm{d}P'(\vartheta', \varphi')}{\mathrm{d}\Omega'} = \frac{1}{4\pi} \frac{e^2}{c^3} a_\perp'^2 (1 - \sin^2 \vartheta' \cos^2 \varphi') \tag{10.65}$$

and shown in Fig. 10.7. Transformed into the laboratory frame K this becomes

$$\frac{\mathrm{d}P(\vartheta, \varphi)}{\mathrm{d}\Omega} = \frac{1}{4\pi} \frac{e^2}{c^3} a_\perp'^2 \frac{1}{(1 - \beta \cos \vartheta)^3} \left\{ 1 - \frac{\sin^2 \vartheta \cos^2 \varphi}{\gamma^2 (1 - \beta \cos \vartheta)^2} \right\} \tag{10.66}$$

as shown in Fig. 10.7 for an extremely mild relativistic case of $\beta = 0.2$.

Inspecting the low velocity limit $\beta \to 0$ represented by (10.65) we see that no radiation is emitted towards $\vartheta = 0, \pi$. Thus we have $u_\perp' = \pm c$ and $u_\parallel' = 0$ for these directions, and in the laboratory frame K they subtend the angle

(a)   (b)
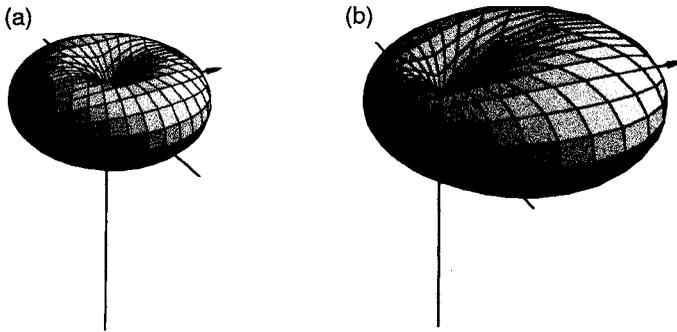
Fig. 10.7 Instantaneous emission cones for an electron gyrating in a homogeneous magnetic field. The electron is moving to the right in the plane of the paper, and the acceleration is directed towards the bottom. θ and ϕ are angles in a polar coordinate system, $\theta = 0$ is the direction of the acceleration, $\phi = 0$ points towards the direction of the motion for the electron. (**a**) shows the emission for an electron with $\beta = v/c \ll 1$, while (**b**) gives that for $\beta = 0.2$. When $\beta \simeq 1$ the emission cone degenerates into a narrow pencil beam subtending the angle $\tan \theta = c/\gamma v$

$$\tan \theta = \frac{u_\perp}{u_\parallel} = \frac{c}{\gamma v} = \frac{1}{\gamma}. \tag{10.67}$$

The power P of (10.66) is then confined to a cone with the angle θ and we obtain a strong beaming effect for relativistic electrons.

10.7.3   The Frequency Distribution of the Emission

The spectrum of synchrotron radiation is related to the detailed form of the variation of the electromagnetic field as seen by the observer; the strong beaming effect that the dipole field experiences for a gyrating relativistic electron is the reason why such electrons emit radiation with fairly high frequencies although the cyclotron frequency ω_G of these particles (according to (10.57)) is very low indeed. An external observer will be able to "see" the radiation from the electron only for a time interval when the cone of emission of angular width $2/\gamma$ includes the direction of observation.

In the comoving frame K' of the electron the time for one gyration is

$$\Delta t' = \frac{2\pi}{\omega_G}.$$

The pulse of radiation is visible, however, only for $2/2\pi\gamma$ due to the beaming effect, so that the pulse emitted lasts only for the time interval

$$\Delta t = \frac{2\pi}{\omega_G} = \frac{2}{2\pi\gamma} = \frac{2}{\gamma \omega_B}.$$

In the relevant part of the orbit the electron moves towards the observer, so that $v_r = -v$, and therefore is the pulse width in the frame K of the observer using the Doppler formula

$$\Delta t' = \gamma \left(1 - \frac{v}{c}\right) \Delta t' = \frac{2}{\omega_G} \left(1 - \frac{v}{c}\right).$$

Now

$$\left(1 + \frac{v}{c}\right) \left(1 - \frac{v}{c}\right) = 1 - \frac{v^2}{c^2} = 1 - \beta^2 = \frac{1}{\gamma^2}$$

and

$$1 + \frac{v}{c} \cong 2 \ \text{ for } \ \gamma \gg 1,$$

so that

$$1 - \frac{v}{c} \cong \frac{1}{2\gamma^2}$$

and

$$\Delta t = \frac{1}{\gamma^3 \, \omega_B} = \frac{1}{\gamma^2 \, \omega_G} \,.$$

If, in addition, the electron has some velocity component $v \neq 0$ parallel to the magnetic field, it will move in a helix with a pitch angle α given by (10.58)

$$\tan \alpha = \frac{|v_\perp|}{|v_\parallel|} \,, \tag{10.68}$$

and we obtain approximately

$$\Delta t = \frac{1}{\gamma^3 \, \omega_B} \frac{1}{\sin \alpha} \,. \tag{10.69}$$

Radiation from the electron is therefore only seen for a time interval $\sim 1/\gamma^3$. If the pulse is Fourier analyzed to derive the frequency distribution of the radiation we see that very high harmonics of ω_B will be present.

Tracing back the derivation of (10.57) we see that the factors γ^2, $\gamma^3 \, \omega_G$ and ω_B arise from the following roots. The relativistic ω_B is lower than ω_G by the factor $1/\gamma$ (10.57) because of the relativistic mass increase. The emitted pulse width becomes shorter by the factor $1/\gamma$ due to relativistic beaming effect, and finally the Doppler effect shortens the emitted pulse by another factor $1/\gamma^2$. Therefore the received pulse width is shorter by $1/\gamma^3$ when the nonrelativistic ω_B is used, while $1/\gamma^2$ enters for ω_G.

In effect, this factor causes energy emitted by the accelerated electron over a fairly long time interval to pile up in one short pulse. This effect dramatically increases both the radiated power and the highest radiated frequency. This frequency distribution can be derived by Fourier analysis of the pulse shape; the details can be found in Westfold (1959) or Pacholczyk (1970).

The total emissivity of an electron of energy E with $\gamma \gg 1$ which has a pitch angle α with respect to the magnetic field is

$$P(\nu) = \sqrt{3} \, \frac{e^3 B \sin \alpha}{m c^2} \frac{\nu}{\nu_c} \int_{\nu/\nu_c}^{\infty} K_{5/3}(\eta) \, d\eta \tag{10.70}$$

and the critical frequency ν_c is defined by

$$\nu_c = \tfrac{3}{2} \gamma^2 \nu_G \sin \alpha = \tfrac{3}{2} \gamma^3 \nu_B \sin \alpha \,, \tag{10.71}$$

where $\nu_G = \omega_G/2\pi$ is the non relativistic gyro-frequency according to (10.58), and $K_{5/3}$ is the modified Bessel function of $5/3$ order.

10.8   The Spectrum and Polarization of Synchrotron Radiation

If a more precise description of the radiation field of highly relativistic electrons orbiting in a homogeneous magnetic field is required, more details of the orbit geometry have to be considered. Let the electron move with the (constant) velocity components $v_\parallel$ and $v_\perp$ measured with respect to the direction of the magnetic field. The emitted radiation will then be strongly beamed in the direction of the instantaneous velocity of the electron. Since this velocity vector describes a cone with the opening angle α given by (10.58) whose axis is the direction of the magnetic field, this is also the direction for the emission pattern of the radiation. The narrowness of the instantaneous beam is determined solely by the energy γ of the electron. The other properties of the radiation field depend on the angle between the observer and the velocity vector.

Obviously radiation is only detected if the direction to the observer is inside the emission beam. The instantaneous radiation is, in general, elliptically polarized, but since the position angle of the polarization ellipse is rotating with the electron, the time-averaged polarization is linear. This is true also for the radiation emitted by an ensemble of monoenergetic electrons moving in parallel orbits.

The details of the Fourier expansion and the integration procedures needed are rather complicated. If expressions for the time averaged radiation integrated over the full sky are required, they are given by Westfold (1959), Ginzburg Syrovatskii (1965) and Pacholczyk (1970). According to these, the spectral radiation density averaged over all directions for linear polarizations parallel and perpendicular to the (projected) magnetic field is given by

$$P_\perp = \frac{\sqrt{3}}{2} \frac{e^3 B \sin \alpha}{m c^2} \left[F(x) + G(x) \right] , \tag{10.72}$$

$$P_\parallel = \frac{\sqrt{3}}{2} \frac{e^3 B \sin \alpha}{m c^2} \left[F(x) - G(x) \right] , \tag{10.73}$$

where

$$\boxed{F(x) = x \int_x^\infty K_{5/3}(t) \, dt} \quad , \tag{10.74}$$

$$\boxed{G(x) = x K_{2/3}(x)} \quad , \tag{10.75}$$

are shown in Fig. 10.8 and

$$\boxed{x = \nu / \nu_c} \quad , \tag{10.76}$$

$$\boxed{\nu_c = \frac{3}{2} \gamma^2 \frac{eB}{mc} \sin \alpha} \quad . \tag{10.77}$$
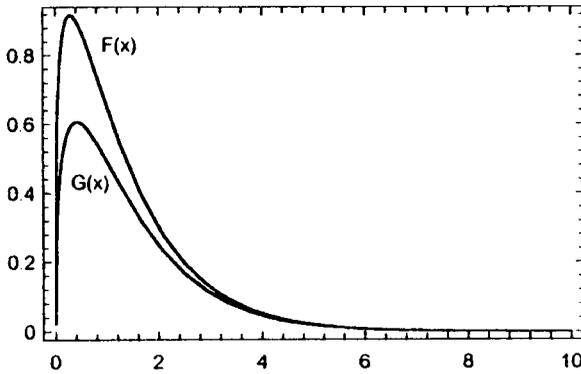
Fig. 10.8 The spectral distribution of the power of synchrotron radiation. The functions $F(x)$ and $G(x)$ are related by (10.72) and (10.73) to the linear polarization components parallel and perpendicular to the magnetic field, the frequency (10.76) is normalized by (10.77)

Table 10.1 Spectral Distribution of the emission from a charged particle moving in a magnetic field

x	$F(x)$	$G(x)$	p	x	$F(x)$	$G(x)$	p
0.00	0.0000	0.0000	0.500	1.00	0.6514	0.4945	0.759
0.01	0.4450	0.2310	0.519	1.10	0.6075	0.4669	0.769
0.02	0.5472	0.2900	0.530	1.20	0.5653	0.4394	0.777
0.03	0.6136	0.3305	0.539	1.30	0.5250	0.4123	0.785
0.04	0.6628	0.3621	0.546	1.40	0.4867	0.3859	0.793
0.05	0.7016	0.3881	0.553	1.50	0.4506	0.3604	0.800
0.06	0.7332	0.4102	0.560	1.60	0.4167	0.3359	0.806
0.07	0.7597	0.4295	0.565	1.70	0.3849	0.3125	0.812
0.08	0.7822	0.4465	0.571	1.80	0.3551	0.2904	0.818
0.09	0.8015	0.4617	0.576	1.90	0.3274	0.2694	0.823
0.10	0.8182	0.4753	0.581	2.00	0.3016	0.2502	0.829
0.12	0.8454	0.4988	0.590	2.50	0.1981	0.1682	0.849
0.14	0.8662	0.5184	0.598	3.00	0.1286	0.1112	0.865
0.16	0.8822	0.5348	0.606	3.50	0.0827	0.07256	0.877
0.18	0.8943	0.5486	0.613	4.00	0.0528	0.04692	0.888
0.20	0.9034	0.5604	0.620	4.50	0.0336	0.03012	0.897
0.22	0.9099	0.5703	0.627	5.00	0.0213	0.01922	0.904
0.24	0.9143	0.5786	0.633	5.50	0.0134	0.01221	0.910
0.26	0.9169	0.5855	0.639	6.00	0.00837	0.00773	0.916
0.28	0.9179	0.5913	0.644	6.50	0.00530	0.00487	0.920
0.30	0.9177	0.5960	0.649	7.00	0.00332	0.00306	0.923
0.40	0.9019	0.6069	0.673	7.50	0.002076	0.00192	0.926
0.50	0.8708	0.6030	0.692	8.00	0.001298	0.00120	0.927
0.60	0.8315	0.5897	0.709	8.50	0.000812	0.000752	0.926
0.70	0.7879	0.5703	0.724	9.00	0.000507	0.000469	0.924
0.80	0.7424	0.5471	0.737	9.50	0.0003177	0.0002920	0.919
0.90	0.6966	0.5214	0.749	10.0	0.0001992	0.0001816	0.912

$K_{5/3}$ and $K_{2/3}$ are modified Bessel functions of (fractional) order (see Abramowitz and Stegun 1964, Chaps. 9 and 10), $P_\parallel$ and $P_\perp$ is the radiative spectral power density for linear polarization.

For the limiting cases of small and large arguments approximate expressions can be given for both $F(x)$ and $G(x)$:

$$F(x) = \frac{4\pi}{\sqrt{3}\Gamma(\frac{1}{3})} \left(\frac{x}{2}\right)^{1/3}, \ x \ll 1 \tag{10.78}$$

$$G(x) = \frac{2\pi}{\sqrt{3}\Gamma(\frac{1}{3})} \left(\frac{x}{2}\right)^{1/3}, \ x \ll 1 \tag{10.79}$$

and

$$F(x) = \sqrt{\frac{\pi}{2}}\, e^{-x}\sqrt{x} = G(x), \ x \gg 1. \tag{10.80}$$

Thus both functions vary like $x^{1/3}$ for small x while they decrease as e^{-x} for large x.

$F(x)$ attains its maximum value of 0.9180 at $x = 0.2858$, while the maximum 0.6070 of $G(x)$ is at $x = 0.4169$.

It is rather difficult to compute numerical values for F and G at intermediate values because this requires extensive use of asymptotic expansion, and it is rather cumbersome to keep track of the precision. The values given in the literature are probably all based on a tabulation given by Oort and Walraven (1956), since they all contain identical inaccuracies. Table 10.1 has been computed independently and should be correct to better than 3 or 4 decimal places.

The linear polarization of the radiation averaged over all directions is remarkably high. This is given by

$$p = \frac{P_\perp - P_\parallel}{P_\perp + P_\parallel} = \frac{G(x)}{F(x)}. \tag{10.81}$$

This varies between 0.5 for $x = 0$ and 1.0 for $x \to \infty$.

10.9   The Spectral Distribution of Synchrotron Radiation from an Ensemble of Electrons

Equation (10.70) gives the spectral density distribution for the emission of a single electron. Obviously the emission of N electrons, all with identical velocity and pitch angle is N-fold this value. But in nature any situation is rarely so simple: the relativistic electrons move with velocities that vary over a wide range and they move in a wide variety of directions. In addition, the magnetic field is frequently inhomogeneous and tangled.

Quite generally the volume emissivity (power per unit frequency interval per unit volume and per unit solid angle) of the relativistic electrons is given by

$$\varepsilon(\nu) = \int_E P(\nu, E) N(E) \, dE, \tag{10.82}$$

Fig. 10.9 The energy
spectrum of cosmic ray
electrons [from Meyer
(1969)]
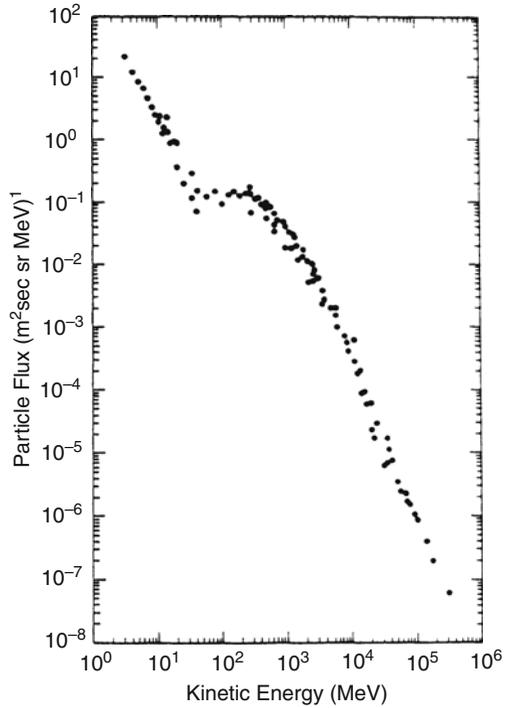
where $P(\nu, E)$ is the total power that one electron with the energy E radiates according to (10.70) and $N(E)\,\mathrm{d}E$ is the number of electrons per unit volume and per unit solid angle moving in the direction of the observer and whose energies lie in the range $E, E + \mathrm{d}E$. From empirical evidence, cosmic ray data (Fig. 10.9) show that $N(E)$ is well described by a power law spectrum. But very little is known about the pitch angle distribution. For many situations of astrophysical interest we assume this distribution to be homogeneous and isotropic, so that

$$N(E)\,\mathrm{d}E = KE^{-\delta}\,\mathrm{d}E \qquad \text{for} \quad E_1 < E < E_2 \tag{10.83}$$

applies independently of the direction of electron motion.

In expression (10.70) for the total radiated power of a relativistic electron in a magnetic field, E (or γ) appears only through ν_c. This alone is sufficient to derive an important result concerning the shape of the synchrotron spectrum emitted by electrons with a power law distribution of their energies. Introducing (10.83) into (10.82) and using (10.63) and (10.77) we find

$$\varepsilon(\nu) = \int_{E_1}^{E_2} P(\nu/\nu_c) E^{-\delta}\,\mathrm{d}E$$

and

$$v_c = \mu E^2 ,$$

(10.84)

where

$$\mu = \frac{3}{4\pi} \frac{eB\sin\alpha}{m^3 c^5} .$$

(10.85)

Substituting

$$x = \frac{v}{\mu E^2} , \qquad dE = -\frac{1}{2\sqrt{\mu}} v^{1/2} x^{-3/2} dx$$

(10.86)

we obtain

$$\varepsilon(v) \propto v^{(1-\delta)/2} \int_{x_1}^{x_2} P(x) x^{(\delta-3)/2} dx ,$$

(10.87)

where

$$v_1 = \mu E_1 , \qquad v_2 = \mu E_2 .$$

(10.88)

If we now put

$$\boxed{n = \tfrac{1}{2}(\delta - 1)}$$

(10.89)

or equivalently

$$\boxed{\delta = 2n + 1}$$

(10.90)

(10.87) can be written as

$$\varepsilon(v) \propto v^{-n} \int_{v/v_2}^{v/v_1} x^{n-1} P(x) \, dx .$$

(10.91)

$$\boxed{\varepsilon(v) \propto v^{-n} \left[Q(n, v/v_1) - Q(n, v/v_2) \right]} \quad ,$$

(10.92)

where

$$\boxed{Q(n, v/v_c) = \int_{v/v_c}^{\infty} x^{n-1} F(x) \, dx} \quad .$$

(10.93)

For $v/v_2 \ll x \ll v/v_1$, Q is independent of the frequency and

$$\boxed{\varepsilon \propto v^{-n}}$$

(10.94)

independently of the pitch angle distribution. For $v > v_2$ the synchrotron spectrum should fall off exponentially, while for $v < v_1$ the spectral index should approach that of the low-frequency approximation (10.78) of $F(x)$, viz. $n = 1/3$. Therefore it should be possible to investigate the shape of the distribution function of relativistic electrons (10.83) from the shape of their synchrotron spectrum. The spectral index

δ of the power-law distribution of the electrons is related to n through (10.89), and the energy limits E_1 and E_2 are given by v_1 and v_2 of (10.77), and they therefore should be observable.

Practical experience with this technique has, however, been rather inconclusive. Bona fide high-frequency breaks in the spectrum that can be interpreted as high-energy cut-offs in the energy distribution function of the relativistic electrons are not observed, and the interpretation of low-frequency breaks is unclear because of optical depth effects either due to absorption in the foreground thermal interstellar plasma or due to synchrotron self-absorption. Another discussion of these problems is presented in Sect. 10.11 and again in Chap. 11, in connection with supernovae.

While the spectral shape of the synchrotron emission is fairly independent of the strength and geometry of the magnetic field in the region, where the radiation is emitted, both the intensity and the state of polarization depend strongly on these quantities. It is rather difficult to obtain general statements valid for all configurations of the magnetic field and pitch angle distributions of the relativistic electrons. Thus we will discuss only two limiting cases: (1) a homogeneous magnetic field with a uniform direction, and (2) a field with completely random orientation. The energy limits E_1 and E_2 where (10.83) ceases to be valid are assumed to be well outside the accessible regions, so that (10.94) with (10.89) applies.

10.9.1 Homogeneous Magnetic Field

Let us assume that the synchrotron emission arises from a region in which the magnetic field is uniform in strength and orientation and which extends for the depth L along the line of sight. Introducing (10.70) into (10.82) we obtain for the total intensity of the emission provided the optical depth is small (Lang 1974)

$$\varepsilon(v) = a(n) \, K \, \frac{\sqrt{3}}{8\pi} \frac{e^3}{mc^2} \left[\frac{3e}{4\pi m^3 c^5} \right]^n (B\sin\alpha)^{n+1} v^{-n} \quad , \qquad (10.95)$$

where

$$a(n) = 2^{n-1} \frac{n+5/3}{n+1} \Gamma\left(\frac{3n+1}{6}\right) \Gamma\left(\frac{3n+5}{6}\right) . \qquad (10.96)$$

Numerical values for $a(n)$ are given in Table 10.2. Inserting CGS values for the physical constants, this becomes

$$I(v) = 0.933 \, a(n) \, K L B_{\perp}^{n+1} \left(\frac{6.26 \times 10^9}{v/\text{GHz}} \right)^n \text{ Jy rad}^{-2} \quad . \qquad (10.97)$$

In the preceding we had shown that the average radiation of monochromatic relativistic electrons is linearly polarized provided the magnetic field is homogeneous.

Table 10.2 Weighting factors for synchrotron emission of a power-law distribution of relativistic electrons

n	δ	$a(n)$	$b(n)$	$p(n)$	n	δ	$a(n)$	$b(n)$	$p(n)$
0.0	1.0	5.236	0.2834	0.600	3.0	7.0	5.016	0.1844	0.857
0.2	1.4	3.156	0.1646	0.643	3.2	7.4	6.240	0.2253	0.863
0.4	1.8	2.320	0.1169	0.677	3.4	7.8	7.866	0.2792	0.868
0.6	2.2	1.913	0.0933	0.706	3.6	8.2	10.039	0.3505	0.873
0.8	2.6	1.705	0.0806	0.730	3.8	8.6	12.963	0.4454	0.878
1.0	3.0	1.612	0.0741	0.750	4.0	9.0	16.923	0.5725	0.882
1.2	3.4	1.596	0.0714	0.767	4.2	9.4	22.325	0.7440	0.886
1.4	3.8	1.642	0.0715	0.783	4.4	9.8	29.744	0.9768	0.890
1.6	4.2	1.744	0.0741	0.796	4.6	10.2	40.003	1.2951	0.894
1.8	4.6	1.905	0.0791	0.808	4.8	10.6	54.288	1.7335	0.897
2.0	5.0	2.133	0.0866	0.818	5.0	11.0	74.310	2.3411	0.900
2.2	5.4	2.441	0.0970	0.828					
2.4	5.8	2.849	0.1109	0.836					
2.6	6.2	3.385	0.1292	0.844					
2.8	6.6	4.089	0.1531	0.851					

The degree of linear polarization will, however, be dependent on frequency. It is remarkable that the radiation emitted by an ensemble of relativistic electrons with a power law distribution of energies still shows linear polarization.

Introducing (10.83) into (10.82) we obtain

$$p = \frac{\int G(x)\,\gamma^{-\delta}\,d\gamma}{\int F(x)\,\gamma^{-\delta}\,d\gamma}.$$

Expressing γ in terms of x, and introducing n through (10.89) this becomes

$$p = \frac{\int G(x)\,x^{-(n+2)}\,dx}{\int F(x)\,x^{-(n+2)}\,dx}. \tag{10.98}$$

Since

$$\int_0^\infty x^\mu F(x)\,dx = \frac{2^{\mu+1}}{\mu+2}\,\Gamma\left(\frac{\mu}{2}+\frac{7}{3}\right)\Gamma\left(\frac{\mu}{2}+\frac{2}{3}\right), \tag{10.99}$$

$$\int_0^\infty x^\mu G(x)\,dx = 2^\mu\,\Gamma\left(\frac{\mu}{2}+\frac{4}{3}\right)\Gamma\left(\frac{\mu}{2}+\frac{2}{3}\right), \tag{10.100}$$

and using the functional equation $\Gamma(x+1) = x\Gamma(x)$, (10.98) can be simplified to

$$p = \frac{n+1}{n+5/3} \qquad . \tag{10.101}$$

Note that the resulting degree of linear polarization is independent of the frequency, and has the *remarkably high value* of 72 % for $n = 0.75$.

10.9.2 Random Magnetic Field

It is only the assumption of a homogeneous and regular magnetic field distribution that leads to this high degree of linear polarization. Adopting a field with random orientation (10.95) must be averaged correspondingly. Using the relation

$$\frac{1}{2} \int_0^{\pi/2} \sin^{n+1}\alpha \, \sin\alpha \, d\alpha = \frac{\pi}{2} \frac{\Gamma\left(\dfrac{n+3}{2}\right)}{\Gamma\left(\dfrac{n+4}{2}\right)} \tag{10.102}$$

we obtain for the average emissivity of the radiation

$$\varepsilon(v) = b(n)\,K\,\frac{e^3}{mc^2}\left[\frac{3e}{4\pi m^3 c^5}\right]^n B^{n+1}\,v^{-n} \quad , \tag{10.103}$$

where $b(n)$ is another function of the (observed) spectral index n

$$b(n) = 2^{n-4}\sqrt{\frac{3}{\pi}}\,\frac{\Gamma\left(\dfrac{3n+1}{6}\right)\Gamma\left(\dfrac{3n+11}{6}\right)\Gamma\left(\dfrac{n+3}{2}\right)}{(n+1)\Gamma\left(\dfrac{n+4}{2}\right)} \tag{10.104}$$

The polarization of the radiation field vanishes. In terms of CGS units (10.103) becomes

$$I(v) = 13.5\,b(n)\,K\,L\,B^{n+1}\left(\frac{6.26\times 10^9}{v/\text{GHz}}\right)^n \; \text{Jy rad}^{-2} \tag{10.105}$$

Therefore, if the observations show partial linear polarization for the radiation, we can be certain that the magnetic field must be uniform in the region where the synchrotron radiation is emitted. Observations show that such regions can extend over ranges of the order of kpc; this is the case both for our own Galaxy and for distant extragalactic systems.

10.10   Energy Requirements of Synchrotron Sources

When physical models for nonthermal radio sources are proposed, one of the first considerations must be the energetics. The total energy content of the source can only be estimated when the physical mechanism that provide their power is known. For nonthermal sources, thermal radiation is usually not seen in the radio range, but may be in the infrared. Thus the thermal energy content is not well determined from radio data alone.

The energy output of nonthermal sources in the radio range is mainly in the form of synchrotron emission, and therefore the energetics of this process are of main interest here. Energy must be present in two different forms for such emission to occur: the kinetic energy of the relativistic particles, W_{part}, and the energy stored in the magnetic field, W_{mag}. We will now estimate how much of these must be available in order to provide the observed flux density S_ν.

If V is the volume of the source then the total energy content that is available for the synchrotron mechanism is, for $\mu = \varepsilon = 1$

$$W_{tot} = W_{part} + W_{mag} = V \left(u_p + u_{mag} \right) . \tag{10.106}$$

Here u_p is the energy density of the relativistic particles, that is, of the electrons and protons (and ions with $Z > 1$). Since protons emit only very little synchrotron radiation compared to electrons of the same kinetic energy E, very little is known about their energy density, and it is customary to assume that

$$u_p = \eta \, u_e , \tag{10.107}$$

where u_e is the energy density of the electrons and $\eta > 1$ is a factor taking all other particles into account. Now if we again assume a power law (10.83) for the energy spectrum of the electrons, the energy density of the synchrotron source in the form of particles will be

$$u_p = \eta K \int_{E_{min}}^{E_{max}} E^{1-\delta} \, dE , \tag{10.108}$$

$$u_{mag} = \frac{1}{8\pi} B^2 . \tag{10.109}$$

An electron with a kinetic energy E will radiate over a wide range of frequencies with a spectral density given by (10.72, 10.73, 10.74, 10.75, 10.76, 10.77), but since radiated power peaks strongly close to ν_c, no large error will be made if we substitute a frequency ν for E in (10.108). From (10.77) and (10.63), there is

$$\nu = \frac{3}{2} \frac{eB}{m^3 c^5} E^2 , \tag{10.110}$$

so that

$$u_p = K \times G \times B^{n-1/2} \tag{10.111}$$

where

$$G = \frac{\eta}{1 - 2n} \left(\frac{e}{m^3 c^5} \right)^{n-1/2} \left(v_{\text{max}}^{1/2-n} - v_{\text{min}}^{1/2-n} \right) \tag{10.112}$$

and

$$W_{\text{tot}} = V \times \left(K \times G \times B^{n-1/2} + \frac{1}{8\pi} B^2 \right) . \tag{10.113}$$

In these expressions K and B are independent variables and the total energy content of the source can vary within wide limits depending on which values for K and B are chosen. We will restrict this wide range of values by making use of the fact that the same electrons that contribute to the energy content also radiate. Their emissivity is given by (10.95) or (10.103); which of these expression is used depends on the large-scale morphology of the magnetic field in the source. But since only order-of-magnitude estimates are used here, the numerical differences of these two formulae are not very critical.

Assuming a spherical shape for the synchrotron source with a volume V, and a source distance R, then the observed flux density will be

$$S_v = KH \frac{V}{R^2} B^{n+1} v^{-n} \tag{10.114}$$

with

$$H = b(n) \frac{e^3}{mc^2} \left(\frac{3e}{4\pi m^3 c^5} \right)^n . \tag{10.115}$$

Using (10.114) K can be eliminated from (10.113) resulting in

$$W_{\text{tot}} = \frac{G}{H} R^2 \left(S_v v^n \right) B^{-3/2} + \frac{V}{8\pi} B^2 . \tag{10.116}$$

Provided that both the distance, R, and the source volume, V, are known, B is the only unknown quantity. Then W_{tot} attains the minimum value for

$$B_{\text{eq}} = \left(6\pi \frac{G}{H} \frac{R^2}{V} S_v v^n \right)^{2/7} , \tag{10.117}$$

and for this magnetic field strength we find

$$\frac{u_p}{u_{\text{mag}}} = \frac{4}{3} , \tag{10.118}$$

so that

$$W_{\text{tot}} = \frac{7}{4} (6\pi)^{-3/7} \left(\frac{G v^n}{H} S_v \right)^{4/7} R^{8/7} V^{3/7} . \tag{10.119}$$

This equation is normally used in high-energy astrophysics to estimate the minimum energy requirements of a synchrotron source.

10.11   Low-Energy Cut-Offs in Nonthermal Sources

The interpretation of the spectrum of nonthermal radio sources as synchrotron emission of relativistic electrons and estimates of the distribution function of the electrons, their energy range, cut-off, etc. as described here is incomplete, since other effects have not been considered. In the low-frequency range, in particular, there are several mechanisms that will affect the observed spectral index of the radiation, and it is difficult to decide for a particular source which of the possible cases does applies.

Synchrotron radiation at frequencies below the low-frequency cut-off ν_1 (cf. (10.88)) should have a spectral index of $n = 1/3$ according to the low-frequency asymptotic form (10.78) of $F(x)$. This interpretation is, however, far from unique, so other interpretations are possible.

In synchrotron radiation fields spontaneous photon emission will be accompanied by absorption and stimulated emission as in any other radiation field. This absorption can become important in compact, high-intensity radio sources at low frequencies when the optical depth becomes large. The spectral distribution of such optically thick sources is $n = -5/2$ (cf. Scheuer 1967, Rybicki and Lightman 1979, p. 190). That this index is different from the value $n = -2$ (the Rayleigh-Jeans value) is an indication of the nonthermal quality of the synchrotron radiation. This is also indicated by the fact that such sources can still show linear polarization if their magnetic field is homogeneous and smooth.

Another effect that could cause a low-frequency cut-off in the synchrotron spectrum has been discussed by Razin (c.f. Rybicki and Lightman (1979) p. 234). This effect is based on the fact that the interstellar medium in a radio source is a plasma. If the frequency of the radiation field comes close to the plasma frequency, the refractive index is

$$n_{\mathrm{P}} = \sqrt{1 - \frac{\nu_{\mathrm{P}}^2}{\nu^2}}\,,$$

so that the speed of light is c/n. Therefore the relativistic beaming, which is needed if frequencies higher than the local cyclotron frequency ν_{G} are to be reached, is not effective. The synchrotron emission, therefore, will be reduced below this frequency. The Razin effect is therefore not absorption, but rather a failure of synchrotron emission.

Finally, there might be a foreground thermal plasma which will absorb the synchrotron emission at lower frequencies. In conclusion, the interpretation of the low-frequency range of the spectrum of nonthermal sources is far from simple and unique. Interpretations may be obtained by combining information obtained from high-resolution aperture synthesis observations, spectral data and variability studies.

10.12 Inverse Compton Scattering

10.12.1 The Sunyaev-Zeldovich Effect

We can have a situation in which photons from a cold source, the 2.7 K background, interact with a hot foreground source, a cluster of galaxies. Such clusters have free electrons with $T_k > 10^7$ K; so the bremsstrahlung radiation peaks in the X-ray range. The net effect of an interaction of the photons and electrons is to shift longer wavelength photons to shorter wavelengths.

Quantitatively, this scattering can be analyzed using the approach described at the beginning of Sect. 10.7, if we use for the acceleration

$$a = \frac{e}{m_0} E \,,$$

where E is the electric field. The Lorentz transformation for the electron is

$$E' = \gamma \left(E + \frac{v}{c} \times B \right) , \tag{10.120}$$

so that the radiation in the zeroth-order rest frame of the electron is

$$P = \frac{2e^2}{3c^3} a_\perp^2 = \frac{2e^4}{3m^2 c^3} (E')^2 . \tag{10.121}$$

Transforming this to the laboratory rest frame,

$$P = 2\frac{8\pi}{3} \left(\frac{e^4}{m_0^2 c^4} \right) \gamma^2 \frac{\left(E + \frac{v}{c} \times B \right)^2 c}{8\pi}$$
$$= \sigma_T \gamma^2 c\, u_{\text{photons}} = 1.99 \times 10^{-14} \gamma^2 c\, u_{\text{photons}} \tag{10.122}$$

where

$$\sigma_T = \frac{8\pi}{3} \left(\frac{e^2}{mc^2} \right)^2 = 6.65 \times 10^{-25} \, \text{cm}^2 \tag{10.123}$$

is the Thomson electron cross section, and u_{photon} is the laboratory rest-frame photon energy density. The power radiated is manifested as an apparent absorption, and the energy lost by the electrons is the energy gained by the photons, namely γ^2. Then

$$\left(\frac{E_{\text{after}} - E_{\text{before}}}{E_{\text{before}}} \right) = (\gamma^2 - 1) \cong \frac{v^2}{c^2} \cong \frac{3kT}{m_0 c^2} . \tag{10.124}$$

The total effect is a loss of photons at wavelengths longer than 1.6 mm (the peak of the 3 K background). The fractional decrease in the background temperature is the product of the Thomson cross section times the column density of the electrons times the energy loss:

$$\frac{\Delta T}{T} = \frac{1}{2} \frac{3kT}{m_0 c^2} \, \sigma_T L N_e \,. \tag{10.125}$$

There is an additional factor of $\frac{4}{3}$ from a detailed analysis using the Kompaneets equation (see, e.g. Rybicki and Lightman (1979) p. 213). Thus, the final result, referred to as the Sunyaev-Zeldovich effect, is:

$$\frac{\Delta T}{T} = \frac{2kT_e}{m_e c^2} \, \sigma_T N_e L = 2.24 \times 10^{-34} \, T_e \, N_e \, L \,. \tag{10.126}$$

The X-ray bremsstrahlung depends on $N_e^2 L$. When combining maps of the X-ray emission and the Sunyaev-Zeldovich or S-Z absorption, these effects can be used to estimate distances independently of assumptions about any red-shift distance relation. Thus, one could obtain another estimate of the Hubble constant. We give an example in the next chapter.

10.12.2   Energy Loss from High-Brightness Sources

We can also use a simple form of the inverse Compton effect to estimate limits on brightness temperature in high-brightness nonthermal sources. From (10.124) we have, for the energy loss of electrons

$$-P = \frac{dE}{dt} = \sigma_T \, \gamma^2 \, c^2 \, u_{ph} = (\gamma m_e c^2) \left(\frac{\gamma e \, \sigma_T}{m_e c^2} \right) u_{ph} \,, \tag{10.127}$$

where m_e is the electron rest mass. Then

$$-\frac{dE}{dt} = E \frac{\sigma_T \gamma}{m_e c} u_{ph} \,, \tag{10.128}$$

$$\frac{1}{E} \frac{dE}{dt} = 2.4 \times 10^{-8} \, \gamma u_{ph} \,. \tag{10.129}$$

We have assumed here that the photon energy is much less than that of the relativistic electrons. This allows us to avoid a more complex expression for the scattering. The expression in (10.129) gives a characteristic time for energy losses. The quantity γ is a free parameter, but the highest observed frequency of synchrotron emission can be combined with (10.77) and a value for $v_G = 17B$, where B is in Gauss and v in MHz, we have a limit on electron lifetimes. This expression should be compared to (10.64), the energy loss for synchrotron emission. Taking the ratio we find

$$\frac{EL_{Compton}}{EL_{synchrotron}} = \frac{u_{ph}}{u_B} \,. \tag{10.130}$$

That is, the ratio of the energy in the photon field to that in the magnetic field. If these are equal, Compton losses dominate.

Problems

1. Suppose an object of radius $100\,\text{m}$, with a uniform surface temperature of $100\,\text{K}$ passes within 0.01 AU of the earth (an astronomical unit, AU, is $1.46 \times 10^{13}\,\text{cm}$).
(a) What is the flux density of this object at $1.3\,\text{mm}$?
(b) Suppose this object is observed with a $30\,\text{m}$ telescope, at $1.3\,\text{mm}$, with a beam-size of $12''$. Assume that the object has a Gaussian shape; calculate the peak brightness temperature by considering the dilution of the object in the telescope beam. Neglect the absorption by the earth's atmosphere.
(c) This telescope is equipped with a bolometer with $\text{NEP} = 10^{-15}\,\text{W\,Hz}^{-1/2}$ and bandwidth $20\,\text{GHz}$; how long must one integrate to detect this object with a 5 to 1 signal-to-noise ratio, if the beam efficiency is 0.5, and the earth's atmospheric optical depth can be neglected?

2. The Orion hot core is a molecular source with an average temperature of $160\,\text{K}$, angular size $10''$, located $500\,\text{pc}$ ($= 1.5 \times 10^{21}\,\text{cm}$) from the Sun. The average local density of H_2 is $10^7\,\text{cm}^{-3}$.
(a) Calculate the line-of-sight depth of this region in pc, if this is taken to be the diameter.
(b) Calculate the column density, $N(H_2)$, which is the integral of density along the line of sight. Assume that the region is uniform.
(c) Obtain the flux density at $1.3\,\text{mm}$ using $T_{\text{dust}} = 160\,\text{K}$, the parameter $b = 1.9$, and solar metallicity ($Z = Z_{\odot}$) in Eq. 10.7.
(d) Use the Rayleigh–Jeans relation to obtain the dust continuum main beam brightness temperature from this flux density, in a $10''$ beam. Show that this is much smaller than T_{dust}.
(e) At long millimeter wavelengths, a number of observations have shown that the optical depth of such radiation is small. Then the observed temperature is $T = T_{\text{dust}}\,\tau_{\text{dust}}$, where the quantities on the right hand side of this equation are the dust temperature and dust optical depth. From this relation, determine τ_{dust}.
(f) At what wavelength is $\tau_{\text{dust}} = 1$ if $\tau_{\text{dust}} \sim \lambda^{-4}$?

3. (a) From Fig. 10.1, determine the "turnover" frequency of the Orion A HII region, that is the frequency at which the flux density stops rising, and starts to decrease. This can be obtained by noting the frequency at which the linear extrapolation of the high and low frequency parts of the plot of flux density versus frequency meet. At this point, the optical depth, τ_{ff}, of free–free emission through the center of Orion A, is unity, that is $\tau_{\text{ff}} = 1$. Call this frequency ν_0.
(b) From Eq. (10.35), there is a relation of turnover frequency, electron temperature, T_{e}, and emission measure. This relation applies to a uniform density, uniform temperature region; *actual HII regions have gradients in both quantities*, so this relation is at best only a first approximation. Determine EM for an electron temperature $T_{\text{e}} = 8300\,\text{K}$.
(c) The FWHP size of Orion A is $2.5'$, and Orion A is $500\,\text{pc}$ from the Sun. What is the linear diameter for the FWHP size? Combine the FWHP size and emission measure to obtain the RMS electron density.

4. A more accurate method to obtain the emission measure of the high electron density core of an HII region such as Orion A is to use $T_B = T_e \tau_{ff}$, where T_B is the brightness temperature of the source corrected for beam dilution.

(a) Use the T_e and source FWHP size values given in the last problem. For $v = 23\,\text{GHz}$, take the main beam brightness temperature, $T_{MB} = 24\,\text{K}$, and the FWHP beamsize as $43''$. Correct the main beam brightness temperature, T_{MB}, for source size to obtain T_B.

(b) Determine τ_{ff}.

(c) Use Eq. (10.35) with $a = 1$ to find v_0 and EM; compare these results to those obtained in the last problem. Discuss the differences. Which method is better for determining the EM value for the core of an HII region at high frequencies?

5(a). For frequencies above $2\,\text{GHz}$, the optical depth of Orion A is small (i.e., the source is optically thin) and τ_{ff} varies as $v^{-2.1}$. Calculate τ_{ff} at $5\,\text{GHz}$, $10\,\text{GHz}$, $23\,\text{GHz}$, $90\,\text{GHz}$, $150\,\text{GHz}$ and $230\,\text{GHz}$.

(b) Next calculate the peak brightness temperature, at the same frequencies, for a telescope beam much smaller than the FWHP source size. Use the expression $T_B = T_e \tau_{ff}$.

(c) With the IRAM radio telescope of $30\,\text{m}$ diameter, one has a FWHP beamwidth in arc seconds of $\theta_b = 2700/v$, where v is measured in GHz. Calculate the main beam brightness temperature at the frequencies given in part **(a)**.

6. (a) Given the characteristics of the sources Orion A (from the last two problems) and Orion hot core (Problem 2), at what frequency will the continuum temperatures of these sources be equal when measured with the $30\,\text{m}$ telescope?

(b) Repeat this calculation for the Heinrich Hertz sub-millimeter telescope, of $10\,\text{m}$ diameter, where now the FWHP beamwidth is $\theta = (8100/v)$ for v measured in GHz. Will T_{dust} equal T_{ff} at a higher or lower frequency?

7. (a) The HII region W3(OH) is $1.88\,\text{kpc}$ from the Sun, has a FWHP size of $2''$ and a turn over frequency of $23\,\text{GHz}$. Determine the RMS electron density if the $T_e = 8500\,\text{K}$. Determine the mass of ionized gas.

(b) There is a molecular cloud of size $2''$ located 7 arcsec East of W3(OH). The dust column density of order $10^{24}\,\text{cm}^{-2}$, with $T_{dust} = 100\,\text{K}$. Given these characteristics, at what frequency will the continuum temperatures of these sources be equal when measured with the $30\,\text{m}$ telescope?

(c) Repeat this calculation for the Heinrich Hertz sub-millimeter telescope. Will T_{dust} equal T_{ff} at a higher or lower frequency?

8. The Sunyaev–Zeldovich (S-Z) effect can be understood in a qualitative sense by considering the interaction of photons in the $2.73\,\text{K}$ black body distribution with much more energetic electrons, with an energy of $5\,\text{keV}$ and density of $\sim 10^{-2}\,\text{cm}^{-3}$.

(a) What is the energy of photons with a wavelength of $1.6\,\text{mm}$ (the peak of the background distribution)? Compare to the energy of the electrons.

(b) Obtain the number of $2.73\,\text{K}$ photons per cm^3 from Problem 20c of Chap. 1.

(c) Assume that the interaction of the $2.73\,\text{K}$ black body photons with the electrons (assumed monoenergetic) in the cluster will lead to the equipartition of energy.

Make a qualitative argument that this interaction leads to a net increase in the energy of the photons. Justify why there is a *decrease* in the temperature of the photon distribution for wavelengths longer than 1.6 mm and an *increase* shorter than this wavelength.

9. The source Cassiopeia A is a cloud of ionized gas associated with the remnant of a star which exploded about 330 years ago. The radio emission has the relation of flux density as a function of frequency shown in Fig. 10.1 in "Tools". For the sake of simplicity, assume that the source has a constant temperature and density, in the shape of a ring, with thickness $1'$ and outer radius of angular size $5.5'$. What is the actual brightness temperature at 100 MHz, 1 GHz, 10 GHz, 100 GHz?

10. Obtain the integrated power and spectral index for synchrotron radiation from an ensemble of electrons which have a distribution $N(E) = N_0$, that is a constant energy distribution from E_{min} to E_{max}.

Chapter 11
Some Examples of Thermal and Nonthermal Radio Sources

As mentioned previously, radio sources can be divided into two classes: thermal and nonthermal. Here we will give examples of each class and discuss the physics involved in some detail.

As an example of a thermal source, we will first describe the emission from the quiet sun. As a second, perhaps more important type of thermal source we discuss galactic H II regions and ionized outflows. H II regions will be discussed later in connection with recombination line radiation. For nonthermal sources, we use supernova remnants as one example. We will present a sketch of their time evolution without discussing the energy sources driving the explosion and the mechanisms producing the highly relativistic electrons and magnetic fields. As additional examples, we consider the radio emission from Cygnus A and an example of the Sunyaev-Zeldovich effect. Finally we discuss some considerations of the time variability of flux density of some sources.

11.1 The Quiet Sun

First attempts to identify the radio emission of the sun were made soon after the detection of radio waves by H. Hertz. Kennelly, one of Edison's coworkers, mentioned in a letter that such investigations took place in 1890. However, these attempts, like those of Scheiner and Wilsing in Potsdam (1896), Sir Oliver Lodge in England (1900) and of Nordman in France (1902), proved unsuccessful. There are two reasons why these experiments failed. In 1900 Planck derived the spectral distribution of thermal radiation so that the expected intensity of solar radiation at radio wavelengths could be computed. This was found to be much too small for the sensitivity of the then available radio receivers. Second, in 1902 Kennelly and Heavyside concluded, from the possibility of transoceanic radio transmission, that there must be a conducting layer high in the earth's atmosphere that reflects radio waves and that this layer would prevent the reception of solar radio radiation at long wavelengths.

More than 40 years later Southworth (1942) detected thermal radiation of the quiet sun, while in the same year Hey found very intense, time variable solar

radiation whose emission was associated with the sunspot phenomenon. Jansky did not detect the Sun because he observed during the period of minimum solar activity. Observations of the time and spatial distributions of these microwave bursts, and a theoretical explanation in terms of concepts of plasma physics belong to solar radio astronomy proper. This is an extensive and specialized field which will not be covered here; several monographs listed in the list of general references for this chapter give a good review. We will be concerned only with the radio emission of the quiet sun, that is, with the thermal radiation of the sun's corona.

As was known from optical measurements there is an extended atmosphere above the solar photosphere in which the gas temperature increases rapidly with height from about 6000 K in the photosphere to a temperature of several million K. The gas density in the corona is lower than that of the photosphere. Because of the lower density, the corona does not stand out in the visual range; its influence is best seen in observations made at solar eclipses. The opposite is true in the radio range: here the radiation of the hot gas is dominant.

Due to the high temperature in the corona the gas is almost fully ionized. The electron density distribution is described by the Baumbach-Allen formula

$$\frac{N_e}{\text{cm}^{-3}} = \left[1.55 \left(\frac{r}{r_0} \right)^{-6} + 2.99 \left(\frac{r}{r_0} \right)^{-16} \right] \times 10^8 \tag{11.1}$$

and for the electron temperature a constant value of $T_e \cong 10^6$ K can be adopted for $h = r - r_0 > 2 \times 10^9$ cm (Fig. 11.1) where r_0 is the solar radius, 7×10^{10} cm.

The optical depth τ_ν and the resulting brightness temperature of the corona could now be computed using (10.35) and (1.37) starting with $\tau_\nu = 0$ at $h \to \infty$. But for
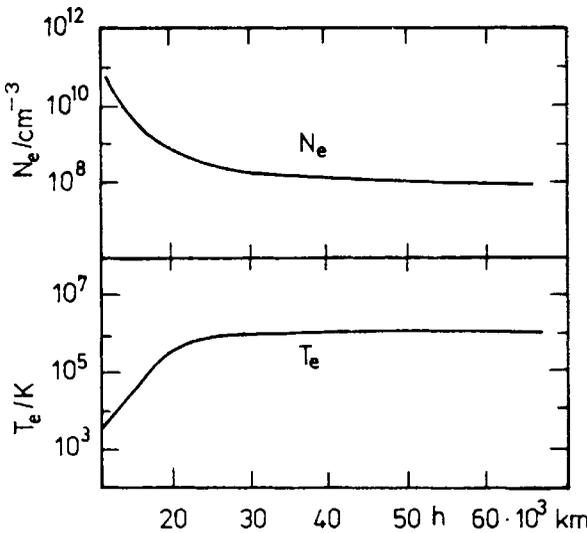
Fig. 11.1 The solar corona after Baumbach and Allen

Fig. 11.2 Refraction in the
solar corona
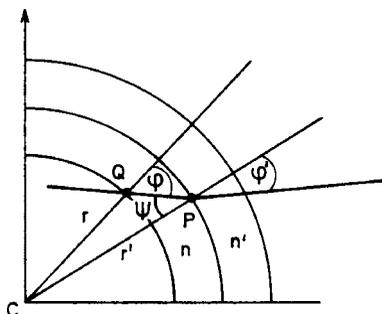
low frequencies not too far away from the plasma frequency (2.77) the index of
refraction (2.81) is significantly less than unity $(n < 1)$, so that refraction effects
must be taken into account in the radio range. The radio waves will propagate along
curved paths. We will now determine the differential equation governing the shape
of these rays.

If the corona is considered to consist of concentric shells of constant index of
refraction n (Fig. 11.2), then the law of refraction gives

$$n' \sin \varphi' = n \sin \psi, \tag{11.2}$$

while in the triangle CPQ

$$\frac{\sin \psi}{r} = \frac{\sin(180° - \varphi)}{r'} = \frac{\sin \varphi}{r'}. \tag{11.3}$$

From this

$$n r \sin \varphi = n' r' \sin \varphi' = \varrho = \text{const}, \tag{11.4}$$

where the parameter ϱ of the ray is equal to the minimum distance from the solar
center that a straight line tangent to the exterior part of the ray would have attained.
The angle $\varphi(r)$ between the ray and the solar radius is closely related to the slope
of the ray, so that (11.4) or the equivalent

$$n(r) r \sin \varphi(r) = \varrho \tag{11.5}$$

is a differential equation for the rays in the solar corona.

If the brightness distribution across the solar disk is to be computed, the equation
of transfer (1.17) has to be solved along a representative sample of rays as shown in
Fig. 11.3. Each ray has a point of closest approach and it is symmetrical around this
point. The radius of this point of closest approach is found from (11.5) by putting
$\varphi = \frac{\pi}{2}$.

The brightness distribution across the solar disk now depends on how the opti-
cal depth is distributed along the ray in relation to this point of closest approach
(Fig. 11.4).

Fig. 11.3 Ray geometry in
the solar atmosphere [after
Jaeger and Westfold (1949)]
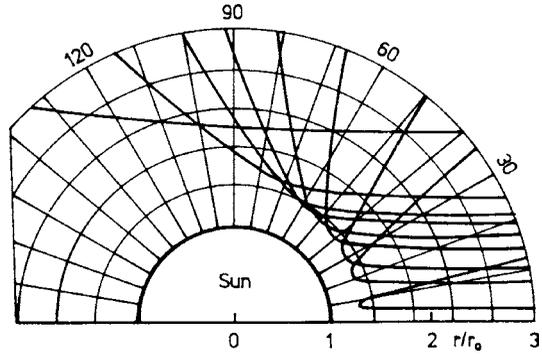
1) For low frequencies $v < 0.1\,\text{GHz}$ ($\lambda > 3\,\text{m}$) the refraction effects in the outer
   parts of the solar corona are so strong that the points of closest approach are
   situated well above those layers where τ_v approaches unity. The solar disk is
   brightest at the center and the brightness decreases smoothly with r reaching
   zero only after several solar radii. Because $\tau_v \ll 1$, the brightness temperature is
   less than the corona temperature.
2) For $0.1\,\text{GHz} < v < 3\,\text{GHz}$, the optical depth reaches unity, $\tau_v \cong 1$, near the point
   of closest approach. Refraction effects can be neglected. The situation can be
   modelled by a luminous atmosphere above a non luminous stellar body, resulting
   in a bright rim.
3) For $v > 3\,\text{GHz}$ conditions approach those met in the optical range. The optical
   thickness produced by the corona is so small that very little radiation is emitted
   by it and there is no bright rim.

The study of thermal radiation from the quiet sun is considered to be rather well
understood. This emission is often negligible compared to the nonthermal, slowly
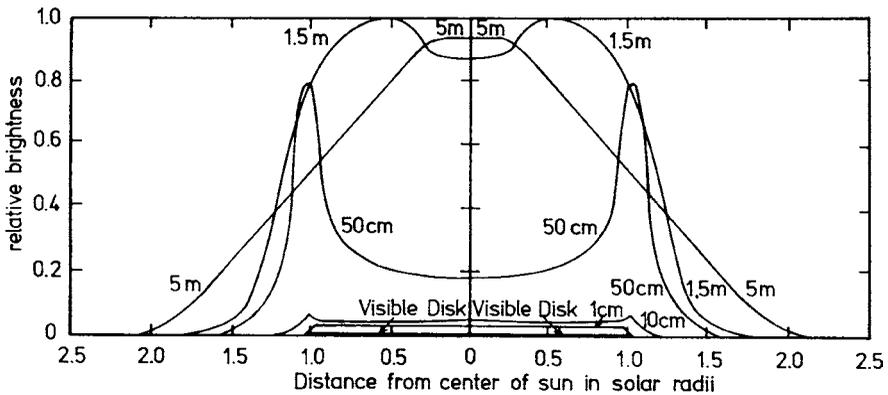
Fig. 11.4 Brightness distribution across the solar disk at different wavelengths [after Smerd
(1950)]

varying component, and much less than the intensity of solar bursts. Active solar events are studied using instruments such as the Nancay and Nobeyama Radio Heliographs. The instruments, observational techniques and models needed to study active regions such as solar flares, coronal mass ejections and non-thermal phenomena differ considerably from those used in the other areas of radio astronomy. Thus this subject has become a separate but important discipline that affects "space weather". A summary of more recent solar radio results is given in McLean and Labrum (1985), Bastian et al. (1998) and Gary and Keller (2004).

11.2   Radio Radiation from H II Regions

11.2.1   Thermal Radiation

The radio radiation from H II regions was discovered later than nonthermal radiation, since the peak intensities are *at most* the electron temperature, T_e. For all H II regions $T_e < 20\,000$ K. In the following, we illustrate the principles given in Sect. 9.4, using data taken for the nearby H II region Orion A, M 42. This well-known, visible source is located about 450 pc from the sun, and is on the front side of a massive molecular cloud. Figure 11.5 contains two radio maps. Both were taken with the 100 m telescope. The lower map was measured at a frequency of 4.8 GHz, or a wavelength 6.2 cm. The beamsize was measured to be 2.45′. In order to recover the detailed structure at the low intensity levels, a variant of CLEAN was applied (see Sect. 9.4.2). The peak intensity in main beam brightness units is 330 K.

The map shown at the top was made at a frequency of 23 GHz, a wavelength of 1.3 cm. The beam size of the 100 m telescope was measured to be 40″. The FWHP size of the core of Orion A is measured to be 2.5′ at 23 GHz. The peak main beam brightness temperature of Orion A at 23 GHz is 24 K. Since the measured size of the core of Orion A is much larger, this is the actual brightness temperature (if small scale structure is not important). At 4.8 GHz, there is some confusion with NGC 1982, to the north, but the size of the core is consistent with the result obtained at 23 GHz. At 4.8 GHz, the size of the beam is comparable to that of the core, so that the observed peak brightness temperature must be corrected for beam size to obtain the actual peak brightness temperature. Then the peak brightness temperature at 4.8 GHz is 650 K. The difference in brightness temperatures at 23 GHz and 4.8 GHz is consistent with (10.35). From (1.17), these results are consistent with the same electron temperature, T_e and small optical depth, τ. From Fig. 11.5, the radio emission of Orion A is far more extended at 4.8 GHz than at 23 GHz. This is not a result of lower receiver sensitivity at 23 GHz, but is caused by the fact that for a given emission measure and T_e, brightness temperatures decrease at shorter wavelengths. This is illustrated in Fig. 11.5. In effect, at shorter wavelengths, the ionized gas becomes optically thin, and we are looking through a more and more transparent medium.
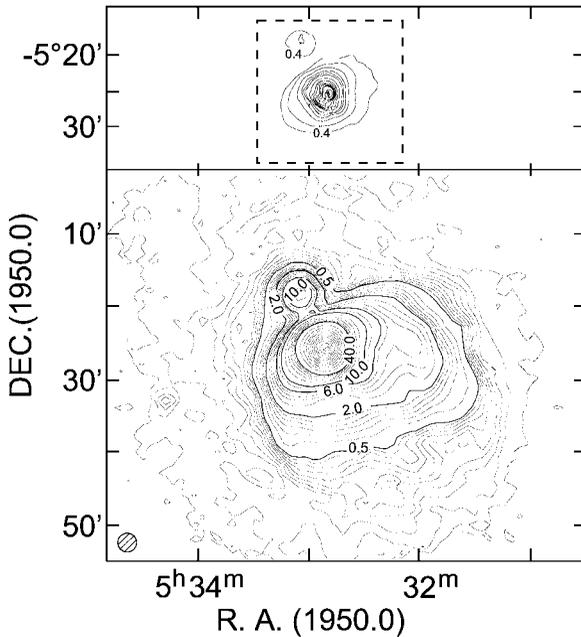
Fig. 11.5 Single-dish radio maps of Orion A, made with the 100 m telescope. In the *upper panel* is the 1.3 cm map (angular resolution 40″). The peak main beam brightness temperature is 24 K (Wilson and Pauls 1984). In the *lower panel*, we show the map made at 6.2 cm (angular resolution 2.45′). The peak main beam brightness temperature is 330 K (Wilson et. al 1997). The angular scale of both maps is identical

From optical and radio spectral line measurements (see Chap. 13), the electron temperature of Orion A is 8 500 K. To obtain the emission measure (EM), we apply (10.36), with $a = 1$, $T_e = 8500$ K, $\tau \ll 1$ at $\nu = 23.0$ GHz to give EM $= 4 \times 10^6$ cm^{-6} pc, averaged over the telescope beam. For a comparison with physical models, and to estimate the mass of the ionized gas, we need the density or column density. For a first approximation based on radio continuum measurements, we can obtain the RMS electron density by dividing the emission measure by the line-of-sight path (A refinement of this method is given later in this section). We first assume that the Gaussian FWHP size of the core, when deconvolved from the beam, is the diameter. In our simple geometrical model, we have *assumed* that the core has a spherical shape. Since the measured diameter is 0.33 pc, this is also the line-of-sight path. Using this value, we obtain an electron density, $\langle N_e \rangle = 3.5 \times 10^3$ cm^{-3}. Small scale structure or nonspherical geometry will give rise to a *larger* value for $\langle N_e \rangle$. The corresponding electron (or proton) column density is 3.4×10^{21} cm^{-2}. Using this value and applying a correction for 36% mass in helium, we find that the mass of the ionized gas in the core of Orion A is 2 solar masses. This is an *upper* limit to the actual value. This mass is about 10% of the mass in the stars in Orion A.

There are three improvements one can apply to the previous analysis. The first two are connected with geometry, and the last with physical conditions. In our estimate of source size, we have tacitly assumed that the diameter of the core is the FWHP deconvolved size. This is true if the source shape is Gaussian. However, our estimate of electron density made use of a spherical uniform density source distribution. For a sphere, there is a general relation between the ratio of the Gaussian circular beam size to the assumed spherical source size and deconvolved spherical source size (Panagia and Walmsley 1978). If the beam is much smaller than the source, the limiting source radius is 0.73 times the deconvolved FWHP size. In the case of the core of Orion A, this is 0.48 pc. The RMS electron density is then $\langle N_e \rangle = 2.9 \times 10^3 \, \text{cm}^{-3}$, and the column density is $4.2 \times 10^{21} \, \text{cm}^{-2}$ and the mass is 5 solar masses.

The second refinement to the calculation of electron density involves relaxing the assumption of spherical geometry. From determinations of electron densities on the basis of optical observations and models of Orion A based on radio recombination lines (see Chap. 13), it has been established that the electron densities in the core of Orion A are $\geq 10^4 \, \text{cm}^{-3}$. To obtain such densities from radio continuum data, the line of sight must be ≤ 0.3 of the measured diameter. This is usually referred to as *clumping*; one could also envision the source as having a structure composed of a series of very thin slabs whose axes are parallel to the line of sight. For a well-studied source such as the core of Orion A, very simple models cannot match the large body of data available, and such complex, multi-layer models are needed.

Finally, in applying (10.36), it was assumed that the H II region is isothermal. This is a fairly good approximation, since both the heating, by photo ionization, and the cooling, by collisions of electrons with ions such as O^{++} and N^{++}, are proportional to N_e^2, and thus T_e is only weakly dependent on density. However, radio recombination line measurements have shown that while the core of Orion A has $T_e = 8\,500 \, \text{K}$, the outer parts have $T_e = 6\,500 \, \text{K}$.

11.2.2   Radio Radiation from Ionized Stellar Winds

The theory for thermal radio emission from ionized stellar winds is a good application of the thermal emission formulas in Chap. 10. Our sketch follows the exposition of Panagia and Felli (1975). The wind from a star is assumed to be ionized. The densities depend on radius r as

$$\rho_{cs} = q\,r^{-2} = \frac{\dot{M}}{4\pi v_w}\,r^{-2}, \tag{11.6}$$

where $\dot{M}$ is the mass loss rate and v_w is the wind velocity. For spherical geometry at distances large compared to the radius we have the result:

$$S_\nu = 8.2 \left(\frac{n_0 r_0^2}{10^{36}} \right)^{4/3} \left(\frac{\nu}{\text{GHz}} \right)^{0.6} \left(\frac{T_e}{10^4 \, \text{K}} \right)^{0.1} \left(\frac{d}{\text{kpc}} \right)^{-2}, \qquad (11.7)$$

where r_0 is the photospheric radius in cm, n_0 is the electron density in cm^{-3}, and S_ν is in units of mJy.

The spectral index is caused by the fact that at different frequencies, larger or smaller portions of the ionized outflow are optically thick. Using the solar data as given in Fig. 11.1, we find that solar-like stars are expected to have a flux density of $< 2 \, \text{mJy}$ at 100 pc. Supergiant stars should be detectable in the radio range at larger distances because of their larger radii. However, at frequencies of $\sim 1 \, \text{GHz}$, some stars radiate by nonthermal processes; this radiation seems to be caused by gyro synchrotron emission in large B fields (for a review of more recent results see Dulk 1985).

11.3 Supernovae and Supernova Remnants

The brightest radio sources at low frequencies ($\nu < 1 \, \text{GHz}$) turn out to be nonthermal. Early radio studies allowed an identification of the source Taurus A with the Crab nebula, the remnant of the supernova explosion observed by Chinese astronomers in 1054. The source Cas A is thought to be the remnant of a supernova that exploded in about 1667 and may have been observed by Flamsteed in spite of interstellar absorption (Fig. 11.6). These early identifications made it plausible that supernova remnants are sources of nonthermal radio emission.

While the Crab nebula radio source Taurus A is an elliptical region completely filled with radio emission, Cassiopeia A is a spherical shell source. In the case of Cas A, the radio emission is confined almost exclusively to a thin shell and observations made at different epochs show that this shell is expanding. Many other nonthermal radio sources were also found to have a similar morphology, although the shell was often incomplete. In some cases, additional evidence was found showing that these sources are indeed remnants of supernova explosions. X-ray emission is quite commonly observed in supernova remnants, and X-ray measurements of SNRs, particularly with the ROSAT satellite, are important inputs for models.

The number of known supernova remnants has increased considerably in the last ten years due to the increasing resolving power and sensitivity of the radio continuum surveys. Usually a procedure consisting of several steps is needed before a source is considered to be a bona-fide supernova remnant. As a first step, one eliminates objects from catalogs of radio sources that have positive spectral indices and thus might be extragalactic. Optical identifications can be used to classify sources away from the galactic plane. Otherwise, where interstellar dust extinction prevents a direct identification, a source of diameter of $< 1'$ might be extragalactic. The remaining sources with negative spectral indices are tentatively SNRs. However, it is possible to confuse filled-center SNRs with H II regions. Before a definite identification is possible, additional evidence is required. This might be high-resolution radio continuum measurements showing a shell structure, distance estimations from
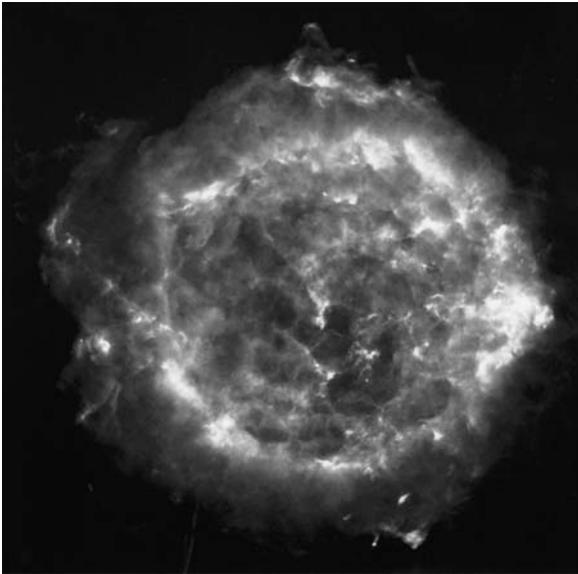
Fig. 11.6 The supernova remnant Cas A; (also known as 3C 461) from observations with the VLA at a wavelength of 6 cm. The array has been used in the configurations A+B+C+D reaching a resolution of 0.2×0.2 arc-sec for a field of 6.1×6.1 arcmin (courtesy NRAO/AUI)

21 cm absorption measurements, or expanding gas filaments visible in the optical range. Today, it is usual to have X-ray observations to confirm the identification. At present more than 200 supernova remnants have been identified in this way. Eventually the total list of such sources in our Galaxy might be doubled, but confusion of low surface brightness objects will set severe limits on old SNRs. The number of pulsars is *much* larger. This result also reflects the greater age of pulsars. Studies of the early time evolution of SNRs in other galaxies is a relatively new research tool in radio astronomy; these can be carried out for times of ≤ 100 years after the SN explosion.

11.4   The Hydrodynamic Evolution of Supernova Remnants

When the core of a star implodes as a supernova, a part of its gravitational binding energy is set free ejecting some fraction of the stellar mass expanding with a high velocity. Initial expansion velocities of $\sim 20\,000\,\mathrm{km\,s^{-1}}$ are observed. This ejected shell will expand freely at early times, because its density is so much larger than that of the density of interstellar gas or circumstellar shells. We expect a collisional shock to form when the shell has swept over a distance about equal to the mean free path in the surroundings. Protons ejected by the SN with $v = 20\,000\,\mathrm{km^{-1}}$ have a kinetic energy of 2 MeV and a mean free path of about 500 pc if the mean density of interstellar neutral hydrogen is $1.2\,\mathrm{cm^{-3}}$ and if all possible loss mechanisms are

taken into account. Obviously, no shock would form under such conditions. But protons from a supernova are charged particles and therefore they will gyrate around interstellar magnetic field lines. A magnetic flux density of $B = 3 \mu$Gauss would result in a Larmor radius of only $R_L \cong 10^{11}$ cm $\cong 10^{-8}$ pc and, even though the energy density of such a field is much too small to affect directly the outward motion of these protons, it will serve as a massless barrier between the supernova gas and the ionized interstellar material so that a hydromagnetic shock will form. The magnetic field is swept up by this process and is collected, along with the gas, by the outward moving shock front. Since the motion of the charged particles along the field lines is not affected, part of the kinetic energy contained in the relativistic particles will be able to diffuse to greater distances. However, the details of the scattering processes for these particles are complicated and are as yet incompletely understood, so that the number of relativistic particles injected by a supernova into the general interstellar medium is rather uncertain.

Four different stages may be distinguished in the evolution of a supernova remnant. This is a rather approximate, empirical classification. Not all SNRs fit into this scheme; the most notable exception is the Crab nebula. The *first stage* is the (approximately) *free expansion*; the mass of the gas swept up by the expanding shell is still less than the initial mass M_s and $R \propto t$:

$$\tfrac{4}{3}\pi r_s^3 \varrho_1 \lesseqgtr M_s \ . \tag{11.8}$$

The details of this phase are governed by the SN explosion. For $\varrho_1 \cong 2 \times 10^{-24}$ g cm^{-3}, $M_s = 0.25 M_0 \cong 0.5 \times 10^{33}$ g, $r_s \cong 1.3$ pc and this stage will last for about 60 years. This phase has been studied in other galaxies.

The *second phase* is the *Sedov or adiabatic phase*. Here the remnant is dominated by swept-up material, but radiative losses are still negligible compared to the total amount of energy that was produced by the supernova. This is therefore expansion under energy conservation and the evolution of the remnant is governed by the interaction between the high-energy particles in the remnant that put pressure on the shell and the surrounding material. The second phase has been investigated for SNRs in our galaxy. In the Sedov phase $R \propto t^{2/5}$.

The *third phase* is the *radiative or snowplow phase*: This occurs when the remnants age is comparable to radiative cooling time scales. Expansion energy is not conserved, but radial momentum is, and so approximately $R \propto t^{1/4}$.

The *fourth phase* is dissipation, when the shock velocity falls below the speed of sound. This happens at $t \sim 10^6$ years.

11.4.1 The Free-Expansion Phase

In the last few years, there has been a great interest in this phase of the SNR evolution due to observational opportunities given by the Very Large Array. The VLA has been used to observe the early development of SNRs in external galaxies. In the case of type II supernovae, a theoretical framework has been developed by

Chevalier (1994). Observationally, classical type I supernovae (so-called type Ia), such as Tycho's SNR, are those showing no H line emission at maximum light, while type Ib are those showing lines of H prominently. Within the class of type II SNRs, there are a number of different subclasses, based on optical light curves. The observations of the early development are in effect the "destructive testing" of the circumstellar environment. The theory presented below is based on measurements obtained for type II and type Ib SNRs.

In the earliest phases of the SN development, we assume that the SN shock wave expands with a constant velocity, v. The progenitor star is believed to have undergone a time-independent mass loss prior to the explosion. The mass-loss rate is taken to be of the order of $\dot{M} = 10^{-4}$ to $10^{-5} M_\odot$ per year. Since the mass loss occurs over a period of about $\sim 10^4$ years, the star is surrounded by a dense circumstellar shell. It is assumed that the density in the circumstellar shell is distributed as

$$\varrho_{cs} = q\,r^{-2} = \frac{\dot{M}}{4\pi v_w} r^{-2}, \tag{11.9}$$

where $\dot{M}$ is the mass-loss rate and v_w is the wind velocity. After the SN explosion, a shell with a mass of 10^{30} g emerges with a velocity of $\sim 10^4\,\mathrm{km\,s^{-1}}$. The density dependence of the SN shell in free expansion is assumed to be:

$$\varrho_{sn} = t^{-3} \left(\frac{r}{t\,U_c} \right)^{-n}, \tag{11.10}$$

where U_c is the expansion velocity of the shell and t the time since the outbreak, n is thought to be between 7 and 12. This has been estimated from fits to density profiles of SNRs at maximum light. Taking this dependence of density on radius, and assuming a spherically symmetric model, we denote the circumstellar shell as component 1, and the SN shell as component 2. Then the deceleration of these components is equal to the pressure difference across the shell:

$$(M_1 + M_2) \frac{\mathrm{d}^2 R}{\mathrm{d}t^2} = 4\pi R^2 \left(P_2 - P_1 \right). \tag{11.11}$$

The masses on the left-hand side of this relation can be obtained by an integration of the density distributions. It is assumed that the lower limits to the integrals of the density for the SN shell, i.e., the photosphere of the progenitor, can be neglected. The pressure terms on the right-hand side are given by $P = \rho\,v^2$. The velocities of the components are given by

$$v_1 = \frac{\mathrm{d}R}{\mathrm{d}t} \tag{11.12}$$

and

$$v_2 = \frac{R}{t} - \frac{\mathrm{d}R}{\mathrm{d}t}. \tag{11.13}$$

Using the mass and pressure relations given above we find that the relation does not involve initial conditions. Assuming a time dependence of radius, $R = K t^m$, where K is a constant, one obtains

$$R = \left[\frac{2U_c^n}{(n-4)(n-3)q}\right]^{1/(n-2)} t^{(n-3)/(n-2)} \sim t^m, \qquad (11.14)$$

where we define $m = (n-3)/(n-2)$. As time increases, the shell radius will tend towards this solution. For the values of n, this is almost a constant velocity. Crucial to understanding this phase of the SNR is the behavior of the energy densities of relativistic electrons and of the magnetic field. From observations, these appear to vary as the thermal energy density. This in turn varies as $1/R^2$, since the volume is increasing as R^3, but mass is being swept up as R. From (10.114), the flux density emitted by the synchrotron process is given by

$$S_v = KH\frac{V}{d^2}B^{n+1}v^{-n}. \qquad (11.15)$$

In order to determine the time evolution of the synchrotron emission, we assume that B varies as t^{-1}, and that the energy of the relativistic electrons varies as R^{-2}. In addition, the radiating volume increases as t^{3m}. Combining these terms, we find that:

$$S_v \sim t^{(5+\delta)/(2+3m)}v^{-(\delta-1)/2}. \qquad (11.16)$$

This time dependence does not match observations at early times. In addition, there is an absorption at lower frequencies. Thus, it has been concluded that thermal ionized gas in the circumstellar shell outside the SNR absorbs the synchrotron radiation. From (10.35), the optical depth of the thermal ionized gas varies as

$$\tau \sim T_e^{-1.35}v^{-2.1}N_e^2 R. \qquad (11.17)$$

The circumstellar electron density N_e varies as R^{-2}, so the variation of τ is R^{-3} or t^{-3m}. Combining the synchrotron emission and circumstellar thermal absorption effects, we have

$$S_v \sim t^{(5+\delta)/(2+3m)}v^{-(\delta-1)/2}\exp\{-T_e^{-1.35}v^{-2.1}N_e^2 R\}. \qquad (11.18)$$

This relation seems to fit the behavior of some SNRs rather well (see Figs. 11.7 and 11.8).

11.4.2 The Second Phase: Adiabatic Expansion

For a realistic treatment numerical models are needed, but many features of the time evolution can be described by the similarity solution of Sedov. Most can be obtained by using simple arguments and first principles if some general properties of the similarity solutions are adopted. If the supernova explosion deposits the total energy E in the remnant, the similarity solution shows that the fraction $K_1 E$ is in the form of heat energy and the remainder in kinetic energy. This factor $K_1 = 0.72$ is a constant in the similarity solution, independent of time. A second result taken from this solution is the ratio $K_2 = 2.13$ of the pressure immediately behind the
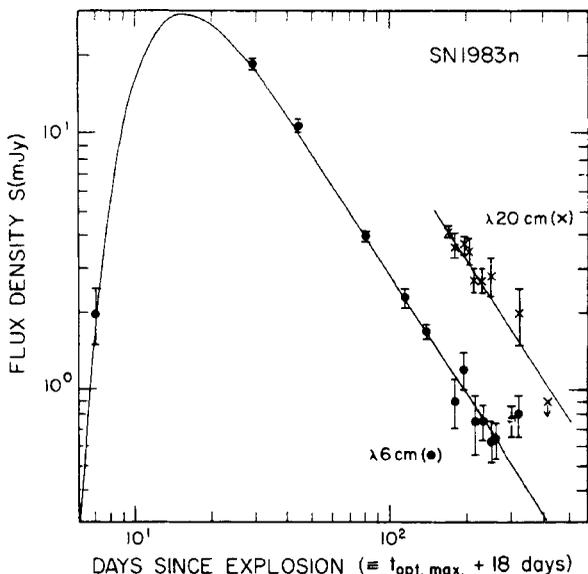
Fig. 11.7 The radio "light curve" for the type Ib supernova 1983N in the galaxy M 83. The data for 20 cm wavelength (*crosses*) and 6 cm wavelength are shown in the same plot. The time scale is measured in days from the estimated date of the explosion, 29 June 1983, eighteen days before maximum light, 17 July 1983. The solid lines represent the best fit [taken from Weiler et al. (1986)]
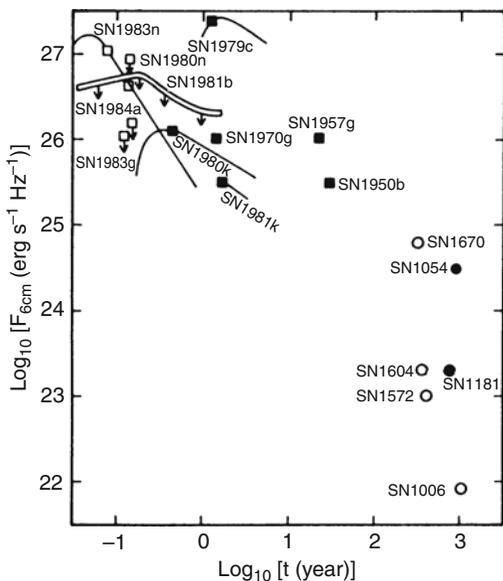
Fig. 11.8 A plot of luminosity versus age for supernovae and young ($< 10^4$ yr) SNRs. The shell SNRs are shown as *open circles*, while filled center (i.e. Crab nebula-like) as *filled circles*. The best fit 6 cm curves are shown for those radio SNRs for which radio "light curves" are available [taken from Weiler et al. (1986)]

shock to the mean pressure of the heated gas within the spherical volume enclosed by the shock. These two factors permit us now to connect this pressure p_2 behind the shock with the total energy of the supernova explosion. For an ideal gas, the specific internal energy e is related to the gas pressure p and the specific volume v by

$$e = \tfrac{3}{2} pv .$$

Solving this for the mean pressure of the gas enclosed by the shock with the volume

$$V = \tfrac{4}{3} \pi r_{\mathrm{s}}^3$$

we find

$$p_2 = K_2 \frac{2}{3} \frac{3K_1 E}{4\pi r_{\mathrm{s}}^3} = \frac{KE}{2\pi r_{\mathrm{s}}^3} \quad (11.19)$$

$$K = K_1 K_2 = 1.53 \quad (11.20)$$

The velocity U_{s} with which the shock is expanding into the surrounding gas of density ϱ_1 and temperature T_1 can be obtained from the "jump conditions" that relate physical quantities on each side of the front. If u_1 and u_2 are the streaming velocities in the gas, conservation of mass requires

$$\varrho_1 u_1 = \varrho_2 u_2 , \quad (11.21)$$

while the conservation of momentum gives

$$p_1 + \varrho_1 u_1^2 = p_2 + \varrho_2 u_2^2 . \quad (11.22)$$

Since we assumed radiation losses to be negligible in this phase, the total energy E in the gas must be conserved during the expansion and the shock must be treated as an adiabatic transition. The third shock condition is then

$$\frac{1}{2} u_1^2 + \frac{p_1}{\varrho_1} + \frac{e_1}{\varrho_1} = \frac{1}{2} u_2^2 + \frac{p_2}{\varrho_2} + \frac{e_2}{\varrho_2} ,$$

which, for an ideal gas, becomes

$$u_1^2 + \frac{2\gamma}{\gamma - 1} \frac{p_1}{\varrho_1} = u_2^2 + \frac{2\gamma}{\gamma - 1} \frac{p_2}{\varrho_2} , \quad (11.23)$$

where $\gamma = c_p/c_v = 5/3$ for a monatomic gas. This γ here should not be confused with $\gamma = (1 - v^2/c^2)^{-1/2}$ of the special theory of relativity in Chap. 10. From (11.21 to 11.23) we obtain after some algebra (see Landau and Lifschitz 1967)

$$\frac{p_2}{p_1} = \frac{2}{\gamma + 1} \frac{\varrho_1}{p_1} u_1^2 - \frac{\gamma - 1}{\gamma + 1} \quad (11.24)$$

and

$$\frac{\varrho_1}{\varrho_2} = \frac{\gamma-1}{\gamma+1} + \frac{2\gamma}{\gamma+1}\frac{p_1}{\varrho_1 u_1^2}. \tag{11.25}$$

For strong shocks with

$$p_2 \gg \frac{\gamma-1}{\gamma-1} p_1 \tag{11.26}$$

we find an asymptotic solution:

$$p_2 = \frac{2\varrho_1}{\gamma+1} u_1^2 \tag{11.27}$$

and

$$\frac{\varrho_1}{\varrho_2} = \frac{\gamma-1}{\gamma+1}. \tag{11.28}$$

Equation (11.28) gives the largest possible jump of the densities for an adiabatic shock, that is, for $\gamma = 5/3$ we obtain $\varrho_2/\varrho_1 \lesssim 4$, even if the ratio of the pressures $p_2/p_1 \to \infty$.

For the expansion velocity of the shock we must put $U_s = u_1$. Substituting (11.19) into (11.27), we find that

$$U_s^2 = \frac{2KE}{3\pi\varrho_1 r_s^3} \qquad . \tag{11.29}$$

The temperature behind the shock is obtained using the equation of state as well as (11.27) and (11.28):

$$T_2 = \frac{\mu\, m_H}{k}\frac{p_2}{\varrho_2} = \frac{\mu\, m_H}{k}\frac{2}{\gamma+1}\frac{\varrho_1}{\varrho_2}u_1^2 = \frac{\mu\, m_H}{k}\frac{2}{\gamma+1}\frac{\gamma-1}{\gamma+1}u_1^2$$

or

$$T_2 = \frac{3}{16}\frac{\mu\, m_H}{k}U_s^2 = 0.061\frac{\mu\, m_H}{k}\frac{E}{\varrho_1 r_s^3} \qquad . \tag{11.30}$$

Here μ is the mean molecular weight per particle for fully ionized gas with a cosmic abundance ($N_H/N_{He} \cong 10$) of $\mu = 0.61$. Since $U_s = dr_s/dt$, (11.29) can be integrated resulting in

$$r_s = \left(\frac{5}{2}\right)^{2/5}\left(\frac{2KE}{3\pi\varrho_1}\right)^{1/5} t^{2/5}$$

or

$$\frac{r_s}{\text{pc}} = 0.26\left(\frac{N_H}{\text{cm}^{-3}}\right)^{-1/5}\left(\frac{t}{\text{yr}}\right)^{2/5} \tag{11.31}$$

for $E \cong 4 \cdot 4 \times 10^{50}$ erg as a representative mean value, and ϱ_1 is the 1 gas density in g cm^{-3}. By using the adiabatic shock conditions we obtain

$$c_s^2 = \frac{2KE}{3\pi\varrho_1 r_s^3}. \tag{11.32}$$

This can be related to T_s using the Boltzmann relation, or T_2, the post-shock temperature from the conservation equation. Then we have

$$\frac{T_2}{K} = 1.5 \times 10^{11} \left(\frac{N_H}{cm^{-3}}\right)^{-2/5} \left(\frac{t}{yr}\right)^{-6/5}. \tag{11.33}$$

T_2 is the temperature immediately behind the shock; for smaller values of r/r_s, T/T_2 decreases as more detailed model computations show. The same applies to the density ϱ/ϱ_2. Consequently, most of the gas of the supernova remnant is collected in a thin layer in which practically all of the radiation, both radio emission and X-rays, is emitted.

According to (11.33), T_2 decreases with time and, when it falls below 10^6 K, the abundant ions C, N and O are able to acquire bound electrons so that they become efficient cooling agents. The thermal energy of the supernova remnant will be radiated away in a short time so that the shock can no longer be treated as adiabatic.

We then enter into the *third phase*. Let t_{rad} be the time at which the remnant has radiated away half of the initial energy E ejected by the supernova. The cooling time then is so short that the matter behind the shock cools quickly and there are no longer any pressure forces to drive the shock. The shell will move at a constant radial momentum piling the swept-up interstellar gas like a snowplow. Constant radial momentum implies that

$$\tfrac{4}{3}\pi r_s^3 \varrho_1 U_s = \text{const},$$

which can be integrated to

$$r_s = r_{rad} \left(\frac{8}{5}\frac{t}{t_{rad}} - \frac{3}{5}\right)^{1/4}. \tag{11.34}$$

Here r_{rad} is the radius of the shell at the time t_{rad} when the adiabatic relation (11.31) ceases to be applicable. The supernova remnant will eventually be lost in the fluctuating density distribution of the interstellar gas when the expansion velocity of the shell approaches $U_s \cong 10$ km s^{-1}, a value close to the RMS velocity dispersion of the interstellar gas. The last phase is the dissipation of the SNR.

11.5   The Radio Evolution of Older Supernova Remnants

Supernova remnants emit radiation in the radio range and thus the question of their radio evolution is of importance. In numerical models the radio emission is computed along with the hydrodynamic evolution; here we will estimate this semi-analytically using ideas first presented by Shklovsky (1960).

As outlined in Sect. 11.3 we will assume that the radio emission of the remnant is synchrotron radiation so that the total flux density of a source assumed to be optically thin is, from (10.114)

$$S_\nu = KH\frac{V}{R^2}B^{n+1}\nu^{-n} \tag{11.35}$$

where V is the volume of the source, B the average magnetic flux density, and ν the observing frequency. K is a constant appearing in the distribution function of the differential number density of the relativistic electrons

$$N(E)\,\mathrm{d}E = KE^{-\delta}\,\mathrm{d}E \tag{11.36}$$

per unit volume with energies between E and $E + \mathrm{d}E$, and H is a constant. The constant K must somehow depend on the total power output of the supernova: the larger the power output the larger K will be. However, no specific relation between these two quantities is known. The problem is: how will S_ν evolve with time if the remnant expands according to (11.31) or (11.34)?

Following (11.33) the temperature T in the supernova remnant remains $> 10^4\,\mathrm{K}$ throughout the whole evolution. Therefore the gas is ionized and of great conductivity, so that any magnetic field remains "frozen" into the gas. If the gas expands, the magnetic field B will decrease. It is difficult to estimate the precise rate, but since expansion along the field lines should be without effect, a dependence

$$B(r) = B_0\left(\frac{r_0}{r}\right)^2 \tag{11.37}$$

seems reasonable.

Let us consider the evolution of the high-energy electrons. These electrons are confined in a volume V whose expansion is caused by their pressure. If the radiation losses are negligible, this expansion is adiabatic, and the work done by the electrons is

$$\mathrm{d}W = -p\,\mathrm{d}V$$

where

$$W = V\int EN(E)\,\mathrm{d}E\,.$$

For a relativistic gas

$$p = \frac{1}{3}e = \frac{1}{3}\frac{W}{V} = \frac{1}{3}\int E N(E)\,dE$$

and

$$\frac{dW}{W} = -\frac{1}{3}\frac{dV}{V}$$

and

$$\frac{dW}{W} = \frac{dE}{E},$$

so that for spherical expansion we have

$$\frac{dE}{E} = -\frac{dr}{r}. \tag{11.38}$$

The number of relativistic electrons remains constant but as the energy of each electron decreases as

$$E(r) = E_0\,\frac{r_0}{r},$$

we obviously must have

$$V_0 \int_{E_1}^{E_2} K_0 E^{-\delta}\,dE = V_0 \left(\frac{r}{r_0}\right)^3 \int_{E_1 r_0/r}^{E_2 r_0/r} K(r)E^{-\delta}\,dE$$

or

$$\boxed{\frac{K(r)}{K_0} = \left(\frac{r}{r_0}\right)^{-(2+\delta)}}. \tag{11.39}$$

The exponent δ remains constant during adiabatic expansion

$$\delta = \delta_0. \tag{11.40}$$

Substituting (11.37, 11.38, 11.39, 11.40) into (11.35) we obtain

$$\boxed{S_v(r) = S_v(r_0)\left(\frac{r}{r_0}\right)^{-2\delta}} \tag{11.41}$$

or, if the time dependence of r in the adiabatic or Sedov phase according to (11.31) is substituted,

$$S_v(t) = S_v(t_0) \left(\frac{t}{t_0} \right)^{-4\delta/5} \qquad . \tag{11.42}$$

This is the time evolution of a supernova remnant. In differential form (11.42) becomes

$$\frac{\dot{S}_v}{S_v} = -\frac{4}{5}\frac{\delta}{t} \tag{11.43}$$

and it relates the spectral index $\alpha = (\delta - 1)/2$ to the time dependence of the source flux. This relation can be tested. For Cas A with $\alpha = 0.77$, $\delta = 2.54$, and $t \cong 300$ years, an annual decrease $\dot{S}_v/S_v = -0.7\,\%$ is predicted, while the observed $\dot{S}_v/S_v = -(1.3 \pm 0.1)\,\%$ is in approximate agreement with this result.

11.6 Pulsars

11.6.1 Detection and Source Nature

Pulsars are radio sources whose emission consist of a train of regularly spaced wideband pulses. Except for the first detected millisecond Pulsar 4C21.53 (Backer et al. 1982), in continuum surveys made before 1967, the time-averaged intensities of even the brightest pulsars were undetectable. The first Pulsars were detected only when Hewish built a high-time-resolution, low-frequency ($v = 81$ MHz) system of great sensitivity. This system was intended for an investigation of the scintillation of small-diameter radio sources by the solar corona. In addition to huge amounts of man-made interference spikes and the searched-for scintillating sources curious spikes spaced at regular time-intervals of 1.3 sec were first recorded on November 28, 1967. These signals were recognized as non-spurious celestial signals by Jocelyn Bell.

A careful search revealed several more sources of this kind, all with different pulse periods, and they eventually were given the name pulsar by the discoverers. When the search started at other radio observatories the total number of known pulsars increased rapidly, reaching 50 by the end of 1969, 500 by 1990 and 1000 in 2000 (see Fig. 11.9). The pulse periods covered a wide range, from $P = 5.1$ s for the longest to $P = 1.5$ ms for the shortest. It was of great importance for the interpretation of the pulsar phenomenon that in 1968 the Crab pulsar with $P = 0.033$ s was detected.

The pulse periods of most pulsars can be measured with a remarkable precision, and it soon became evident that many, if not all, showed a slight systematic lengthening of P with time. Average derivatives $\dot{P}$ of 10^{-12} to 10^{-21} seem to be typical, and any detailed model must account for this. The region from which the radiation is emitted by the pulsar must be smaller in extent than the speed of light times the transit time of the sharpest pulse observed, that is, less than 10 000 km for the Crab
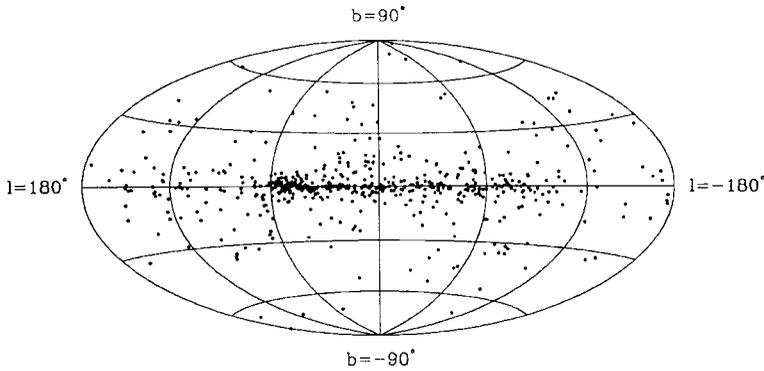
Fig. 11.9 Distribution of 558 pulsars in Galactic coordinates (from Taylor et al. 1993)

pulsar. This diameter can be further reduced if we consider that some kind of clock device is needed for the pulsing. A suitable system would be a rotating body held together by its own gravitation. The centrifugal acceleration then has to be less than the gravitational attraction, that is $(2\pi/P)^2 \leq GM/R^2$. This results in

$$\frac{\bar{\varrho}}{\mathrm{g\,cm^{-3}}} \geq 1.41 \times 10^8 \left(\frac{P}{\mathrm{s}}\right)^{-2} \tag{11.44}$$

for the average density. If an oscillating body instead of a rotating one is considered, the $\geq$ has to be replaced by $=$, and a slightly different numerical constant results. Otherwise the expression remains unchanged.

The Crab pulsar thus requires $\bar{\varrho} \geq 1.3 \times 10^{11}\,\mathrm{g\,cm^{-3}}$, a mean density well outside the possible range for white dwarfs, so only neutron stars can be considered to be feasible models. This is true for all pulsar periods shorter than about $1/2$ s. Only rotating neutron star models need to be considered, because then (11.44) forms a lower limit, while for oscillating star models the equals sign applies. A third observational feature that is typical for all pulsars was found in the very first pulsar observations. The pulse arrival times for a given pulsar depend strongly on frequency. The lower the frequency, the larger the delay of a selected pulse. Quantitative measurements showed the delay to be given by (2.87) with high precision, so that the *dispersion measure* (2.85) is caused by a low-density plasma, that is, by the intervening interstellar medium.

11.6.2 Distance Estimates and Galactic Distribution

Pulsars are designated by PSR Jhhmm±ddmm for **p**ulsating **r**adio **s**ource followed by data for the position in the J2000 coordinate system. Other coordinates such as B1950 coordinates should be prefixed by "B" The last two digits, representing

minutes of declination, can be omitted if the positional accuracy is lower and if there is no danger of confusion with other pulsars. Well-known objects are often quoted by their detection names, that is CP for pulsars detected at Cambridge, MP for those from Molonglo, JP for pulsars from Jodrell Bank, AP for Arecibo pulsars, etc.

The most recent published catalogs of pulsar data are in the book by Lyne and Graham-Smith (2006). More modern data are available on the Internet. Plotting these positions in an equal area map, there is a strong concentration of the pulsars towards the galactic plane as well as an increase of the pulsar numbers for the longitude range $60° \leq \ell \leq 300°$, but part of this non-uniformity may be due to variations in the sensitivity of the pulsar searches. The distribution of pulsars is that of Pop I objects. See the distribution in Fig. 11.9.

Fundamental to any understanding of pulsar physics is a good estimate of the distances. Fortunately there exist several independent methods that supplement each other. Taken together, these give a good estimate of the distance scale.

- Annual parallax measurements have been made for a handful of nearby pulsars. Five pulsars have interferometric parallax measurements, three are obtained from extremely precise timing data and one, the parallax of the Crab pulsar, is based on optical measurements. These data involve state-of-the-art measurements. The possibilities for systematic errors abound, but these form the best available direct determinations of pulsar distances.
- About 26 young pulsars are associated with supernova remnants. Their distances should be about the same as that of the remnants.
- The distance to low-latitude pulsars can be estimated on the basis of $\lambda\,21\,\mathrm{cm}$ neutral hydrogen (H I) absorption measurements (see Chap. 12). About 50 pulsar distances have been successfully investigated in this way. Pulsars are rather unique background sources in view of the precision with which the neutral hydrogen absorption can be measured. This is because the expected line emission at the position of the pulsar can be determined with a precision not possible with other background sources. By timing the line profile measurements to those moments when the pulsar signal is off, the expected H I emission line profile can be measured directly for the pulsar position. This avoids all interpolation of the emission profile from that of neighboring positions. By comparing the shape of the emission and the absorption profile, the pulsar can be positioned in front of or behind a particular cloud. Such clouds are measured in H I emission line surveys. A model for galactic rotation is then used to convert the radial velocities into distance. While individual distance estimates obtained by such means may differ by a large factor, the average pulsar distance scale can be calibrated quite well.
- In the paper announcing the discovery of the first pulsars the dispersion measure DM was used as an indication for the distance by Hewish et al. (1968). Distance estimates using adopted values for N_e are much better than corresponding estimates using a value for the total gas density, N_{tot}, because the values for N_e are much less variable that those of N_{tot}. Pulsar distances, therefore, are usually determined from the measured dispersion measure, DM. Most often the electron densities from the model of Taylor and Cordes (1993) are used. In this model a

smooth run of N_e with galactic radius is assumed and in addition a contribution from spiral arms is included. Individual pulsar distances thus estimated may be in error by a factor of 2, but the pulsar distance scale should be within 30% of that found with other methods.

If the distances are known, estimates for the galactic distribution of pulsars can be constructed. For the *radial distribution* $\varrho_R(R)$, there is a gradual increase towards the galactic center. In the solar neighborhood near 8 kpc, there are about 30 ± 6 observable pulsars per square kiloparsec. This value increases to about 200 at $R = 4$ kpc. In the anticenter direction at $R = 12$ kpc, the density declines to 10–15 pulsars/kpc^2 .

The *z distribution* $\varrho_z(R)$ is approximately exponential with a scale height of 600 pc. This is much larger than that of any other class of Pop I objects, but it can be understood if the *velocity distribution* is considered: Precise position data for pulsars usually also provide quite good estimates for proper motions. Distance information then results in estimates for the transverse velocity. As early as 1982 it was clear that transverse velocities of 100–200 km s^{-1} are quite common for pulsars. It is not quite so easy to derive unbiased estimates, but Lorimer et al. (1995) obtained a mean space velocity of 450 km s^{-1} for a sample of 99 pulsars.

At such velocities about half of the galactic pulsars will escape the gravitational field of the Galaxy, and it is conceivable that the z distribution of pulsars will be much wider than their Pop I progenitors. The fundamental problem of the origin of this formidable space velocity still remains unsolved. Two possible mechanisms are: (1) the disruption of a close binary by a supernova, and (2) asymmetric supernova explosions.

11.6.3 Intensity Spectrum and Pulse Morphology

Pulsar radio emission is investigated with two quite different research goals in mind. One is to find out what pulsars are and what kind of physics is controlling their behavior. A concise summary of the current majority views will be outlined in the remaining parts of this section. A few remarks should, however, be made on a different aspect.

Pulsar radiation is often used as a research tool for investigations of the interstellar medium between pulsar and observer. Several of the constituents of the interstellar medium, both matter and magnetic field, exert influence on the propagation of radiation, and the characteristics of the pulsar emission are such that this influence can be clearly measured. Some of these methods have already been discussed in preceding chapters of this book, so a short reference should suffice here.

In the discussion of the *dispersion measure* in Sect. 2.8 it was shown how the mean electron density N_e could be determined, and in Eq. (3.70) a similar relation was given for the *rotation measure* of the Faraday rotation, depending on both the electron density N_e and the longitudinal magnetic field strength. By combining both

DM and RM, average values for the mean longitudinal magnetic field (3.72) can be determined without too many restrictive assumptions.

Another advantage of the periodic time structure of the pulsar radiation is that in addition to the various mean values, a multitude of time variations can be measured. Interpretations of these data require rather complicated theories of wave propagation in a medium of moving stochastic scattering screens. We will not go further into this interesting field but only mention that by these means it is possible to measure several interesting quantities, such as, the transverse pulsar velocity. These measurements confirm the high pulsar space velocities found from proper motion studies.

Pulsars are nonthermal radio sources with a spectrum that usually can be represented by a power law $S \sim \nu^{-\alpha}$, with a spectral index α close to 1.6, but which show large variations both from one object to the other and, in time, for a single object. For frequencies below 400 MHz there is usually a break in the spectrum.

The total luminosity emitted in the radio range by a typical pulsar averaged over many pulse periods is only $\sim 10^{30} \, \mathrm{erg \, s^{-1}}$. This low *average* luminosity is caused by the small pulse duty cycle. Thus it is rare that pulsars were found in low frequency surveys. A prominent exception is the first millisecond pulsar found (Backer et al. 1982), which was identified as 4C21.53. Pulsar emission was detected only because it occurs in pulse form. If it were not for pulse dispersion, the optimal frequency for detecting and investigating the pulsar phenomenon would be 400 MHz. Strongly dispersed pulsars suffer not only from pulse smearing across the receiving band – this could be removed by some dispersion-removal device, at least for some selected range of dispersion measures – but also from irregular scattering of the pulses at all frequencies. Here the only remedy is to make the measurement at higher frequencies, where the effects of the dispersion by the interstellar plasma are less. The reduction in dispersion more than compensates for the lower signal strength caused by the power law spectrum. Therefore searches for high-dispersion pulsars are now conducted at frequencies in the GHz range.

The pulses are emitted by highly directional beams rotating with the spinning neutron stars. Different pulse periods for different pulsars are found because these stars rotate with different rates. The average total-intensity profiles of most pulsars often contain up to three distinct peaks, and these usually result from a combination of higher than average intensity in a few pulses and lower than average intensity in many. The total pulse width when measured in degrees of rotational phase is roughly independent of the rotational period. The emission cone is remarkably narrow $w_{50} \simeq 30 \, \mathrm{mP} \simeq 10°$, where the pulse width is expressed in fractions of the pulse period.

Pulse profiles are different because:

- the beam cuts the direction towards the observer at different angles for different pulsars and
- the radiation in the emission cones will be different.

Most pulse profiles change slowly with observing frequency and also the component intensity and spacing may change. A representative sample is shown in Fig. 11.10. Such features can easily be accommodated into the polar cap emission model if the emissivity is assumed to depend on the height above the neutron star
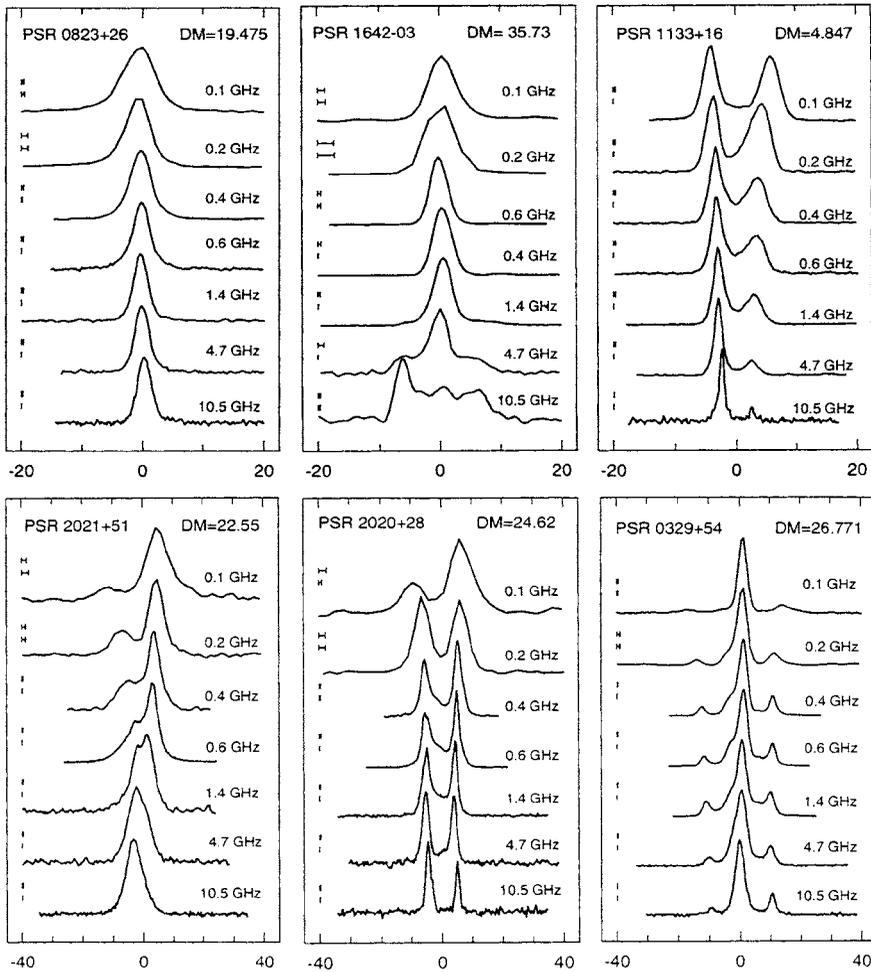
Fig. 11.10 Integrated pulse profiles plotted in the same rotation scale for each pulsar for frequencies 0.1–10.5 GHz. The pulse dispersion has been removed (figure from data in Kuzmin et al. 1998)

surface. So far no consistent model applicable to more than only a selected few pulsar candidates has been presented.

Pulsar radiation is often strongly polarized. Radhakrishnan and Cooke (1969) detected a rapid monotonic change of the linear polarization angle across the pulse of the Vela pulsar PSR 0833-45, at a rate of $6.2°$ polarization angle per degree of rotation. They proposed that pulsar radiation is produced by groups of particles streaming along curved magnetic field lines in the vicinity of the dipole axis. The radiation is linearly polarized with its electric vector in the plane of the field curvature. When the pulsar rotates the polarization angle will also rotate, and this can be

used to determine the angle between the rotation axis and the magnetic axis of the neutron star.

This model of polar cap radio emission with a rotating polarization angle is a basic piece of evidence for intense magnetic fields in the radio emission regions of pulsars. Other observational support is given by the braking mechanism for the spinning neutron stars. This is needed to explain the empirical fact that all pulsars show a noticeable slowing down in their periods, and hence spin rate.

11.6.4 Pulsar Timing

Pulsars are extremely precise clocks. Timing measurements require not only the best terrestrial clocks but also an application of relativistic celestial mechanics to describe the effects of the motion of both the planetary and the pulsar system. This has become part of astrometry and here only a rough indication of the methods must suffice.

Arrival times are measured by sampling the pulsar signal in a narrow receiver passband at a rate of about 1000 samples per pulse. Narrow passbands must be used in order to avoid pulse smearing caused by interstellar dispersion. Alternately some dispersion removal scheme must be employed. If the sampling is synchronized with the local clock setting this results in a determination of the local pulse phase.

If, however, the intrinsic pulsar properties are to be determined, the pulse phase must be converted into the pulsar coordinate system. To do this several different effects have to be taken into account. In order not to neglect any important effects a full relativistic treatment is needed. Here it should suffice to quote the result.

The plasma properties of the intervening interstellar medium are compensated for by eliminating the total dispersion delay using (2.87) with $v_2 \rightarrow \infty$:

$$\frac{\Delta \tau_D}{\mu s} = 4.148 \times 10^9 \left[\frac{DM}{cm^{-3}\,pc} \right] \left(\frac{v}{MHz} \right)^{-2} \quad . \tag{11.45}$$

Similarly it is important to eliminate the motion and the gravitational interaction of the emitter and the receiver with their respective surroundings. Here effects of general relativity theory (GRT) must be included up to order c^{-2} using the PPN system of the parameterized post Newtonian coordinate transformations of Will (1981). Photons from the pulsar will follow null geodesics. It is sufficient to neglect deviations of the signal path by the gravitational field of the sun and only include the gravitational redshift. The time of flight of the pulse N from a pulsar at $\mathbf{R}_N$ to the earth at $\mathbf{r}_N$ is

$$c(t_N - T_N) = |\mathbf{R}_N - \mathbf{r}_N| - \sum_p \frac{2Gm_p}{c^2} \ln \left[\frac{\mathbf{n} \cdot \mathbf{r}_{pN} + r_{pN}}{\mathbf{n} \cdot \mathbf{R}_{pN} + R_{pN}} \right] \quad . \tag{11.46}$$

where $\mathbf{r}_{pN}$ is the position of the receiver and $\mathbf{R}_{pN}$ is the pulsar's position relative to body p, and $\mathbf{n}$ is a unit vector in the direction of the pulsar.

When these time transformation equations are applied to the pulsar timing obser-
vations, it will usually be sufficient to use an appropriate linearization. This allows
one to obtain a timing model for the pulsar. The resulting series of corrected pulse
phases can, in most cases, be represented by a power series in T. The resulting value
for $\dot{P}$ measured in units of $\mathrm{s\,s^{-1}}$ is, in all cases, positive with

$$0 \leq 10^{-20} < \dot{P} < 10^{-11}.$$

The vector $\mathbf{n}$ defines the direction of the wavefront emitted by the pulsar. Quite
often these data can be used to improve $\mathbf{n}$. For some nearby pulsars it has even been
possible to measure the spherical shape of this wavefront by using the positional
variation of the Earth in its orbit around the Sun. This is equivalent to measuring the
pulsar's parallax.

The precision of the timing for most pulsars is very good, with timing uncer-
tainties of a few microseconds. Some even compete with the best terrestrial clocks.
Such a case is pulsar PSR 1937+21, whose timing noise continues to drop as the
time scale of the timing data increases (see Fig. 11.11).

Most pulsars show a regular slowdown that is steady and predictable. However,
some show erratic behavior described as *glitches*. These are a sudden increase of
the rotation rate usually followed by an exponential recovery back to the pre-glitch
frequency. This behavior was first detected in the Vela pulsar, which showed 6 such
glitches with $10^{-6} < \Delta\Omega/\Omega < 3 \times 10^{-6}$ during a period of 14 years (see Fig. 11.12).
Glitches have also been detected in other pulsars. The Crab pulsar has shown a series
with $\Delta\Omega/\Omega \sim 10^{-8}$ to 10^{-7}. In about $2-3\%$ of all pulsars this effect is observable.
The exceptions are the millisecond pulsars.

Possible explanations for glitches might be star quakes of the partly crystalline
neutron star. This abruptly changes the elliptic shape of the pulsar. An alternative is
the interaction of the vortex structure of the superfluid inner parts with the crystal
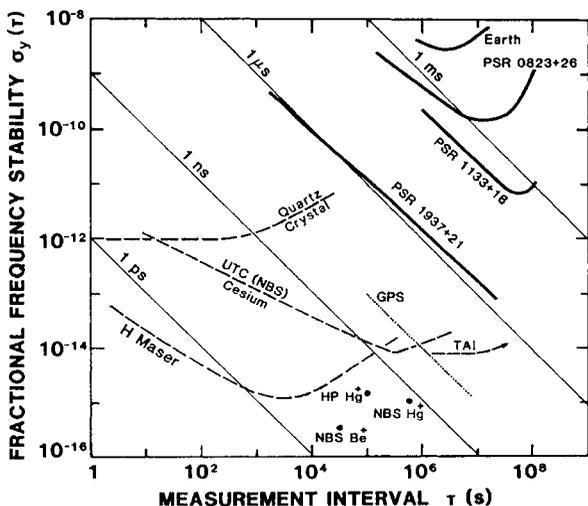
Fig. 11.11 Fractional frequency stability for different technical and natural clock systems (from
Backer and Hellings 1986)
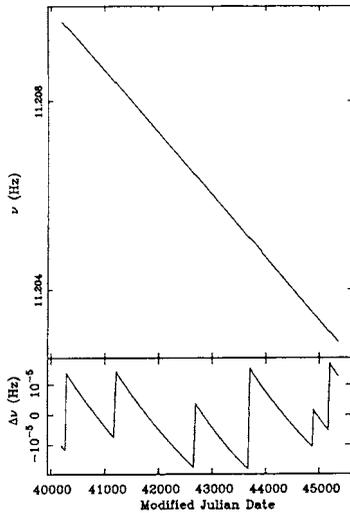
Fig. 11.12 The slowdown of the Vela pulsar over 14 years before (*upper panel*) and after (*lower panel*) subtraction of a constant value of $\dot{v}$ ($v = 1/P$) (after Lyne 2006)

lattice. It is, however, difficult to state how far these explanations go beyond a qualitative description of the observed facts and permit rigorous physics.

As noted, no glitches have yet been observed in millisecond pulsars with $\dot{P} < 10^{-16}$. This is remarkable, since the linear rotational velocity at the surface of these objects must reach a moderate percentage of c, and the shape of such pulsars should be rather elliptical. On the contrary, some of these, like PSR 1937+21, are the most precise celestial clocks known.

11.6.5 Rotational Slowdown and Magnetic Moment

If we adopt the model of a rotating neutron star for a pulsar, an increase of the pulsar period $\dot{P} > 0$ corresponds to a decrease in the rotational energy

$$\frac{\mathrm{d}\mathscr{E}_{\mathrm{rot}}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}I\Omega^2\right) = I\Omega\dot{\Omega} = -\frac{4\pi^2 I\dot{P}}{P^3}\,.$$

This energy must be transported away from the neutron star by some means. The most plausible is magnetic dipole radiation, since neutron stars should usually possess rather strong magnetic fields. This is the result of flux conservation in the collapse of stars.

Let the magnetic moment of the neutron star be $\mathbf{m}_\perp$; then this configuration will experience an energy loss

$$\left(\frac{\mathrm{d}\mathscr{E}}{\mathrm{d}t}\right)_{\mathrm{mag}} = -\frac{2}{3c^3}\left(\frac{\mathrm{d}^2\mathbf{m}_\perp}{\mathrm{d}t^2}\right)^2\,.$$

This is the formula corresponding to the Poynting flux of an electric dipole as given by (6.40). If the magnetic moment $\mathbf{m}_\perp$ is varying with the angular velocity $2\pi/P$ we obtain

$$\frac{\mathrm{d}^2\mathbf{m}_\perp}{\mathrm{d}t^2} = -\mathbf{m}_\perp \cdot \Omega^2 = -\mathbf{m}_\perp \cdot \left(\frac{2\pi}{P}\right)^2 .$$

Adopting a magnetic moment $\mathbf{m}_{\perp 0}$ perpendicular to the rotation axis

$$\mathbf{m}_{\perp 0} = B_0 R^3 \sin\alpha ,$$

where B_0 is the surface magnetic field of the neutron star, R its radius and α the angle between rotation and magnetic axis, we obtain

$$\left(\frac{\mathrm{d}\mathscr{E}}{\mathrm{d}t}\right)_{\mathrm{mag}} = -\frac{2}{3c^3}\left(\frac{4\pi^2 B_0 R^3 \sin\alpha}{P^2}\right)^2 . \tag{11.47}$$

If we now require the loss of rotational energy to be caused by the magnetic dipole radiation, this gives a condition for the magnetic field strength $B_0 \sin\alpha$, or, since $|\sin\alpha| \le 1$,

$$B_0 \ge \left(\frac{3Ic^3 P\dot{P}}{8\pi^2 R^6}\right)^{1/2} . \tag{11.48}$$

For a neutron star with a radius $R = 10^6$ cm and a moment of inertia $I = 10^{45}\,\mathrm{g\,cm^2}$ we then obtain

$$B_0 \ge 3.2 \times 10^{19}(P\dot{P})^{1/2} . \tag{11.49}$$

All pulsar magnetic field strengths given in the literature are computed following this recipe, even when this is not so explicitly stated. It is strictly a lower limit, because, depending on the value of α, fields with larger field strengths could be tolerated.

When the magnetic field strength of pulsars is computed from the observed values of P and $\dot{P}$ by using (11.49), a range of $10^8 \le B_0 \le 10^{14}$ G is obtained, where all those known to be associated with SNRs possess large magnetic fields while binary and millisecond pulsars usually have fields of $B_0 \le 10^{11}$ G.

Taking B_0 as a constant (11.48) gives us a differential equation for the time evolution of P:

$$P\dot{P} = k . \tag{11.50}$$

Integrating and eliminating k gives

$$P^2 - P_0^2 = P_0^2 \left(\frac{2 \dot{P}_0}{P_0} \right) (t - t_0) \,,$$

which results in a timescale

$$\tau = \frac{P_0}{2 \dot{P}_0} \,, \tag{11.51}$$

the *characteristic age* of the pulsar. For the Crab pulsar this gives an age of 1260 years, a reasonable approximation to the known age of 940 years.

Equation (11.50) defines the evolutionary tracks of pulsars provided the basic assumption of $B_0 = \mathrm{const}$ is valid. These evolutionary tracks are straight lines with a slope of -1 in the $\log P$–$\log \dot{P}$ diagram, while lines of constant age are straight lines with the slope $+1$. If one accepts the concept of large values of B_0 as well as large initial values for P from the collapse of the progenitor, very young pulsars should start at positions near the Crab pulsar in the P–$\dot{P}$ diagram. However, if this is to apply to the majority of the observed pulsars, the evolutionary process outlined here cannot be entirely correct. A decay of B_0 with time is not plausible for a variety of reasons, so this is an unsolved problem.

This situation becomes even more acute when millisecond pulsars are considered. These are situated in the lower left corner of the P–$\dot{P}$ diagram with characteristic ages between 10^8 and 10^9 years. Their evolutionary rate is minute. Since most of these are members of binary systems the idea has generally been accepted that they are recycled pulsars, whose rotational speed has been increased by a large factor due to mass transfer from a binary companion. If this mass is first assembled in a Keplerian disk from where it then passes on to the neutron star, a plausible scenario can be constructed.

Some millisecond pulsars seem to be single stars. Here the model is rescued by supposing that these pulsars were formerly members of a binary system which eventually broke up due to supernova explosions.

11.6.6 Binary Pulsars and Millisecond Pulsars

In July 1974 the pulsar PSR B1913+16 with the then remarkable short period $P = 0.059$ s was detected by Hulse and Taylor at the Arecibo Observatory. When they attempted to measure the pulse arrival times they met with curious "irregularities". After accounting for all known gravitational and kinematic effects from the solar system they found the pulse phase jumped by up to $80\,\mu$s from day to day, sometimes even by $8\,\mu$s over 5 min. The observed effects therefore had to be caused by the dynamics of the pulsar. By converting the phase shifts into equivalent radial velocities, Hulse and Taylor could explain their data by a model of a binary pulsar moving in an elliptical orbit. The orbit parameters could be determined from the data by using the classical Thiele-Innes method for a single-line spectroscopic binary.

The orbital elements describing the system (based on the 1974 data) are:

- projected semi-major axis $a_1 \sin i = 6.96 \cdot 10^5$ km
- eccentricity $e = 0.615$
- orbital period $P_b = 0.323^d$
- velocity curve semi amplitude $K_1 = 199 \, \mathrm{km \, s^{-1}}$.

Since $\beta = v/c \simeq 7 \times 10^{-4}$, relativity effects should become rather important for this system. This becomes clear when the expected advance of the periastron due to relativistic effects is calculated. From any textbook on the theory of general relativity, one can obtain the shift of the periastron in radians per orbital revolution:

$$\Delta \omega = \frac{6\pi G(m_1 + m_2)}{c^2 a (1 - e^2)} \quad . \tag{11.52}$$

If the measured orbital elements are inserted, and account is taken of Kepler's third law,

$$\frac{a^3}{P_b^2} = \frac{G(m_1 + m_2)}{4\pi^2} \quad , \tag{11.53}$$

we obtain the rate of the periastron advance

$$\dot{\omega} = 2.1° \, (m_1 + m_2)^{2/3} \, \mathrm{yr}^{-1} \quad . \tag{11.54}$$

From estimated masses for the pulsar binary, the periastron advance is more than 10^5 times that of the orbit of planet Mercury ($0.43"\mathrm{yr}^{-1}$). This system therefore provides the unprecedented chance of a high precision test of the general theory of relativity. This opportunity has been seized by Taylor in subsequent investigations.

In order to carry out such an investigation, high precision values for the pulsar system are needed, and an analysis along the classical lines by converting pulse phases into equivalent radial velocity data is not appropriate. What we need is an arrival time analysis as given in (11.46), but now including the dynamics of the binary pulsar.

We will continue to assume that the pulsar clock controlling the pulse emission is described by a power series in T. But this clock is now carried on an orbit that is elliptic in the Newtonian approximation and described by an osculating ellipse with a rotating line of the apsides if the GRT applies. So, if the pulse phase of the received pulses is to be determined properly, a transformation of the emission time sequence to the barycentric coordinate system must be determined. If terms up to and including v^2/c^2 are included, we obtain from the Schwarzschild solution

$$c(t_N - T_N) = |R_N - r_N| + c\,T_1 \sin E + \frac{2Gm_2}{c^2} \ln \left(\frac{2\,r_N}{r - R} \right) \quad , \tag{11.55}$$

where r_N is the (relative) position of the pulsar in its orbit at emission of the pulse N, and R is the barycentre of this orbit. E is the eccentric anomaly of the pulsar system, which is related to time by the Kepler equation

$$\frac{2\pi}{P_b}(t - t_0) = E - e\sin E$$

and by

$$r = a(1 - e\cos E)$$

to the orbital geometry. The advantage of using E is that the corresponding expressions resulting from a fully relativistic treatment are formally identical and use expressions for a and e that deviate from the classical formulae only by exceedingly small factors.

The quantity T_1 has the dimension of time and is given by

$$T_1 = \frac{\sqrt{Ga}}{c^2}\,e\,m_2\,\frac{m_1 + 2m_2}{(m_1 + m_2)^{3/2}} \qquad . \tag{11.56}$$

T_1 has been determined from a long series of timing observations of PSR 1913+16 with an accuracy of better that 10%. If the semi-major axis a is eliminated from (11.56) we obtain

$$\frac{T_1}{s} = 1.74 \times 10^{-3}\,m_2\,\frac{m_1 + 2m_2}{(m_1 + m_2)^{3/2}} \tag{11.57}$$

where the masses are in solar masses. Introducing the observed value $\dot{\omega} = 4.226°$ yr^{-1} into (11.54) we obtain $m_1 + m_2 = 2.85\,M_\odot$, while (11.57) then results in $m_1 = 1.43\,M_\odot$, $m_2 = 1.42\,M_\odot$ and $\sin i = 0.72$.

This analysis of the orbit of the binary pulsar with the help of GRT has another implication. Since the system is emitting gravitational waves it must lose energy, and therefore the size of the orbit will shrink. Using

$$\mathcal{E} = -\frac{G(m_1 + m_2)}{2a} \,,$$

for the *total energy of the orbit*, and Kepler's third law (11.53), we obtain

$$\dot{P}/P = -\frac{3}{2}\mathcal{E}^{-1}\mathrm{d}\mathcal{E}/\mathrm{d}t \qquad .$$

Assuming that the total energy is changed only by gravitational quadrupole radiation we find

$$\dot{P}_b = -\frac{96}{5}P_b\,\frac{1 + 73/24\,e^2 + 37/96\,e^4}{(1 - e^2)^{7/2}}\,\frac{G^3 m_1\,m_2^5}{(m_1 + m_2)^3 a_1^4\,c^5} \qquad . \tag{11.58}$$

Introducing the measured orbital parameters, one obtains $\dot{P}_b = (-2.40 \pm 0.09) \times 10^{-12}$. A direct fit to the observed phase shift of the pulse arrival times results in

$\dot{P}_b = -2.40 \cdot 10^{-12}$. This provides striking confirmation of Einstein's GRT and gives the first empirical evidence for the existence of gravitational waves.

Since 1975, almost 50 binary and millisecond pulsars have been discovered, in addition to about 30 in globular clusters. In none of these are the relativistic effects nearly as spectacular as in PSR 1913+16. However for two systems, PSR B1534-12 and PSR J1518+1904, interesting results are expected in the future.

The binary pulsar PSR 1913+16 discussed in the preceding sections is a remarkable system for other reasons: it has a remarkably short pulse period, $P = 59$ ms. At the time of its detection it was the second shortest known; only the Crab pulsar was a faster rotator. However, in contrast to the Crab pulsar, the magnetic field strength of PSR 1913+16 is with $B_0 = 2.3 \times 10^{10}$ G. This is almost 4 orders of magnitude smaller. A similar system, PSR B1937+21 with $P = 1.55$ ms, $\dot{P} = 1.05 \times 10^{-4}$ was detected 1981 by Backer. It has an even smaller magnetic field strength, only $B_0 = 4 \times 10^8$ G, and this pulsar is not known to be a member of a binary system. Its characteristic age computed according to (11.51) is rather large (2×10^8 yr), so obviously these systems must have had a history very different from that of the Crab pulsar.

The ms pulsars, which populate the lower-left part of the P–$\dot{P}$ diagram (see Fig. 11.13), form a population different from the regular pulsars. This is also shown by the fact that many of these system are detected in galactic globular clusters, which are generally considered to be rather old objects.

11.6.7 Radio Emission Mechanism

Pulsars were discovered more than 30 years ago. Extensive observational and theoretical efforts have gone into attempts to understand the mechanism of the pulsar emission. The cause for the pulsing is quite clear – it is the rotation of the neutron stars together with their magnetospheres. The central problem is to understand what "makes them shine" as Taylor and Stinebring (1986) put it. Obviously this must be connected with both their large magnetic fields and their fast rotation, but this is not sufficient to select the appropriate radiation mechanism.

The rotation of the pulsar must have a profound influence on the structure of the corotationg magnetosphere, because at the radius $r_c \Omega = 2\pi r_c / P = c$, the field lines reach the velocity of light, and these field lines as well as those originating closer to the magnetic poles cannot be closed. In 1968 T. Gold noticed this feature and pointed out that it should be of importance to the radiation emission mechanism.

The general idea is that the radiation is emitted from the polar caps with the open field lines, and that the pulse width is related to the opening angle. A fundamental problem is the exceedingly high brightness temperature of pulsars, resulting in a ratio $kT_b / h\nu \simeq 10^{28}$. Because of that, this region must be entirely free of matter, otherwise thermal emission occurring at frequencies with $h\nu / kT_b \simeq 1$, which are in the X-ray region, should be observed. As Gold put it "there can be no dirt in the waveguide". The high brightness temperature of the emitted radiation could
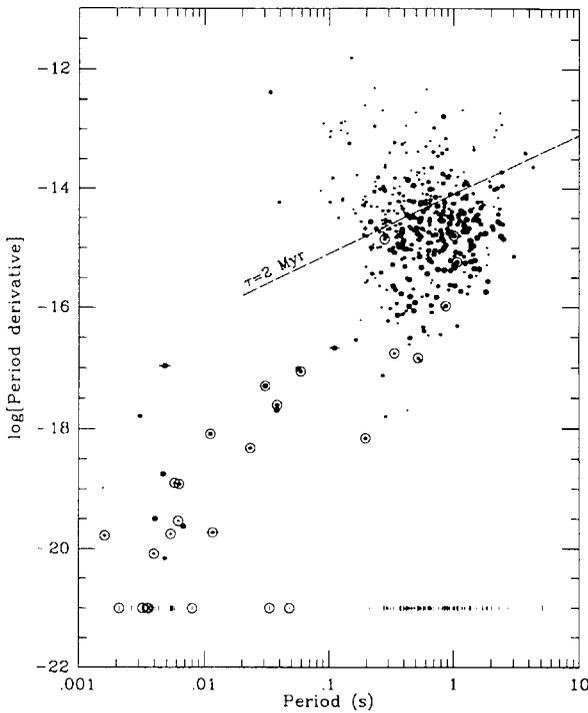
Fig. 11.13 Distribution of periods P and period derivatives $\dot{P}$ for 466 galactic pulsars. *Dots surrounded by circles* denotes pulsars that are members of binary systems. Symbol size denotes relative radio luminosity. (After Taylor et al. 1993)

be caused by a large number of charges emitting in phase, that is, by a coherent emission of radiation.

Another possibility is that the observed high brightness temperature is caused by a masering process in some kind of periodic structure. A number of theories have been proposed, but as yet none is really convincing.

Another characteristic feature we do not yet understand is the large variation in pulse intensity. High time resolution studies have shown strong intensity variations with time scales from 10 to $1000\,\mu$s that reach intensities of 10^4 Jy corresponding to brightness temperatures of 10^{32} K if sizes corresponding to the time scale are adopted for the emission region. In some pulsars the emission is quite erratic, showing strong variations from pulse to pulse so that the smooth average pulse profile may be obtained only after a rather long time. Some other pulsars may switch off altogether, or null, for intervals ranging from a few to several hundred periods, and others show a drifting pattern of sub-pulses. For all these effects, ad hoc explanations have been advanced, but no coherent model valid for all pulsars has yet emerged.

11.7   Extragalactic Sources

Most radio sources are extragalactic, but the radio emission from most of these is considerably weaker than from radio sources in our galaxy. Among the most powerful extragalactic sources are Perseus A (3C84), Cygnus A and Centaurus A. The first two were contained in the initial source surveys made before 1950. Sources like these make up only a few percent of all extragalactic sources. Most of these have active galactic nuclei (AGN), which have luminosities of $\sim 10^{46}$ erg s^{-1}. These AGNs include quasars, Seyfert galaxies, radio galaxies and BL Lacertae objects. AGNs are observed from the radio range up to the gamma ray (~ 100 MeV) region, although it seems clear that most quasars are radio quiet. It is believed that the "prime movers" powering the AGNs are *black holes*; but as yet there is no definite proof of this hypothesis. Most radio sources show evidence for *jets*, that is, elongated features which emanate from the nucleus ending in one of the regions of extended emission. A scheme to unify Quasars, BL Lac, Seyfert and radio galaxies in terms of a single type of object has been proposed by Scheuer and Readhead (1979). In this hypothesis, the viewing angle is the most important parameter: The most rapidly variable sources have jets pointing in our direction.

For these radio sources, the current questions are how the energy is transmitted to the jets from the "prime mover", how the jets are maintained, and the interaction of the radio emission with the surroundings. For the last 25 years, the generally accepted theory for luminous extragalactic radio sources has been the following: a cold, highly collimated, high-velocity flow is formed in or near the parent. This flow propagates without large losses to the ends of the radio lobes where they terminate in shock interactions, converting much of the kinetic energy into relativistic electrons and magnetic fields. There are bright *hot spots* where the flow impinges on the external medium, and the momentum moves the *hot spots* forward, so that the lobes grow in size and luminosity.

In the following, we will consider three extreme examples of extragalactic sources: *radio galaxies*, as represented by Cygnus A, a *cluster of galaxies* which is itself radio quiet, but interacts with the 2.7 K background, and the case of *extreme time variability*.

11.7.1   Radio Galaxies: Cygnus A

Let us apply (10.114) to Cygnus A, which we take as a typical radio galaxy. This source is at a distance of 170 Mpc (for a Hubble constant of 100 km s^{-1} Mpc^{-1}). After optical identification, Cygnus A was thought to be a collision of two galaxies. This was shown to be incorrect, since the radio emission is much more extended than the optical image. This source is an example of an FR2 region (Fanaroff and Riley 1974), defined as a source for which the centroid of the brightness distribution of each lobe is further than half way from the nucleus to its outer boundary, i.e. the source is edge brightened. Conversely FR1 sources are center brightened. FR2 sources are more luminous than FR1's at low radio frequencies. In the case

of Cygnus A the emission consists of two lobes, each of diameter 20 kpc. Within each of the lower brightness lobes are high-brightness compact regions, so-called *hot spots*; these were first found in the 1970s. These maximum intensity peaks are located about 60 kpc from the parent galaxy. In the cm wavelength range, the spectral index, n, has the value 0.75. In (10.112), ν_{max} is not important, but we must set $\nu_{min} = 10$ MHz. We take $\nu_{max} = 10$ GHz. At 1 GHz, the radio luminosity per Hertz is $L_\nu = 4.3 \times 10^{27}$ W Hz^{-1}. Using this result in (10.117), we obtain a minimum magnetic field B of about 10^{-5} Gauss. From (10.118), the particle energy density is 4/3 the magnetic energy density. Then, given the source volume, V, we can obtain an estimate of the minimum electron energy. For the two lobes this is

$$U_{el} = 2.6 \times 10^{57} \text{ erg}$$

where it is assumed that the entire volume is filled uniformly. It appears that the intrinsic rotation measure is < 10 radian m^{-2}. Then from (3.71) $NB_\parallel < 1.23 \times 10^{-9}$ where N is in cm^{-3} and B in Gauss. If we take $B = 10^{-5}$ Gauss, then $N \approx 10^{-4}$ cm^{-3}. The radio lobes must be associated with the parent galaxy. Assuming the lobes to be identical, but one moving towards us with $v \simeq 0.2\,c$, and the other receding with the same velocity (see Sect. 10.6.3), Doppler boosting (11.66) gives a flux density ratio of 4.6 : 1. This difference in the flux densities increases very rapidly with approach (and recession) velocity; in addition the nearer component is observed at a later stage in its development. There are a fairly large number of sources similar to Cygnus A. These are doubles with equal radio flux density, and this requires that the lobes are moving with $v \leq 0.2\,c$. For a lobe-parent separation of $R_s = 100$ kpc, the lobes must have lifetimes of $5R_s/c = 2 \times 10^6$ yr. If the electrons are transported from the central galaxy, these must survive for this time period. For $B \sim 10^{-5}$ Gauss, $\nu_g \approx 170$ Hz, so $\gamma = 1.3 \times 10^4$ for the radiation to reach $\lambda = 1$ cm. Then from (10.129)

$$\frac{1}{E}\frac{dE}{dt} = 1.99 \times 10^{-14} \left(8 \times 10^{-12}\right) \left(1.3 \times 10^4\right) = 2 \times 10^{-21} \,.$$

This lifetime is very long, so an age of $\gg 10^6$ yr is no problem. How to contain this hot mixture of magnetic fields and relativistic particles is not clear. It *cannot* be gravitationally bound! From the volume and density given above the mass of each lobe is $\sim 10^7$ M$_\odot$; the gravitational energy cannot balance the kinetic energy.

Carilli et al. (1991) have presented arc second resolution multi-frequency studies of Cygnus A with the VLA. The analytic methods used to determine source ages are similar to those used for supernova remnants. The minimum energy analysis indicated a source age of 6 Myr, and an expansion velocity of 0.06c. If the B field is a factor of 3 weaker, the lobe separation velocity is 0.01c and the source age is 30 Myr. As in previous studies, the synchrotron spectrum in the *hot spots* is flatter than in regions closer to the parent galaxy; this is evidence for aging of the electrons. Apparently electrons are accelerated in the hot spots and lose energy in the lobes.

11.7.2 An Example of the Sunyaev-Zeldovich Effect: Clusters of Galaxies

In another context, the 2.7 K background photons can interact with energetic electrons in clusters of galaxies by means of the Sunyaev-Zeldovich effect (Sect. 10.12.1). As an example, at 1 cm wavelength, there is weak absorption of order $-700\,\mu$K for the cluster CL 0016 and $-600\,\mu$K toward the cluster Abell 773. For the cluster CL 0016+16, at redshift $z = 0.541$, the X-ray data give 1.6×10^8 K, FWHP size $= 30 \pm 19''$ and RMS electron density $1.2 \times 10^{-2}\,\mathrm{cm}^{-3}$. When combined with the radio data these parameters give reasonably consistent values for a Hubble constant $H_0 \leq 50\,\mathrm{km\,s}^{-1}\,\mathrm{Mpc}^{-1}$. More accurate radio and X-ray data may allow an estimate of H_0 which is independent of Cepheid variable parameters.

If the source is resolved, one can carry out a calculations in which images of X ray emission from electrons is compared to images of Synchrotron emission. One can assume that the relativistic electrons give rise to X rays by means of the inverse Compton effect also give rise to Synchrotron emission. Combing these data, one can determine quantities such as the *B* field without recourse to the equipartition assumption. Results from the X ray satellites *Chandra* or *XMM* have been used in such investigations. An example of this analysis is given in the paper by Goodger et al. (2008).

11.7.3 Relativistic Effects and Time Variability

In the mid-1960s, it was noted that the radio output of QSOs varied on time scales of months. A simple conversion of time scale to distance gave problems with brightness temperatures (these are presented in the next section). M. Rees pointed out that an analysis of the time scale of these changes must take the finite speed of light into account.

Suppose material is ejected from a source S with a constant speed (cf. Fig. 11.14). If $v \ll c$, after a time t, the material lies on a sphere of radius vt. This changes dramatically if $v \approx c$. Then, for an observer at O, the distance to which the material appears to have moved depends on the direction. The apparent motion as seen from O can be understood in terms of 2 very short light flashes. The first is at the start. Then the travel time for SO is t. Later a second flash is emitted from R'. This arrives at O at t_2. At R', the source has moved by vt and is closer to O by $vt \cos\theta$ and therefore R is smaller by $vt \cos\theta$, so

$$t_2 = t_1 + t - \frac{vt}{c}\cos\theta$$

and the apparent expansion speed is

$$v_{\mathrm{app}} = \frac{r}{t_2 - t_1} = \frac{v}{1 - \dfrac{v}{c}\cos\theta}. \tag{11.59}$$

Fig. 11.14 A sketch to illustrate the geometry of light propagation in an expanding source

For transverse velocities this is

$$(v_{app})_{tranverse} = \frac{v \sin\theta}{1 - \frac{v}{c}\cos\theta}. \tag{11.60}$$

Clearly those parts of the source approaching at an angle $\cos\theta \approx v/c$ make the largest contribution to the continuum emission. The observed flux density from a region moving towards us will be increased over the intrinsic flux density by the relativistic Doppler effect. For a region of intrinsic luminosity $L_0(v)$, a flux density between v_0 and $v_0 + dv_0$ crosses a shell of radius R_0 in the emitter frame:

$$S(v_o)\,dv_0 = \frac{L_0(v)\,dv}{4\pi R_0^2}.$$

For a Doppler shift

$$(1+z) = \frac{\lambda_{obs}}{\lambda_{emitted}} = \frac{1 + \frac{v}{c}\cos\theta}{\sqrt{1 - \frac{v^2}{c^2}}},$$

the flux density is

$$S(v)\,dv = \frac{1}{(1+z)^2} S(v_0)\,dv_0 = \frac{1}{1+z} S_0(v(1+z))\,dv. \tag{11.61}$$

Both the energy *and* the rate of photons are modified by the same factor; this can also be seen by transforming the Poynting vector. If R_0 is the distance in the emitter frame, then the distance to the emitter, R, measured by the observer is $R = R_0/(1+z)$. Then

$$S(v) = \frac{L_0(v(1+z))}{4\pi R^2 (1+z)^3}. \tag{11.62}$$

If

$$L_0(v) = k v^{-\alpha} \tag{11.63}$$

then

$$L_0(v(1+z)) = k v^{-\alpha}(1+z)^{-\alpha}. \tag{11.64}$$

So

$$S(\nu) = \frac{L_0(\nu)}{4\pi R^2} \left[(1+z)^{-3-\alpha}\right] ; \tag{11.65}$$

that is, the observed flux density, S, is

$$\boxed{S = S_0 \, (1+z)^{-(3+\alpha)}} \quad . \tag{11.66}$$

This effect is known as *Doppler boosting*. As an extreme example, we consider a plasma ejected with $\gamma = 6$, $\beta = 0.986$, at an angle $9.6°$ with our line of sight. For a receding region with $\alpha = 1$, the observed flux density would be 0.000051 times the intrinsic source flux density. For a region approaching with $\gamma = 6$, the observed flux density would only be 1291 times the intrinsic value.

If it were not for these relativistic effects, one could use observations of time variability to set limits on the source size, (i. e., a large variation in 3 months requires a size of 3 light months) and this should allow one to set limits on the B field using the self absorbed synchrotron emission from a determination of $\nu_G = eB/m = 17B\,\mu$Gauss. However, beaming effects require a detailed knowledge of source structure and kinematics. This is a current area of study using VLBI.

Irrespective of these questions, (10.129) provides an effective upper limit on source brightness temperatures. If, for example, the brightness temperature is determined to be 10^{12} K, at 3 cm, then the radiation energy density is, roughly,

$$u_{ph} = 8\pi \, \frac{k\Delta T\Delta\nu}{\lambda^2 c} \cong 10^{-4}\,\mathrm{erg\,cm^{-3}\,s^{-1}}.$$

Then

$$\frac{1}{E}\frac{dE}{dt} = -10^{-12}\,\gamma.$$

To have electrons radiate at $\lambda = 3\,\mathrm{cm} = 10^{10}\,$Hz by the synchrotron process with $B \sim 10^{-5}$ Gauss, then $\nu_G = 170\,$Hz, so $\gamma \cong \sqrt{\nu/\nu_G} \cong 7700$ and

$$\frac{1}{E}\frac{dE}{dt} = -7.7 \times 10^{-9}\,\mathrm{s^{-1}}$$

or about 4 years. This means that the energy of such electrons must be replenished on this time scale. Clearly, much larger brightness temperatures are out of the question, unless there is a nearby source which can accelerate the electrons on short time scales.

We can turn the problem around and determine the largest γ which can be maintained in the presence of the 2.7 K photons. Here the photon energy density is a blackbody, so

$$u = \frac{4\pi\sigma}{c}\,T^4 = 2.376 \times 10^{-14}\,T^4\,\mathrm{erg\,cm^{-3}\,s^{-1}} = 1.26 \times 10^{-12}\,\mathrm{erg\,cm^{-3}\,s^{-1}}.$$

Then

$$-\frac{1}{E}\frac{dE}{dt} = 3.03 \times 10^{-20}\,\gamma.$$

For time scales of 10^6 years, this would limit γ to 10^6. For cosmic rays, this limits proton energies to $E \sim 10^{15}$ eV.

Problems

1. (a) Use Eq. (11.1), with $r_0 = 7 \times 10^{10}$ cm to determine the emission measure of the quiescent solar atmosphere.
(b) Determine the optical depth of the quiescent solar atmosphere, looking at the center of the Sun, using Eq. (10.35), with $T_e = 10^6$ K and for a frequency 100 MHz. What is the brightness temperature of the Sun?

2. (a) At 5 GHz, the brightness temperature in the outer parts of Orion A is $\sim$0.5 K. Use the assumption of an optically thin, smooth Bremsstrahlung emission from a region with $T_e = 6500$ K which fills the telescope beam completely to calculate the brightness temperature of these regions at 23 GHz (use Eq. (10.36)). In the upper part of Fig. 11.5 is the 23 GHz map of Orion A. The RMS noise in this map is 0.1 K. Would this emission from the outer parts of Orion A be detected in the 23 GHz map?
(b) At what frequency would the outer regions of Orion A have an optical depth of unity?

3. Suppose a solar type star is to be detected at the 1 μ Jy level at $\lambda = 3$ mm. Given that $T_e = 5700$ K and $r_0 = 7 \times 10^{10}$ cm, what is the maximum distance that such a star can be detected?

4. (a) Calculate the radio continuum flux density at $\nu = 10$ GHz for a B3 supergiant $(T = 1.6 \times 10^4$ K, $r_0 = 3.6 \times 10^{12}$ cm). Use an electron and ion density of 10^{10} cm^{-3} and Eq. (11.7) with $r = r_0$ for such a star which is 3 kpc distant.
(b) Is this source detectable with the 100 m telescope if the receiver noise is 50 K, if 1 Jy corresponds to 1.3 K, T_A, and the receiver bandwidth is 500 MHz? Do not consider confusion effects.
(c) With the VLA at 23 GHz, a source was found to have a continuum flux density of 27 mJy. This is at a distance of 7 kpc. What would $n_0 r_0^2$ have to be if this emission be caused by an ionized outflow of $T = 20000$ K?
(d) If $n_0 = 10^{10}$ cm^{-3}, what is r_0?

5*. Reformulate equation (11.7) by substituting the mass loss rate for a steady ionized wind. The product of electron density and radius squared can be related to

$$n_e(r) = \frac{\dot{M}}{4\pi r^2 v_w \mu m_H}.$$

Substitute this relation into the equation in Problem 4, where $\dot{M}$ is the mass loss rate in $10^{-5}\dot{M}$ (per year) and v_w is the wind velocity, in units of 1000 km s^{-1}. μ is

the average mass as a multiple of the mass of the hydrogen atom, m_H. Use this result to find the mass loss rate for the source analyzed in Problem 4
(a) if $v_w = 100\,\mathrm{km\,s^{-1}}$.

6. The parameters of a B0 zero age main sequence (ZAMS) star are $T = 3.1 \times 10^4$ K, luminosity $L = 2.5 \times 10^4\,L_\odot$ and radius $r = 3.8 \times 10^{11}$ cm. Suppose this object has a mass loss rate of $10^{-6}M_\odot$ per year and is 7 kpc distant. What is the flux density for a frequency of 10 GHz? Is this source detectable with the 100 m telescope? With the VLA?

7. From the flux density at 100 MHz in Fig. 10.1, calculate the peak brightness temperature of the Crab nebula, if the FWHP angular size of this source is $5'$, and the source shape is taken to be Gaussian. Repeat this calculation for a frequency of 10 GHz, using the same angular size. If the maximum brightness temperature for Bremsstrahlung emission from a pure hydrogen HII region is considered to be 20 000 K, is the emission from the Crab nebula thermal or non-thermal?

8. If Cassiopeia A has an angular diameter of $5.5'$, determine the present-day linear size of Cassiopeia A if this source is 3 kpc from the Sun. If the explosion occurred in 1667, and if the expansion velocity has been constant, what is v_{exp}? The VLA can measure positions of "point" sources to $0.05''$ accuracy. If there are such "point" features in Cas A, and if these features do not change shape with time, but merely move with v_{exp}, over what time scale would you have to carry out VLA measurements to observe expansion?

9. (a) Use (Eq. 11.43) with $\delta = 2.54$, to extrapolate the radio flux density of Cassiopeia A to a time when this source was 100 years old; that is, what was S_ν at 100 MHz in 1777? See Fig. 10.1 for the flux today. What would be the angular size if the expansion is linear with time?
(b) Calculate the peak brightness temperature in 1777 assuming that this source was Gaussian, using (Eq. 8.20).

10. There is a sharp decrease in the flux density of Cassiopeia A at a frequency of about 10 MHz. If this source is 3 kpc from the Sun, and the average electron density is $0.03\,\mathrm{cm^{-3}}$, calculate whether the cause of the fall off is free–free absorption by electrons along the line of sight. These will have an effect only if $\tau = 1$. Use Eq. (10.35) with $T_e = 6000$ K.

11. (a) Make use of the *minimum energy theorem* to estimate the magnetic fields and relativistic particle energies on the basis of synchrotron emission, using (Eq. 10.119) to obtain a numerical result if the spectral index, n, is 0.75, and $b(n) = 0.086$. For the maximum frequency, take ν_{max} equal to 50 GHz and for the minimum frequency, ν_{min}, equal to 0.1 GHz. Finally, take η (the ratio of other relativistic particles to that of electrons) to be 10. With these parameters, show that the expression for the B field is

$$B_{eq} = 1.2 \times 10^{-5} \left(\frac{S_\nu[\mathrm{Jy}]\,R^2[\mathrm{Mpc}]\,\nu^{0.75}[\mathrm{GHz}]}{V[\mathrm{kpc}]} \right)^{2/7}.$$

12. Assume that the galaxy NGC 253 is similar to our Milky Way. The radius of the synchrotron-emitting halo is 10 kpc at a distance of 3.4 Mpc. At $\nu = 8.7$ GHz,

the integrated flux density is 2.1 Jy and the spectral index is $n = 0.75$ ($S_\nu = S_0(\nu/\nu_0)^{-0.75}$). Take $\nu_{max} = 50\,\text{GHz}$ and $\nu_{min} = 10\,\text{MHz}$ to calculate the B field and estimate the relativistic particle energy assuming that the minimum energy condition holds, i.e. using (Eq. 10.119).

13. Assume that the distance to Cygnus A is 170 Mpc. This source has a flux density of 10^4 Jy at 100 MHz. Assume that the electrons radiate over a frequency range from 10 MHz to 50 GHz with a spectral index $n = 0.75$. Find the power, P, radiated by the electrons via the synchrotron process, using

$$P = 4\pi R^2 \int_{\nu_{min}}^{\nu_{max}} S_\nu d\nu \ .$$

Compare to the total energy of the radio lobes, 2×10^{57} erg, calculated under the assumption of equipartition. What is the lifetime of these relativistic electrons if synchrotron emission is the only loss mechanism? Compare this to the expected lifetime of the source if the lobes are 7×10^4 pc apart and are thought to be moving with a speed $<0.2c$. What do you conclude about the need to replenish the energy of the electrons?

14. (a) The quasi-stellar radio source 3C273 has a red shift of 0.16. Take the Hubble constant, H_0, to be $70\,\text{km}\,\text{s}^{-1}$ per Mpc. Find this distance. The flux density varies on a time scale of months. Use a simple relation of $R = ct$ to determine the source size, without taking any relativistic effects into account. What is the angular size? Next, using this angular size, convert the flux density at 20 GHz, which is $\sim$20 Jy, into a source brightness temperature. What is the result? Does this exceed the maximum temperature of 10^{12} K, the limit predicted by the inverse Compton effect? This is an indication that relativistic beaming effects are important.
(b) Refer to (Eq. 11.59); for transverse velocities this is given by (Eq. 11.60). What is the angle at which the apparent transverse velocity is a maximum? What is the apparent velocity at this angle? If the apparent expansion velocity is $7c$, what is the beaming angle? There is no counter jet. Explain why not, taking "Doppler boosting" into account, using (Eq. 11.66).

Chapter 12
Spectral Line Fundamentals

12.1 The Einstein Coefficients

In local thermodynamic equilibrium (LTE) the intensities of emitted and absorbed radiation are not independent but are related by Kirchhoff's law (1.14). This applies to both continuous radiation and line radiation. The Einstein coefficients give a convenient means to describe the interaction of radiation with matter by the emission and absorption of photons.

Consider a cavity containing atoms with discrete energy levels E_i. According to Einstein (1916) a system in the excited level E_2 will return spontaneously to the lower level E_1 with a certain probability A_{21} such that $N_2 A_{21}$ is the number of such spontaneous transitions per second in a unit volume if N_2 is the density in the state E_2 (Fig. 12.1). The energy levels E_1 and E_2 have a finite energy spread so that $E_2 - E_1$ will have a certain energy distribution. Converting this into frequencies using $E_2 - E_1 = h\nu$, the absorption line will be described by a line profile function $\varphi(\nu)$ which is sharply peaked and normalized so that

$$\int_0^\infty \varphi(\nu)\,d\nu = 1\,. \tag{12.1}$$

If the intensity of the radiation field is I_ν (see Sect. 1.3) we can define an average intensity by

$$\bar{I} = \int_0^\infty I_\nu\, \varphi(\nu)\,d\nu \tag{12.2}$$

and the probability of the absorption of a photon is $B_{12}\bar{U}$ such that the number of absorbed photons is $N_1 B_{12} \bar{U}$ where $\bar{U} = 4\pi\bar{I}/c$ is the average energy density of the radiation field. Einstein found that to derive Planck's law another emission process proportional to $\bar{U}$ was needed. This is $N_2 B_{21} \bar{U}$, equal to the number of photons emitted by *stimulated* emission.

If the system is in a stationary state, the number of absorbed and emitted photons must be equal, so that

Fig. 12.1 Transitions
between the states 1 and 2
and the Einstein probabilities
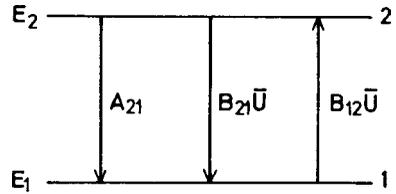
$$N_2 A_{21} + N_2 B_{21} \bar{U} = N_1 B_{12} \bar{U} \qquad .\qquad (12.3)$$

The Einstein coefficients A_{21}, B_{21} and B_{12} are not independent as can be seen
if we consider a system in full thermodynamic equilibrium (TE). The systems in
the cavity are not all in one state but are distributed over different states so that the
different atomic levels are populated according to the Boltzmann distribution

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} \exp\left(-\frac{h\nu_0}{kT}\right), \qquad (12.4)$$

where g_1 and g_2 are the statistical weights of the states and T is the temperature of
the cavity in Kelvin. Solving (12.3) for $\bar{U}$ gives

$$\bar{U} = \frac{A_{21}}{\dfrac{N_1}{N_2} B_{12} - B_{21}} . \qquad (12.5)$$

From this and (12.4) we have

$$\bar{U} = \frac{A_{21}}{\dfrac{g_1}{g_2} \exp\left(\dfrac{h\nu_0}{kT}\right) B_{12} - B_{21}} . \qquad (12.6)$$

But in TE we know that $\bar{U}$ must be given by the Planck function (1.13):

$$\bar{U} = \frac{4\pi}{c} B_\nu(T) = \frac{8\pi h \nu_0^3}{c^3} \frac{1}{\exp\left(\dfrac{h\nu_0}{kT}\right) - 1} . \qquad (12.7)$$

The expressions (12.6) and (12.7) should be identical; this is so only if

$$g_1 B_{12} = g_2 B_{21} \qquad (12.8)$$

and

$$A_{21} = \frac{8\pi h \nu_0^3}{c^3} B_{21} \qquad . \qquad (12.9)$$

   In deriving these relations no reference is made to any thermodynamic property of the cavity. Therefore they must be valid for systems independent of the assumption of a TE environment.

12.2   Radiative Transfer with Einstein Coefficients

When the radiative transfer was considered in Sect. 1.4, the material properties were expressed as the emission coefficient ε_ν and the absorption coefficient κ_ν. Both ε_ν and κ_ν are macroscopic parameters; for a physical theory these must to be related to atomic properties of the matter in the cavity. If line radiation is considered, the Einstein coefficients in Sect. 12.1 are very useful because these can be linked directly to the properties of the transition responsible for the spectral line. For radiative transfer ε_ν and κ_ν are needed, so we must investigate the relation between κ_ν and A_{ik} and B_{ik}. This is best done by considering the possible change of intensity I_ν passing through a slab of material with thickness ds just as in Sect. 1.4. Now we will use A_{ik} and B_{ik}.

   According to Einstein there are three different processes contributing to the intensity I_ν. Each system making a transition from E_2 to E_1 contributes the energy $h\nu_0$ distributed over the full solid angle 4π. Then the total amount of energy emitted spontaneously is

$$dE_e(\nu) = h\nu_0 N_2 A_{21} \, \varphi_e(\nu) \, dV \frac{d\Omega}{4\pi} \, d\nu \, dt \, . \tag{12.10}$$

   For the total energy *absorbed* we similarly obtain

$$dE_a(\nu) = h\nu_0 N_1 B_{12} \frac{4\pi}{c} I_\nu \, \varphi_a(\nu) \, dV \frac{d\Omega}{4\pi} \, d\nu \, dt \tag{12.11}$$

and for the *stimulated* emission

$$dE_s(\nu) = h\nu_0 N_2 B_{21} \frac{4\pi}{c} I_\nu \, \varphi_e(\nu) \, dV \frac{d\Omega}{4\pi} \, d\nu \, dt \, . \tag{12.12}$$

   The line profiles $\varphi_a(\nu)$ and $\varphi_e(\nu)$ for absorbed and emitted radiation could be different, but in astrophysics it is usually permissible to put $\varphi_a(\nu) = \varphi_e(\nu) = \varphi(\nu)$. For the volume element we put $dV = d\sigma \, ds$, where $d\sigma$ is the unit area perpendicular to the beam direction. For a stationary situation, we find

$$dE_e(\nu) + dE_s(\nu) - dE_a(\nu) = dI_\nu \, d\Omega \, d\sigma \, d\nu \, dt$$
$$= \frac{h\nu_0}{4\pi} \left[N_2 A_{21} + N_2 B_{21} \frac{4\pi}{c} I_\nu - N_1 B_{12} \frac{4\pi}{c} I_\nu \right] \varphi(\nu) \, d\Omega \, d\sigma \, ds \, d\nu \, dt \, .$$

   The resulting equation of transfer with Einstein coefficients is

$$\frac{dI_v}{ds} = -\frac{h v_0}{c} \left(N_1 B_{12} - N_2 B_{21} \right) I_v \, \varphi(v) + \frac{h v_0}{4\pi} N_2 A_{21} \, \varphi(v) \qquad . \qquad (12.13)$$

Comparing this with (1.9) we obtain agreement by putting

$$\kappa_v = \frac{h v_0}{c} N_1 B_{12} \left(1 - \frac{g_1 N_2}{g_2 N_1} \right) \varphi(v) \qquad\qquad (12.14)$$

and

$$\varepsilon_v = \frac{h v_0}{4\pi} N_2 A_{21} \, \varphi(v) \qquad , \qquad (12.15)$$

where we used (12.8) to relate B_{12} and B_{21}. The factor in brackets in (12.14) is the correction for stimulated emission. In radio astronomy, where the stimulated emission almost completely cancels the effect of the true absorption, this is important. How this comes about is best seen if we investigate what becomes of (12.13–12.15) if LTE is assumed.

From (12.14) and (12.15) we find by substituting (12.8) and (12.9) that

$$\frac{\varepsilon_v}{\kappa_v} = \frac{2h v^3}{c^2} \left(\frac{g_2 N_1}{g_1 N_2} - 1 \right)^{-1} .$$

But for LTE, according to (1.14), this should be equal to the Planck function (1.13), resulting in

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} \exp\left(-\frac{h v_0}{kT} \right) \qquad . \qquad (12.16)$$

In LTE, the energy levels are populated according to the same Boltzmann distribution (12.4) for the temperature T that applied to full TE. Then the absorption coefficient becomes

$$\kappa_v = \frac{c^2}{8\pi} \frac{1}{v_0^2} \frac{g_2}{g_1} N_1 A_{21} \left[1 - \exp\left(-\frac{h v_0}{kT} \right) \right] \varphi(v) \qquad , \qquad (12.17)$$

where we have replaced the B coefficient by the A coefficient, using

$$B_{12} = \frac{g_2}{g_1} A_{21} \frac{c^3}{8\pi h v^3} .$$

This last relation is obtained from (12.8) and (12.9). In (12.17) the expression in brackets is the correction for stimulated emission. Since

$$\frac{h}{k} = 4.799\,27(15) \times 10^{-11}\,\mathrm{K\,Hz^{-1}}, \qquad (12.18)$$

Table 12.1 Physical line parameters at different frequencies

Line	v/Hz	$\dfrac{hv}{k}$/K	$T_{10\%}$/K
Ly cont.	3.29×10^{15}	1.58×10^5	6.9×10^4
Hα	4.57×10^{14}	2.19×10^4	9.5×10^3
21 cm	1.42×10^9	6.82×10^{-2}	3.0×10^{-2}

we obtain the values in Table 12.1. In it we give for a few transitions the temperatures for which stimulated emission will be important. In the second column, we give the line frequency, in the third $T_0 = hv/k$, and in the fourth the temperature at which the correction for stimulated emission is 10%.

For radiation in the ultraviolet and visual range, the correction for stimulated emission is small, and only absorption is relevant. Only when $hv/kT < 1$ must stimulated emission be taken into account, as in the radio range. When $hv/kT \ll 1$ it is sufficient to use the first term of the Taylor series

$$1 - \exp\left(-\frac{hv}{kT}\right) \cong hv/kT - \frac{1}{2}(hv/kT)^2 + \ldots . \tag{12.19}$$

Thus stimulated emission cancels most of the absorption; for molecular line radiation in the mm wavelength range of low-temperature regions ($T < 10\,\text{K}$) some of the higher terms in (12.19) or even the full exponential function might be needed. The last column in Table 12.1 gives the temperature at which the correction for stimulated emission amounts to 10%, that is, for which $\exp\{-hv/kT\} = 0.1$.

12.3 Dipole Transition Probabilities

The simplest sources for electromagnetic radiation are oscillating dipoles. Radiating electric dipoles have already been treated classically in Chap. 6, but it should also be possible to express these results in terms of the Einstein coefficients. There are two types of dipoles that can be treated by quite similar means: the electric and the magnetic dipole.

Electric Dipole. Consider an oscillating electric dipole

$$d(t) = ex(t) = ex_0 \cos \omega t . \tag{12.20}$$

According to electromagnetic theory, this will radiate. The power emitted into a full 4π steradian is, according to (10.11),

$$P(t) = \frac{2}{3} \frac{e^2 \dot{v}(t)^2}{c^3} . \tag{12.21}$$

Expressing $x = d/e$ and $\dot{v} = \ddot{x}$, we obtain an average power, emitted over one period of oscillation of

$$\langle P \rangle = \frac{64\pi^4}{3c^3} v_{mn}^4 \left(\frac{ex_0}{2}\right)^2 . \tag{12.22}$$

This mean emitted power can also be expressed in terms of the Einstein A coefficient:

$$\langle P \rangle = h\nu_{mn} A_{mn} . \tag{12.23}$$

Equating (12.22) and (12.23) we obtain

$$\boxed{A_{mn} = \frac{64\pi^4}{3hc^3} v_{mn}^3 |\mu_{mn}|^2} , \tag{12.24}$$

where

$$\mu_{mn} = \frac{ex_0}{2} \tag{12.25}$$

is the mean electric dipole moment of the oscillator for this transition.

Strictly speaking, expression (12.24) is applicable only to classical electric dipole oscillators. It turns out to be valid for quantum systems also. Thus the transition probability for atomic hydrogen close to the Lyman limit is about 10^9 s^{-1}. This is obtained by putting $x_0 = a_0 = 5.29 \times 10^{-9}$ cm = Bohr radius, $v_{mn} = cR_\infty = 3.29 \times 10^{15}$ Hz = frequency at the Lyman limit, and $\mu_{mn} = ea_0/2 = 4.24 \times 10^{-19}$ [cgs].

Magnetic Dipole. For a magnetic dipole,

$$m(t) = m_0 \cos \omega t , \tag{12.26}$$

the corresponding Einstein A coefficient is

$$A_{mn} = \frac{64\pi^4}{3hc^3} v_{mn}^3 |\mu_{mn}^*|^2 , \tag{12.27}$$

where μ_{mn}^* is the mean magnetic dipole moment of the oscillator for this transition. If we again apply this relation to the hydrogen atom and compute

$$|\mu_{mn}^*| = \frac{e\hbar}{2m_e c} \cong 9.27 \times 10^{-21} \text{ erg Gauss}^{-1} \tag{12.28}$$

in terms of the magnetic moment of the lowest Bohr orbit we obtain

$$A_{mn} \cong 10^4 \text{ s}^{-1} . \tag{12.29}$$

The transition probability for a magnetic dipole is thus smaller than that of an electric dipole by a factor of 2092 provided all other parameters of the two dipoles are identical; this is because the typical dipole moment of a magnetic dipole is a factor 46 smaller than that of an equivalent electric dipole.

The Einstein coefficients for transitions in atomic systems in which the electric dipole moment changes are therefore much larger than those for transitions in which only the magnetic dipole moment or the electric quadrupole moment changes.

Electric dipole transitions therefore are referred to as "allowed", while the others are termed "forbidden".

12.4 Simple Solutions of the Rate Equation

In order to compute absorption or emission coefficients in (12.14) and (12.15), both the Einstein coefficients and the number densities N_i and N_k must be known. In the case of LTE, the ratio of N_i to N_k given by the Boltzmann function (12.16) of the local temperature leads to (12.17). If LTE does not apply, the individual processes that lead to the population or depopulation of an energy level have to be considered. Usually such processes involve not only the two levels, giving rise to the transition in question but the whole system of all transitions.

Let R_{jk}^y be the transition probability for the transition $j \to k$ caused by the process y and let N_j be the number density in the state j. Then

$$\frac{dN_j}{dt} = -N_j \sum_k \sum_y R_{jk}^y + \sum_k N_k \sum_y R_{kj}^y.$$

(12.30)

For a stationary situation $dN_j/dt = 0$. Depending on which processes cause the transitions, the solution of (12.30) can be rather complicated. Here we will consider two simple cases; a slightly more complicated situation will be met in the case of radio recombination lines in Chap. 14 and of molecular lines in Chap. 15.

First we consider the case of two states, 1 and 2 where the only way to change states is by the emission and absorption of radiation. We do not, however, assume that LTE applies. The transition rates in (12.30) are given by the Einstein coefficients. For a stationary situation we must have

$$N_1 B_{12} \bar{U} = N_2 (A_{21} + B_{21} \bar{U}),$$

(12.31)

where $\bar{U}$ is given by (12.2). $\bar{U}$ is a single number and can be formally expressed by a radiation density

$$\bar{U} = \frac{4\pi}{c} \bar{I} = \frac{8\pi h v_0^3}{c^3} \frac{1}{\exp\left(\dfrac{h v_0}{k T_b}\right) - 1}$$

(12.32)

resulting in a brightness temperature

$$T_b = \frac{h v_0}{k} \frac{1}{\ln\left(\dfrac{8\pi h v_0^3}{c^3 \bar{U}} + 1\right)}.$$

(12.33)

Combining (12.31) with (12.9), we have

$$\frac{N_2}{N_1} = \frac{B_{12}}{B_{21}} \frac{\bar{I}}{\dfrac{2h v_0^3}{c^2} + \bar{I}} = \frac{g_2}{g_1} \exp\left(-\frac{h v_0}{k T_b}\right).$$

(12.34)

The number densities N_1 and N_2 thus will be described by a Boltzmann distribution as in the case of LTE. The temperature T_b in this distribution describes the radiation density at the frequency corresponding to the transition $2 \to 1$. This need have nothing to do with a thermodynamic temperature of the system.

Another simplification of (12.30) is possible if it is used to describe the number densities of the cool parts of the interstellar gas. Let us assume that only the two lowest states are populated to any extent, but that collisions, not radiation governs the transition rates.

If C_{12} and C_{21} are the collision probabilities for the transitions $1 \to 2$ and $2 \to 1$, respectively, then the rate equation (12.30) for a stationary situation will be

$$N_1 (C_{12} + B_{12} \bar{U}) = N_2 (A_{21} + B_{21} \bar{U} + C_{21}), \tag{12.35}$$

where C_{ij} are the probabilities per particle (in $\mathrm{cm^3\,s^{-1}}$).

$$C_{ik} = N_i C_{ik} = N_i \int_0^\infty \sigma_{ik}(v)\, v\, f(v)\, dv, \tag{12.36}$$

where σ_{ik} is the collision cross section and $f(v)$ the velocity distribution function of the colliding particles. If collisions dominate, the principle of detailed balance leads to

$$\frac{C_{12}}{C_{21}} = \frac{N_2}{N_1} = \frac{g_2}{g_1} \exp\left(-\frac{h\nu_0}{kT_K}\right), \tag{12.37}$$

where T_K is the temperature describing the velocity distribution, $h\nu_0 = E_2 - E_1$ and

$$f(v) = \left(\frac{2}{\pi}\right)^{1/2} v^2 \left(\frac{m_r}{kT_K}\right)^{3/2} \exp\left(-\frac{m_r v^2}{2kT_K}\right), \tag{12.38}$$

with

$$m_r = \frac{m_a m_b}{m_a + m_b}$$

the reduced mass of the colliding particles. T_K is the *kinetic temperature*.

Substituting (12.36) and (12.37) together with (12.32) into (12.35) we obtain

$$\boxed{\begin{aligned}
\frac{N_2 g_1}{N_1 g_2} &= \exp\left(-\frac{h\nu}{kT_{ex}}\right) \\
&= \exp\left(-\frac{h\nu}{kT_b}\right) \frac{A_{21} + C_{21} \exp\left(-\dfrac{h\nu}{kT_K}\right)\left[\exp\left(\dfrac{h\nu}{kT_b}\right) - 1\right]}{A_{21} + C_{21}\left[1 - \exp\left(-\dfrac{h\nu}{kT_b}\right)\right]}
\end{aligned}} \tag{12.39}$$

where we characterize N_2/N_1 by a formal excitation temperature T_{ex} defined

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} \exp\left(-\frac{h\nu}{kT_{ex}}\right). \tag{12.40}$$

This excitation temperature is a mean between the radiation temperature T_b and the kinetic temperature T_K. When T_{ex}, T_b and $T_K \gg h\nu/k$, and if we use for abbreviation

$$T_0 = \frac{h\nu}{k} \tag{12.41}$$

we have

$$T_{ex} = T_K \frac{T_b A_{21} + T_0 C_{21}}{T_K A_{21} + T_0 C_{21}} \tag{12.42}$$

If radiation dominates the rate equation ($C_{21} \ll A_{21}$), then $T_0 \to 0$, (12.39) tends to (12.34) and $T_{ex} \to T_b$. If on the other hand collisions dominate ($C_{21} \gg A_{21}$), then $T_{ex} \to T_K$. Since C_{ik} increases with increasing N collisions will dominate the distribution in high-density situations and the excitation temperature of the line will be equal to the kinetic temperature. In low-density situations $T_{ex} \to T_b$. The density when $A_{21} \approx C_{21} \approx N^* \langle \sigma v \rangle$ is called the *critical density*. The smaller A_{21}, the lower is N^*. We will return to this point in Chap. 14.

Problems

1. Use Eq. (12.4) to estimate T for a two-state system with equal statistical weight factors and level populations $N_1 = 1.01 N_2$ (upper state is 2).

2. We now investigate the variation of T_{ex} with the collision rate, C_{21}, and the spontaneous decay rate, A_{21}, for a two-level system. Equation (12.42) gives the dependence on the kinetic temperature, T_K, and the temperature of the radiation field, T_b, and the ratio of collision rates to A coefficients. Suppose that the collision rate, C_{21}, is given by $n \langle \sigma v \rangle$, where the value of $\langle \sigma v \rangle$ is $\sim 10^{-10}$. When $n \langle \sigma v \rangle = A_{21}$ for the transition involved, this is referred to as the "critical density", n^*. For the 21 cm line, $A_{21} = 2.85 \times 10^{-15}$ s^{-1}. Find n^* for this transition. For neutral hydrogen, in most cases, only two levels are involved in the formation and excitation of the 21 cm line since the $N = 2$ level is 9 eV higher. Less secure is any result for multi-level systems. However, to obtain an order of magnitude estimate, repeat this calculation for the $J = 1 - 0$ transition of the molecule HCO$^+$, modelling the molecule as a two-level system in which the Einstein A coefficient is $A_{21} = 3 \times 10^{-5}$ s^{-1}. What is the value of n^*? Compare this to the value for the 21 cm line. For HCO$^+$, take $T_K = 100$ K; find the value of the local density for which $T_{ex} = 3.5$ K. $T_b = 2.7$ K. For the same density, calculate n^* for the $J = 1 - 0$ transition of the carbon monoxide molecule, CO, modelling this as a two-level system with $A_{21} = 7.4 \times 10^{-8}$ s^{-1}.

3*. Line shapes can be obtained using a semi-classical model of the atom. Use the model of a classical oscillator, but now with a loss term proportional to velocity: $\ddot{x} = -\omega_0^2 x - \gamma \dot{x}$.

(a) Solve for x under the assumption that $\gamma \ll \omega_0$, using $x = x_0 e^{\alpha t}$.

(b) Determine the electric field caused by the motion of the oscillating charge.

(c) Determine the line shape using the Fourier transform (F.T.) of the electric field.

(d) Obtain the line intensity as the absolute value of the square of the F.T. of the electric field. This is the *Lorentzian* line shape.

(e) Determine the line shape if the thermal motion of the atoms, described by $f(v) = (m/2kT)^{3/2} \exp(-mv^2/2kT)$, is combined with the relation for the Doppler shift, $\Delta v/c = \Delta \nu/\nu_0$.

(f) Assume that the areas of these two line profiles are equal, and plot the line shapes. Discuss the difference in intensities of the line wings.

(g) Compare values of γ and ω_0 for the Lyman α line, given that the line frequency, ν, is 3.29×10^{15} s^{-1} and the A coefficient is 5.4×10^9 s^{-1}. Take γ as A, the Einstein coefficient for spontaneous decay. Repeat this for the 1.420 GHz line of hydrogen, emitted by hydrogen atoms in regions of density 1 cm^{-3}, 10^5 cm^{-3} and 10^{19} cm^{-3} if $\gamma = 2.87 \times 10^{15}$ s^{-1}.

4. The energy of the ground state of the hydrogen atom can be obtained using the following analysis, which is closer to the spirit of quantum mechanics than the usual semi-classical orbit analysis. Assume that the nucleus has a very large mass, and charge e. The electron has a mass m and change $-e$. The electron moves with a momentum p at a distance x from the nucleus.

(a) Write down the energy equation for this situation.

(b) Use the relation obtained in Problem 7 in Chap. 2, namely $\Delta x \Delta k = 1$. Use the de Broglie relation $k = p/\hbar$ in the energy equation. Differentiate the energy equation, and set the result to zero to obtain the minimum value of x. What is this value? Compare to the lowest Bohr orbit. Calculate the energy. The value x is the lowest orbit of the electron. The radius increases with n^2, where n is the principal quantum number. Calculate the energy of the lowest two orbits. Now take the difference and set the energy difference equal to $h\nu$. What is the value of ν? Compare this to the frequency of the Lyman α line.

5. Evaluate the constants in Eq. (12.24) to show that

$$A_{\text{ul}} = 1.165 \times 10^{-11} \times \nu_{\text{ul}}^3 |\mu_{\text{ul}}|^2 \qquad (12.43)$$

where ν_{ul} is in GHz, and μ_{ul} is in 10^{-18} esu units.

Chapter 13
Line Radiation of Neutral Hydrogen

Most atomic transitions give rise to spectral lines at wavelengths in the infrared or shorter. With the exception of radio recombination lines (Chap. 14), atomic radio lines are rare. The energy levels are described by the scheme $^{2S+1}L_J$. In this description, S is the total spin quantum number, and $2S+1$ is the multiplicity of the line, that is the number of possible spin states. L is the total orbital angular momentum of the system in question, and J is the total angular momentum. For the lighter elements, the energy levels are best described using LS coupling. This is constructed by vectorially summing the orbital momenta to obtain the total $\boldsymbol{L}$, then combining the spins of the individual electrons to obtain $\boldsymbol{S}$, and then vectorially combining $\boldsymbol{L}$ and $\boldsymbol{S}$ to obtain $\boldsymbol{J}$. If the nucleus has a total spin, $\boldsymbol{I}$, this can be vectorially combined with $\boldsymbol{J}$ to form $\boldsymbol{F}$. For an is olated system, all of these quantum numbers have a constant magnitude and also a constant projection in one direction. Usually the direction is arbitrarily chosen to be along the z axis, and the projected quantum numbers are referred to as M_F, M_J, M_L and M_S.

Compared to the famous 21 cm hyperfine line of H I, most other fine structure and hyperfine structure lines are not very intense. We give a list of the quantum assignments together with line frequencies, Einstein A coefficients and critical densities in Table 13.1. A few comments about this table is in order. First, the 327 MHz line of D I is the deuterium analog of the 21 cm line of H I. There have been intense searches over a 50 year time period. These searches were rewarded with the detection of a weak line; an overview of the results is to be found in Rogers et al. (2007). The measurements were made toward the galactic anticenter where hydrogen and deuterium are found only in the atomic form. The D/H ratios from these data are consistent with Big Bang nucleosynthesis. In other regions, it is likely that much of the D is in the molecule HD. The ^{3}He$^+$ hyperfine transition has been has been studied in a number of galactic sources. The relative abundance of this isotope of helium provides an important observational constraint to standard big bang nucleosynthesis, and is also produced in solar mass stars from D. Unlike D, helium is chemically inert, so cannot be incorporated in molecules. There are radio lines of neutral carbon at 492 and 809 GHz. These lines arise from the somewhat protected molecular regions. In less dense regions, ionized carbon, C$^+$ or C II is present. This ion has a fine structure line at 157 μm and is expected to be a dominant cooling line in denser clouds.

Table 13.1 Parameters of some atomic lines

Element and ionization state	Transition	v/GHz	A_{ij}/s^{-1}	Critical density n^*	Notes
DI	$^2S_{1/2}, F = 3/2 - 1/2$	0.327	4.65×10^{-17}	~ 1	a,b
HI	$^2S_{1/2}, F = 1 - 0$	1.420	2.87×10^{-15}	~ 1	a,b
$^3\text{He}^+$	$^2S_{1/2}, F = 0 - 1$	8.665	1.95×10^{-12}	~ 10	a
CI	$^3P_1 - {}^3P_0$	492.16	7.93×10^{-8}	5×10^2	b
CI	$^3P_2 - {}^3P_1$	809.34	2.65×10^{-7}	10^4	b
CII	$^2P_{3/2} - {}^2P_{1/2}$	1900.54	2.4×10^{-6}	5×10^3	b
OI	$^3P_0 - {}^3P_1$	2060.07	1.7×10^{-5}	$\sim 4 \times 10^5$	b
OI	$^3P_1 - {}^3P_2$	4744.77	8.95×10^{-5}	$\sim 3 \times 10^6$	a,b
OIII	$^3P_1 - {}^3P_0$	3392.66	2.6×10^{-5}	$\sim 5 \times 10^2$	a
OIII	$^3P_2 - {}^3P_1$	5785.82	9.8×10^{-5}	$\sim 4 \times 10^3$	a
NII	$^3P_1 - {}^3P_0$	1473.2	2.1×10^{-6}	$\sim 5 \times 10^1$	a
NII	$^3P_2 - {}^3P_1$	2459.4	7.5×10^{-6}	$\sim 3 \times 10^2$	a
NIII	$^2P_{3/2} - {}^2P_{1/2}$	5230.43	4.8×10^{-5}	$\sim 3 \times 10^3$	a,b

[a] ions or electrons as collision partners.
[b] H_2 as a collision partner.

Although these lines might be considered as part of infrared astronomy, the heterodyne techniques have reached the $1\,000\,\text{GHz} = 1\,\text{THz}$ frequency range ($=300\,\mu\text{m}$).

13.1   The 21 cm Line of Neutral Hydrogen

The interstellar medium pervades the whole galactic system; neutral interstellar gas is present practically everywhere as shown by the 21 cm observations of neutral hydrogen, H I. The structure of this medium is rather irregular. On the one hand, there are large regions with extremely low gas density and, on the other hand, we find large cloud complexes (Fig. 13.1).

This medium is not in a state of equilibrium; violent internal motions are superposed on the general differential galactic rotation field. The physical state of this medium varies markedly from one region to the other because the gas temperature has a dependence on the local energy input and cooling processes. There exist large, cool cloud complexes in which both dust grains and many different molecular species are abundant, and warmer regions where only atomic lines are found.

The 21 cm line is the transition between the hyperfine structure levels $1^2S_{1/2}$, $F = 0$ and $F = 1$ of neutral hydrogen or H I. The energy of these differs slightly due to the interaction of the spin of the nucleus and that of the electron. The frequency of the resulting line has been measured with high precision in the laboratory; it is, in fact, one of the most precisely measured physical quantities with a mean relative error of only 2×10^{-11}:

Fig. 13.1 In the upper diagram, a 21 cm line profile for $l = 41.9°, b = 0°$ measured with the Effelsberg 100 m telescope (8′) is shown, below is a profile of the CO $J = 1 \rightarrow 0$ taken with the same resolution, for comparison. (CO spectrum courtesy of T. Dame, S. Hüttemeister and P. Thaddeus). Note that the CO profile is more structured. This is an indication that CO exists mostly in clouds, while H I is also present in the general interstellar medium
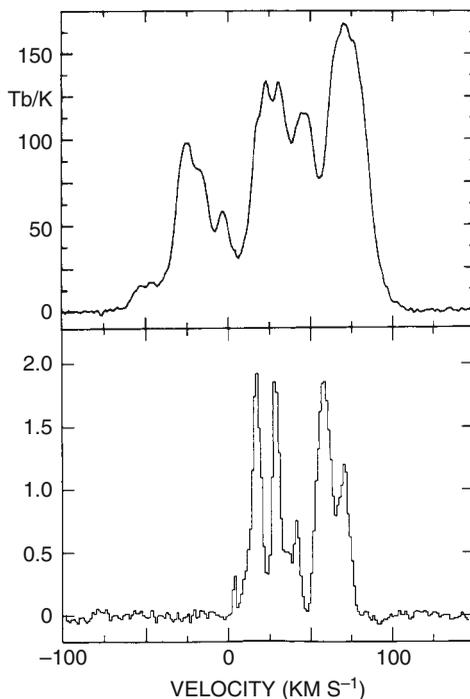
$$\nu_{10} = 1.420\,405\,751\,786(30) \times 10^9\,\text{Hz} \qquad (13.1)$$

(Peters et al. 1965). This radiation is caused by a magnetic dipole transition with a dipole matrix element of one Bohr magneton. Substituting (13.1) into (12.27) we then obtain

$$A_{10} = 2.86888(7) \times 10^{-15}\,\text{s}^{-1}. \qquad (13.2)$$

This transition probability is about a factor 10^{23} smaller than that of an allowed optical transition, mostly due to the difference frequency, which enters as ν^3 in (12.24); an additional factor of 5×10^5 comes from the small magnitude of the magnetic dipole moment.

The spontaneous mean half-life time of the $F = 1$ state is

$$t_{1/2} \cong 1/A_{10} = 3.49 \times 10^{14}\,\text{s} \cong 1.11 \times 10^7\,\text{yr}.$$

Since a typical interstellar hydrogen atom will change the spin of the electron due to collisions about every 400 years, only a very small fraction of all collisions will give rise to the emission or absorption of a photon. Thus in practically all astronomical situations the relative population of the hyperfine structure levels will be determined by collisions.

Let the relative population of the levels be described by an excitation temperature which in this case is usually called the *spin temperature* T_s

$$\frac{N_1}{N_0} = \frac{g_1}{g_0} \exp\left(-\frac{h\,v_{10}}{k\,T_{\mathrm{s}}}\right), \tag{13.3}$$

with

$$T_0 = \frac{h\,v_{10}}{k} = 0.0682\,\mathrm{K}, \tag{13.4}$$

and

$$\frac{N_1}{N_0} = \frac{g_1}{g_0} = 3, \qquad \text{for} \quad T_{\mathrm{s}} \gg T_0.$$

The exponential function in (12.17) can be replaced by the first two terms of the Taylor series, so that

$$\kappa_v = \frac{3\,c^2}{32\pi} \frac{1}{v_{10}} A_{10} N_{\mathrm{H}} \frac{h}{k\,T_{\mathrm{s}}} \varphi(v) \tag{13.5}$$

where the total number of neutral hydrogen atoms (per unit volume) has been introduced by $N_{\mathrm{H}} = N_0 + N_1 = 4N_0$.

Equation (13.5) gives the absorption coefficient per unit frequency interval. Since in radio astronomy the line shapes are usually given in terms of the corresponding Doppler velocities

$$\frac{v_{10} - v}{v_{10}} = \frac{v}{c}, \tag{13.6}$$

(13.5) can be transformed into

$$\begin{aligned}
\mathrm{d}\tau\left(\frac{v}{\mathrm{km\,s^{-1}}}\right) &= -\kappa_v(s)\,\mathrm{d}\left(\frac{s}{\mathrm{cm}}\right)\\
&= -5.4873(10)\times 10^{-19}\left(\frac{N_{\mathrm{H}}}{\mathrm{cm^{-3}}}\right)\left(\frac{T_{\mathrm{s}}(s)}{\mathrm{K}}\right)^{-1}\left(\frac{\varphi(v)}{\mathrm{km^{-1}\,s}}\right)\mathrm{d}\left(\frac{s}{\mathrm{cm}}\right).
\end{aligned} \tag{13.7}$$

If the spin temperature T_{s} is constant along the line of sight, we obtain from (13.7) by integrating both over s and over v

$$\int_{-\infty}^{\infty} \tau(v)\,\mathrm{d}\left(\frac{v}{\mathrm{km\,s^{-1}}}\right) = 5.4873(10)\times 10^{-19}\left(\frac{T_{\mathrm{s}}}{\mathrm{K}}\right)^{-1}\int_{0}^{\infty} N_{\mathrm{H}}(s)\,\mathrm{d}s$$

or

$$\frac{\mathcal{N}_{\mathrm{H}}}{\mathrm{cm^{-2}}} = 1.8224(3)\times 10^{18}\left(\frac{T_{\mathrm{s}}}{\mathrm{K}}\right)\int_{-\infty}^{\infty} \tau(v)\,\mathrm{d}\left(\frac{v}{\mathrm{km\,s^{-1}}}\right) \tag{13.8}$$

if we define the *column density* $\mathcal{N}_{\mathrm{H}}$ by

$$\boxed{\frac{\mathcal{N}_{\mathrm{H}}}{\mathrm{cm}^{-2}} = \int\limits_{0}^{\infty} \left(\frac{N_{\mathrm{H}}(s)}{\mathrm{cm}^{-3}} \right) \mathrm{d}\left(\frac{s}{\mathrm{cm}} \right)} \qquad . \tag{13.9}$$

In general, τ is defined as

$$\tau = -\ln\left(1 - \frac{T_{\mathrm{L}}}{T_{\mathrm{s}} - T_{\mathrm{BG}}} \right), \tag{13.10}$$

where T_{L} is the observed line brightness temperature.

13.2 The Zeeman Effect

From Faraday rotation and a value of the electron density, one can estimate the value
of the line-of-sight component of the interstellar B field (Sect. 3.5). Another estimate
of the line-of-sight B field strength can be obtained from the Zeeman effect using
the magnitude of the frequency shift of the circularly polarized components of an
H I line profile. These two components are the $\Delta F = \pm 1$ transitions and they have
an opposite sense of circular polarization. This polarization and the line frequency
shift allow an unambiguous identification of a Zeeman shifted line. Zeeman mea-
surements are carried out by switching between two senses of circular polarization.
In the switched power spectrum, the result is usually given by the difference of the
intensities of the right and left polarized line intensities, corrected for the instrumen-
tal bandpass. The resulting profile should be "S" shaped if the H I line is split by the
Zeeman effect. Since the shift in frequency is 2.8 Hz per µGauss for the $\lambda = 21$ cm
line, the Zeeman effect measurements are limited by both noise and systematic ef-
fects. More than 30 results have been published; absorption line results are easier
to measure and more reliable, but emission line results are needed to determine the
direction of the magnetic field in a particular H I cloud. Zeeman effects also occur
in molecules with unpaired electron spins.

The most popular molecule for studying the Zeeman effect is OH. The Zeeman
effect is present in both the ground state and in rotationally excited states of OH.
In the OH ground state, the shift is 3.27 Hz per µGauss for the 1.665 GHz line and
1.96 Hz per µGauss for the 1.667 GHz line. We show an energy level diagram of the
OH molecule in Fig. 15.9.

13.3 Spin Temperatures

The excitation temperature for a given transition in a stationary state will be some
average between the brightness temperature describing the ambient radiation field at
the wavelength of the transition considered and the kinetic temperature describing

the local velocity distribution of the colliding particles. Because $T_s \gg T_0$ we can use (12.42)

$$T_s = T_K \frac{T_b A_{10} + T_0 C_{10}}{T_K A_{10} + T_0 C_{10}} \tag{13.11}$$

or

$$\boxed{T_s = \frac{T_b + y T_K}{1 + y}} \quad , \tag{13.12}$$

where

$$\boxed{y = \frac{h \nu_{10}}{k T_K} \frac{C_{10}}{A_{10}}} \quad . \tag{13.13}$$

The y values are given in Table 13.2. T_s is thus a weighted mean of the kinetic gas temperature and the brightness temperature of the radiation field. The weighting factor y depends on the collision probabilities of the colliding partners H I – H I and H I – e which have to be computed by quantum mechanical methods. A survey of the methods and the results are given by Purcell and Field (1956), Field (1958) and Elwert (1959) show that for gas with the density $N_H > 1\,\text{cm}^{-3}$, $T_s \cong T_K$ is always true irrespective of whether the gas is mainly neutral or ionized, and this also applies for low-density gas $(N_H < 0.1\,\text{cm}^{-3})$ if it is partly ionized. Therefore we can safely adopt $T_s = T_K$ for neutral hydrogen gas. Note that the weighting factor y given by Kulkarni and Heiles (1988) is defined as the inverse of (13.13). Although their definition is more plausible in some ways than that used here, we have preferred to keep the definition introduced originally by Purcell and Field (1956) which is used throughout the literature.

The spin temperature T_s will quite often vary with s because the line-of-sight intersects clouds of different kinetic temperature. Then it is possible to determine an average spin temperature which, according to Kahn (1955), will be the harmonic mean value of the temperatures encountered. Let

Table 13.2 Weighting factors for the determination of the spin temperature of neutral hydrogen

$\dfrac{T_K}{K}$	$y_H \Big/ \dfrac{N_H}{\text{cm}^{-3}}$ [a]	$y_e \Big/ \dfrac{N_e}{\text{cm}^{-3}}$ [b]
1	1200.0	6700
3	490.0	3900
10	190.0	2100
30	85.0	1200
100	35.0	650
300	16.0	350
1000	6.7	130
3000	3.9	66
10000	1.3	18

[a] computed for collisions with neutral hydrogen atoms.
[b] computed for collisions with electrons.

$$N_v(s) = N_{\mathrm{H}}(s)\,\varphi(v|s)$$

be the space density of neutral hydrogen atoms with velocities between v and $v + dv$ at the position s. Equation (13.7) then can be integrated from 0 to s yielding

$$\tau_v(s) = 5.4873 \times 10^{-19} \int_0^s \frac{N_v(s)}{T_{\mathrm{s}}(s)}\,\mathrm{d}s$$

$$= 5.4873 \times 10^{-19} \left\langle \frac{1}{T_{\mathrm{s}}(v)} \right\rangle \int_0^s N_v(s)\,\mathrm{d}s\,,$$

where $\langle 1/T_{\mathrm{s}} \rangle$ is the appropriate mean value of the inverse spin temperature

$$\left\langle \frac{1}{T_{\mathrm{s}}(v)} \right\rangle = \frac{\displaystyle\int_0^s \frac{N_v(s)}{T_{\mathrm{s}}(s)}\,\mathrm{d}s}{\displaystyle\int_0^s N_v(s)\,\mathrm{d}s}\,. \tag{13.14}$$

This average is a weighted harmonic mean value where the neutral hydrogen gas density is the weighting factor and, since N_v depends on the velocity, the harmonic mean spin temperature depends on v also.

13.4 Emission and Absorption Lines

We now will estimate $\mathscr{N}_{\mathrm{H}}$ from 21 cm measurements. Consider an isothermal cloud of gas in front of some background source. The solution of the equation of radiation transfer (1.37) in terms of the brightness temperature is then

$$T_{\mathrm{b}}(v) = T_{\mathrm{s}}\left[1 - \mathrm{e}^{-\tau(v)}\right] + T_{\mathrm{c}}\,\mathrm{e}^{-\tau(v)}, \tag{13.15}$$

where T_{s} is the spin temperature of the cloud, T_{c} the brightness temperature of the background source, and $\tau(v)$ the optical depth of the cloud at the radial velocity v. For positions without a background source, $T_{\mathrm{c}} = 0$ and we observe a pure *emission line profile*. If $\tau(v) \ll 1$, quadratic and higher terms in the Taylor series $\mathrm{e}^{-\tau} = 1 - \tau + \tau^2/2 - \ldots$ can be neglected resulting in

$$T_{\mathrm{b}}(v) = T_{\mathrm{s}}\,\tau(v) \qquad \text{for } \tau(v) \ll 1. \tag{13.16}$$

Substituting this into (13.8) we find

$$\frac{\mathcal{N}_H}{cm^{-2}} = 1.8224(3) \times 10^{18} \int\limits_{-\infty}^{\infty} \left(\frac{T_b(v)}{K}\right) d\left(\frac{v}{km\,s^{-1}}\right) \qquad (13.17)$$

where T_b is the main beam brightness temperature and v is the radial velocity. Thus, for optically thin radiation the column density is independent of the spin temperature of the gas. N_s can be determined unambiguously from the integral over the emission line. If the optical depth is not very small, one can correct the column density using the factor F, where

$$F = \frac{\tau}{1 - e^{-\tau}}.$$

This is also treated in regard to molecular lines in (15.38).

While (13.17) is valid only for optically thin radiation, (13.8) is applicable irrespective of the value of τ. But then the column density depends critically on the adopted value for T_s. Solving (13.15) with $T_c = 0$ for τ and substituting this into (13.8) we find that

$$\frac{\mathcal{N}_H}{cm^{-2}} = -1.8224(3) \times 10^{18} \left(\frac{T_s}{K}\right) \int\limits_{-\infty}^{\infty} \ln\left[1 - \frac{T_b(v)}{T_s}\right] d\left(\frac{v}{km\,s^{-1}}\right). \quad (13.18)$$

Precise measurements of T_s for an actual cloud of gas are rather difficult. Limits can be estimated from (13.15): $T_b(v) \rightarrow T_s$ as $\tau(v) \rightarrow \infty$ (Fig. 13.2). This is the basis for the "classical" value $T_s = 125$ K reported by Dutch radio astronomers. This harmonic mean depends sensitively on the relative amount of low- and high-temperature gases.

Determinations of parameters for the interstellar gas are more accurate if there is a background source with a small angular size, measured by a large telescope. But the relationship between the source geometry and the telescope beam plays a role.

13.4.1 The Influence of Beam Filling Factors
and Source Geometry

The following formalism provides the detailed analysis needed to understand spectral line transfer for homogeneous clouds. Whether a given line is seen in emission or in absorption depends critically on geometric factors. If the telescope beam is small compared to the cloud size, which is small compared to the size of the background source, the observed line temperature T_L is given by

$$T_L = \frac{c^2}{2k\,v_{10}^2} \left[B_v(T_s) - B_v(T_C)\right] (1 - e^{-\tau}), \qquad (13.19)$$
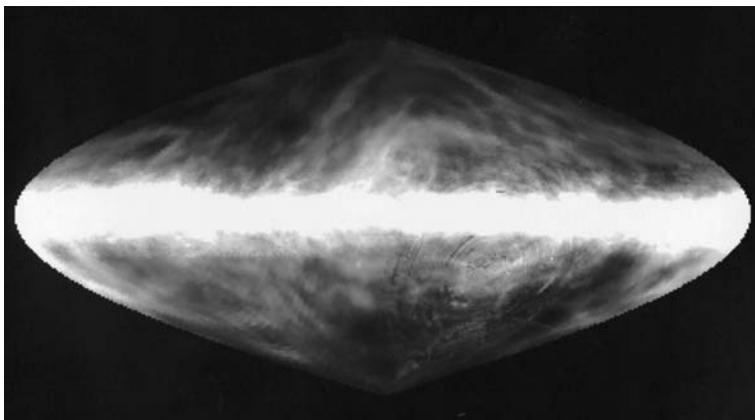
Fig. 13.2 The brightness temperature T_b of galactic H I over the entire sky. This emission has been integrated over the radial velocity [after Dickey and Lockman (1990)]. Recently a survey with 36 arcmin angular resolution, the Leiden/Argentine/Bonn survey, has been published by Kalberla et al. (2005)

where T_C is the true brightness temperature of the continuum source, T_s is the spin temperature and τ is the optical depth. The physics of the line enters only through T_s and τ. In all following considerations, we assume $T_C \gg 2.7\,\text{K}$, that is, we will neglect the presence of the microwave background radiation. Using the Rayleigh-Jeans approximation this equation becomes

$$T_L = (T_s - T_C)(1 - e^{-\tau}). \tag{13.20}$$

Now we will relax the assumption about filling factors, f, for both the line radiation and the continuum source. Consider a cloud of the size Ω_{cl} and a continuum source Ω_C only partly filling the telescope beam Ω_B. The continuum source is situated behind the cloud and partially covered by it.

In this general case, one has:

$$T_L = f_{cl}\, T_s\,(1 - e^{-\tau}) - f_0\, f_C\, T_C\,(1 - e^{-\tau}) \tag{13.21}$$

$$= (f_{cl}\, T_s - f_0\, f_C\, T_C)(1 - e^{-\tau}). \tag{13.22}$$

Here, f_{cl} and f_C are the beam filling factors of the cloud and continuum source, respectively, and f_0 denotes the fraction of the continuum source covered by the cloud. The relations $0 \leq f_{cl}, f_C, f_0 \leq 1$ must be fulfilled. If both the sources and the beams are Gaussian shaped, f_{cl} and f_C are given by

$$f_{\text{source}} = \frac{\Omega_{\text{source}}}{\Omega_{\text{source}} + \Omega_{\text{beam}}}, \tag{13.23}$$

where Ω_{source} and Ω_{beam} denote the solid angles covered by the source and the beam, respectively. The influence of the different factors become more obvious when one considers special cases.

1) *Optical Depth*. For an optically thin line (13.21) becomes

$$T_L = (f_{cl} T_s - f_0 f_C) \tau, \tag{13.24}$$

while for an optically thick line one has

$$T_L = f_{cl} T_s - f_0 f_C T_C. \tag{13.25}$$

2) *Degree to which the continuum source is covered*. If the continuum source is not covered by the cloud, is in front of the cloud or if there is no continuum source, one has $f_0 = 0$ and

$$T_L = f_{cl} T_s (1 - e^{-\tau}). \tag{13.26}$$

Then the line will always appear in emission. If, on the other hand, the continuum source is completely covered by the cloud ($f_0 = 1$), one has

$$T_L = (f_{cl} T_s - f_C T_C)(1 - e^{-\tau}) \tag{13.27}$$

and the probability of absorption dominating emission is largest.

3) *Source size*. In the simplest case possible, both the line radiation emitting cloud and the continuum source are significantly larger than the telescope beam ($\Omega_{cl}, \Omega_C \gg \Omega_B$). All filling factors approach unity and we recover (13.20). If absorption lines are observed and there is reason to assume $T_s \ll T_C$, it is possible to calculate optical depths from the ratio of the observed line temperatures and the continuum temperature by using

$$\tau \equiv \tau_{\text{app}} = -\ln\left(1 - \frac{|T_L|}{T_C}\right). \tag{13.28}$$

This method works best if the optical depths are fairly low. In the case of high τ, T_L/T_C will be very close to unity. Then the exact value of τ becomes fairly uncertain. If the cloud has no fine structure and is larger that the beam, but the continuum source is smaller ($\Omega_{cl} \gg \Omega_B > \Omega_C \Rightarrow f_{cl} = 1, f_0 = 1$), (13.21) becomes

$$T_L = (T_s - T_{MC})(1 - e^{-\tau}) \tag{13.29}$$

with the main beam continuum brightness temperature defined as $T_{MC} \equiv f_C T_C$. Emission is now very likely.

If the beam size is larger than both Ω_C and Ω_{cl}, without further assumptions we get the general case as described by (13.23). Assuming, however, $\Omega_{cl} = \Omega_C$ and for simplicity $f_0 = 1$, resulting in $f_{cl} = f_0 f_C = f$, (13.21) becomes

$$T_L = f(T_s - T_C)(1 - e^{-\tau}). \tag{13.30}$$

Now absorption is more likely. Finally, Ω_{cl} may be smaller that both Ω_C and Ω_B. One gets

$$0 \le f_0 \le \frac{f_{cl}}{f_C} \tag{13.31}$$

and if $f_0 = f_{cl}/f_C$, that is, if the line emitting cloud is completely in front of the larger continuum source, (13.21) becomes

$$T_L = f_{cl}(T_s - T_C)(1 - e^{-\tau}). \tag{13.32}$$

Of course, many other situations are possible but these are combinations of the simple cases shown here: sometimes there will be several gas clumps with different properties and geometries present in the beam. The only possibility to disentangle such a mixture is if the clouds have different radial velocities or positions.

13.5   The Physical State of the Diffuse Interstellar Gas

The physical properties of the neutral hydrogen regions are widely different. However the H I line is always excited by collisions. This permits a great simplification in the analysis of H I line radiation. The mean free path for neutral hydrogen atoms is

$$\ell = (N\pi a_0^2)^{-1} \approx 10^{16}\left[\frac{N}{cm^{-3}}\right]^{-1} cm \approx \frac{1}{300}\left[\frac{N}{cm^{-3}}\right]^{-1} pc \tag{13.33}$$

and the mean free time, τ, between collisions for hydrogen atoms in a gas with a Maxwellian velocity distribution corresponding to a temperature T is given by

$$\tau \approx 10^{12}\left[\frac{T}{K}\right]^{-1/2}\left[\frac{N}{cm^{-3}}\right]^{-1} s \approx 3.2 \times 10^4 \left[\frac{T}{K}\right]^{-1/2}\left[\frac{N}{cm^{-3}}\right]^{-1} years. \tag{13.34}$$

For $10^{-3} < N/cm^{-3} < 1$ and $10 < T/K < 3\,000$ we thus find $1/3\,000 < \ell/pc < 3$ and $1\,000 < \tau/yr < 6 \times 10^5$ so that we can safely adopt a single Maxwellian distribution for each volume element of the interstellar gas, and that a single kinetic temperature is sufficient to describe the gas at each position.

These considerations can have far reaching consequences. An example shows this: The large magellanic cloud (LMC) is a dwarf galaxy that is a close satellite of the galaxy with a distance of only about 50 kpc. It is a barred irregular system with a flat disk seen nearly face-on with a tilt of $i \simeq 33° \pm 6°$. However, we encounter problems if we try to incorporate the observed line profile shapes for H I in the LMC into this simple model. Surveys of the 21 cm line in the LMC show double-peaked line profiles for a large percentage of the positions. From a detailed analysis of these data one can associate one of these peaks with a flat, gaseous disk in differential

rotation extending over all of the LMC, while the other peak arises from another gas sheet at lower radial velocity extending over roughly 40% of the cloud area. Both structures are large-scale features compared to the main beam of the telescope. If the disk model of the LMC was also applicable to the H I gas, then the two gas features, the so-called D and the L components, would occupy the same volume. But due to the collision time argument this is not possible. Thus, given their large angular extent, only a line-of-sight superposition is possible. The gas of the LMC therefore cannot all belong to a flat disk; the L component must be situated either in front or behind. In any case, the gas is distributed in three-dimensions and not only a disk.

The value of the local kinetic gas temperature will be determined by a balance between energy gain and loss. For monatomic gas the thermal energy is given by $\frac{3}{2}kT_kN$, so that this balance is governed by

$$N\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{3}{2}kT_k\right) - kT_k\frac{\mathrm{d}N}{\mathrm{d}t} = \Gamma - \Lambda \tag{13.35}$$

where Γ is the energy per cm^{-3} gained, and Λ the corresponding loss function. An extensive discussion of the different processes involved can be found in the review of observations and theory by Kulkarni and Heiles (1988). Below, we give a short summary.

Cooling of the gas occurs by collisional excitation of fine structure transitions. The excitation energy is subsequently lost from the gas by spectral line emission, provided the gas is transparent. Neutral hydrogen itself contributes little to the cooling for $T < 10^4$ K; in this temperature range rarer heavy elements in the interstellar medium are much more efficient. In particular the 157.7 μm fine structure line of singly ionized carbon is usually dominant, but depletion of carbon on dust grains may complicate the situation, so cooling lines of other elements may be important.

Heating processes convert external energy into thermal motion of the gas particles. There are many different heating mechanisms. On a microscopic scale, heating by ionization is probably most important. If a bound electron is stripped off an atom or grain it is usually ejected with excess kinetic energy, and this then is shared with the other gas particles by collisions. The ionization could be caused by radiation; then the frequency is of paramount importance. Another mechanism is heating by high-energy particles. Depending on which process is considered, different properties of the ISM are of importance, but in any case the detailed theory is fairly complicated.

On a macroscopic scale a heating source might be the collision of gas clouds or streams with different velocities. There are many situations conceivable in which such collisions could happen – stellar winds, collisions with supernova ejecta or high-velocity clouds are only some examples. The observed line width of the 21 cm line emission usually is ~ 6–9 km s^{-1}. This value is much too high to be interpreted as thermal in origin; thus one must consider the processes mentioned above. In any case, the heating depends on gas density, while cooling depends on density squared. Thus, in a steady state solution of (13.35), T_k will depend on density.

Observations of the galactic interstellar medium show a wide number of components. Certainly these are not in thermodynamic equilibrium and it is often not even certain whether pressure equilibrium can be adopted. We can divide the phenomena into four classes according to temperature and degree of ionization.

CNM (cold neutral medium) appears as narrow absorption features in the 21 cm line spectrum in front of strong galactic sources. Comparing emission and absorption, a gas temperature $T < 50\,\text{K}$ and a high-volume gas density ($N > 1 - 10\,\text{cm}^{-3}$) is found. At higher local densities and column densities larger than $10^{20}\,\text{cm}^{-2}$ the H I is converted to H_2 on grains (see Chap. 15).

WNM (warm neutral medium). Estimating upper limits for the optical depth of 21 cm line radiation it can be shown that a considerable part of the gas must have $T > 200\,\text{K}$. This gas usually has emission with a large line width, it is present practically everywhere near the galactic plane.

WIM (warm ionized medium): sensitive observations of H_α emission show this to have a fairly widespread distribution in the galaxy not confined to the vicinity of hot stars. Therefore ionized hydrogen gas with $T_k \approx 10^4$ K must be present in large portions in the interstellar medium.

HIM (hot ionized medium): the detection of diffuse soft X-ray emission and of an O VI absorption line in the UV requires the presence of yet another hot ($T \approx 10^6$ K) gas component in the ISM.

The distribution, physics and evolutionary history of the interstellar medium forms one of the most active branches of present day astrophysics, that incorporates observations of many branches of observational astronomy as well as theoretical astrophysics. At this moment there is no comprehensive model which incorporates all major observational facts, so each case must be considered individually.

13.6   Differential Velocity Fields and the Shape of Spectral Lines

Velocity fields in a line emitting gas will affect the appearance of this radiation in a number of different ways. The bulk velocity of the gas will shift the mean frequency; random velocities of the gas atoms will also influence the line shapes. In H I, line emission arises from a large volume so that large-scale velocity gradients are of great importance.

Let us assume that neutral hydrogen gas has a bulk velocity $U(s)$ that is a function of the position s along the line of sight. For the sake of simplicity we will assume that the local line shape does not depend on s. Then the line-of-sight element ds contributes $d\tau$ at the velocity v according to (13.7) and

$$d\tau(v) = -w \frac{N_H(s)}{T_K(s)} \, \varphi\left[v - U(s)\right] ds. \qquad (13.36)$$

The coefficient w depends on the units used for s and v; if s is in cm and v in $\text{km}\,\text{s}^{-1}$ we have $w = 5.4873(10) \times 10^{-19}$, while $w = 1.6932$ for s in pc and v in

km s^{-1}. For the total optical depth at the velocity v for gas between 0 and s we therefore obtain

$$\tau(v,s) = w \int\limits_0^s \frac{N_H(x)}{T_K(x)} \varphi\left[v - U(x)\right] dx \tag{13.37}$$

or

$$\tau(v,U) = w \int\limits_{U(0)}^{U(s)} \frac{N_H[s(U)]}{T_K[s(U)]} \varphi\left[v - U\right] \frac{dU}{\left|\dfrac{dU}{ds}\right|} \tag{13.38}$$

if

$$U = U(s). \tag{13.39}$$

A simple example illustrates the use of these expressions. We assume that a Gaussian line shape is emitted in a homogeneous medium with a *quadratic velocity field*:

$$U(s) = U_c + b\,(s - s_c)^2\,;$$

then for $N_H = $ const, $T_K = $ const and $s \to \infty$ we have

$$\tau(v) = \frac{wN_H}{T_K} \left[\int\limits_0^{s_c} \varphi\left[v - U(x)\right] dx + \int\limits_{s_c}^{\infty} \varphi\left[v - U(x)\right] dx \right] = I + II\,.$$

For the range of the integral I we find

$$s = s_c - \frac{1}{\sqrt{|b|}} |U - U_c|^{1/2}\,,$$

while for II we have

$$s = s_c + \frac{1}{\sqrt{|b|}} |U - U_c|^{1/2}\,,$$

and therefore

$$I = -\frac{1}{2} \frac{wN_H}{T_K} \frac{1}{\sqrt{|b|}} \int\limits_{U_c + bs_c^2}^{U_c} \varphi(v - U) \frac{dU}{\sqrt{|U - U_c|}}$$

and

$$II = \frac{1}{2} \frac{wN_H}{T_K} \frac{1}{\sqrt{|b|}} \int\limits_{U_c}^{\infty} \varphi(v - U) \frac{dU}{\sqrt{|U - U_c|}}\,.$$

Assuming

$$\varphi(v) = \frac{1}{\sigma\sqrt{2}} \exp\left(-\frac{v^2}{2\sigma^2}\right)$$

and

$$b s_c^2 \gg \sigma$$

so that we can adopt effectively $U_c + b s_c^2 \rightarrow \infty$, *I* and *II* can be taken together and we obtain

$$\tau(v) = \frac{w}{\sqrt{2\pi}} \frac{N_H}{\sigma T_K} \frac{1}{\sqrt{|b|}} \int\limits_{U_c}^{\infty} \frac{1}{\sqrt{|U - U_c|}} \exp\left[-\frac{(v-U)^2}{2\sigma^2}\right] dU .$$

Substituting

$$\sigma x = U - U_c$$

we find

$$\tau(v) = \frac{w}{\sqrt{2\pi}\sigma} \frac{N_H}{T_K} \frac{1}{\sqrt{|b|}} P\left(\frac{v - U_c}{\sigma}\right) \qquad\qquad (13.40)$$

$$P(\xi) = \int\limits_0^{\infty} \frac{1}{\sqrt{x}} \exp\left[-\frac{1}{2}(\xi - x)^2\right] dx \qquad\qquad (13.41)$$

For $\xi < 1$ the shape of $P(\xi)$ is quite similar to a shifted Gaussian. The values are given in Table 13.3 (Fig. 13.3) since they will be used to describe the shape of the 21 cm line emission near the galactic equator close to the radial velocities that are

Table 13.3 The function $P(\xi) = \int\limits_0^{\infty} \frac{1}{\sqrt{x}} \exp\left[-\frac{1}{2}(\xi - x)^2\right] dx$

ξ	P	ξ	P	ξ	P
−5.0	0.0000	−1.0	0.9015	1.0	2.5263
−4.0	0.0003	−0.8	1.1467	1.2	2.4532
−2.8	0.0110	−0.6	1.4096	1.4	2.3516
−2.6	0.0202	−0.4	1.6762	1.6	2.2333
−2.4	0.0358	−0.2	1.9304	1.8	2.1095
−2.2	0.0610	0.0	2.1560	2.0	1.9893
−2.0	0.1002	0.2	2.3388	2.2	1.8793
−1.8	0.2414	0.4	2.4690	2.4	1.7835
−1.6	0.3545	0.6	2.5420	2.6	1.7026
−1.4	0.5014	0.8	2.5591	2.8	1.6349
−1.2	0.6844	1.0	2.5263	3.0	1.5758

Fig. 13.3 The function $P(\xi)$
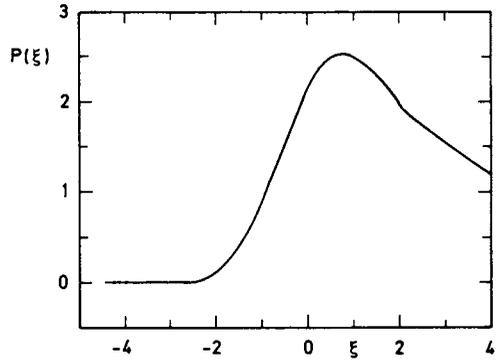
measured at the tangential or subcentral points in the longitude range $270° < l < 360°$ and $0° < l < 90°$.

13.7   The Galactic Velocity Field in the Interstellar Gas

Neutral hydrogen gas is one of the main constituents of the interstellar medium; it is distributed over the whole Galaxy. Since H I is thermalized at low densities, it should be always observable, so the 21 cm line forms an almost ideal tool for the study of galactic kinematics. Its only shortcomings are the problems often encountered when the exact distance of features must be determined. Therefore, models are needed to provide the possible structure; however, these cannot give a conclusive solution.

The large-scale kinematics of galactic interstellar gas are governed by galactic rotation, except for non circular motions close to the galactic center. In addition, the velocity field may be perturbed by streaming velocities, and on a more local scale, supernova explosions which may introduce irregularities in the velocity field. Here we will describe the large scale field and how it influences the radial velocity and shape of the line radiation.

For the sake of simplicity we will assume that all motions are axially symmetric if seen from the galactic center. If $\Theta(r)$ is the (linear) rotational velocity at the galactic radius r, and $\Pi(r)$ the corresponding motion along r as defined by

$$\Theta(r) = r\Omega(r) \quad \text{and} \quad \Pi(r) = rH(r) \tag{13.42}$$

respectively, and if $\Omega(r)$ is the angular velocity and $H(r)$ is the expansion rate, then the radial velocity of a point P relative to the local standard of rest is

$$v_r = \Theta(r)\sin(l+\vartheta) - \Theta(r_0)\sin l - \Pi(r)\cos(l+\vartheta) + \Pi(r_0)\cos l .$$

Now

$$r_0 \sin l = r \sin(l + \vartheta) \quad \text{and} \quad r \cos(l + \theta) = r_0 \cos l - s$$

so that

$$v_r = r_0 \left[\Omega(r) - \Omega(r_0) \right] \sin l - r_0 \left[H(r) - H(r_0) \right] \cos l + H(r) s \qquad .$$

(13.43)

This is the law of differential galactic rotation. Experience has shown that $H(r_0)$ is very small in the solar neighborhood so that the velocity field is well described by pure rotation.

Another observational result is that $\Theta(r)$ varies very slowly with r outside the immediate surroundings of the galactic center. Therefore the series expansion

$$\Theta(r) = \Theta(r_0) + \left. \frac{d\Theta}{dr} \right|_0 (r - r_0) \tag{13.44}$$

should be a good representation of $\Theta(r)$ in the solar neighborhood. Now, from

$$r^2 = r_0^2 + s^2 - 2 r_0 s \cos l, \tag{13.45}$$

the approximations

$$r = r_0 \left(1 - \frac{s}{r_0} \cos l \right) \quad \text{and} \quad \frac{1}{r} = \frac{1}{r_0} \left(1 + \frac{s}{r_0} \cos l \right)$$

are obtained so that

$$v_r = r_0 \left[\Theta_0 \left(\frac{1}{r} - \frac{1}{r_0} \right) - \left. \frac{d\Theta}{dr} \right|_0 \frac{s}{r_0} \cos l \right] \sin l$$

$$= \left(\frac{\Theta_0}{r_0} - \left. \frac{d\Theta}{dr} \right|_0 \right) s \cos l \sin l,$$

and

$$v_r = s A(r_0) \sin 2l \tag{13.46}$$

when using

$$A(r_0) = \frac{1}{2} \left(\frac{\Theta_0}{r_0} - \left. \frac{d\Theta}{dr} \right|_0 \right) = -\frac{1}{2} r_0 \left. \frac{d\Omega}{dr} \right|_0 \qquad . \tag{13.47}$$

This is the famous Oort $\sin 2l$ relation describing the differential galactic rotation in the solar neighborhood. For given l, v_r is proportional to s, but obviously this is

Fig. 13.4 Radial velocity in
km s^{-1} caused by differential
galactic rotation and
expansion for $l = 20°$. The
abscissa is the distance
$\Delta s = s - s_c$ from the
subcentral point defined by
(13.49)
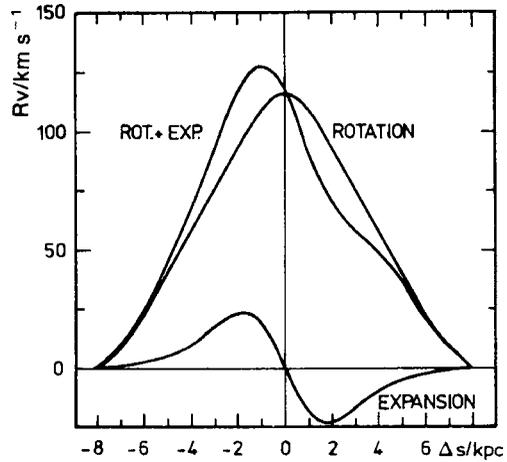
valid only close to the sun. In the first galactic quadrant $(0° < l < 90°)\, v_r$ reaches
a maximum (see Fig. 13.4) and in the fourth quadrant $(270° < l < 360°)$ a mini-
mum. This can be shown formally in the following way: Along a given line of sight
$(l = \text{const})$

$$\frac{dv_r}{ds} = \frac{dv_r}{dr}\frac{dr}{ds}$$

and from (13.43) with $H \equiv 0$ using (13.47)

$$\frac{dv_r}{dr} = r_0 \left.\frac{d\Omega}{dr}\right|_0 \sin l = -2A(r_0)\sin l, \tag{13.48}$$

while (13.45) gives

$$\frac{dr}{ds} = \frac{s - r_0 \cos l}{r} = \frac{s - s_c}{r}$$

so that

$$\frac{dv_r}{ds} = -2A(r_0)\sin l\, \frac{s - s_c}{r}.$$

Therefore $dv_r/ds = 0$ for

$$s = s_c = r_0 \cos l \quad \text{and} \quad r_c = r_0|\sin l|. \tag{13.49}$$

The measured radial velocity thus adopts an extreme value; at the subcentral
point, the radial velocity from (13.43) is

$$v_{\rm c} = \left[\Theta(r_0 |\sin l|) - \Theta_0 |\sin l| \right] \frac{\sin l}{|\sin l|} . \tag{13.50}$$

This relation can be used to construct the rotation curve point by point for $r < r_0$ from the extremes of the measured radial velocity for each longitude. Equation (13.49) gives the galactic position of the gas that emitted this radiation, and

$$\Theta(r_0 |\sin l|) = v_{\rm c} \frac{\sin l}{|\sin l|} + \Theta_0 |\sin l| . \tag{13.51}$$

If the series expansion (13.44) is introduced into (13.50) we obtain using (13.47)

$$v_{\rm c} = 2A(r_0)\, r_0 \left(1 - |\sin l|\right) \frac{\sin l}{|\sin l|} \qquad . \tag{13.52}$$

This relation is a good approximation for the conditions in our galaxy for $|\sin l| > 0.5$, that is, for $r > r_0/2$, and so a linear function in this range must be a good approximation for $\Theta(r)$, (see Fig. 13.5). Most investigations of the galactic velocity field use a formula differing slightly from (13.52), but it was shown by Gunn et al. (1979) that (13.52) is the correct expression for galaxies with almost flat rotation curves.

Close to the subcentral point (where gas is closest to the center) the run of radial velocity along the line of sight will not be represented well by a linear function; higher terms of the appropriate Taylor series have to be included.
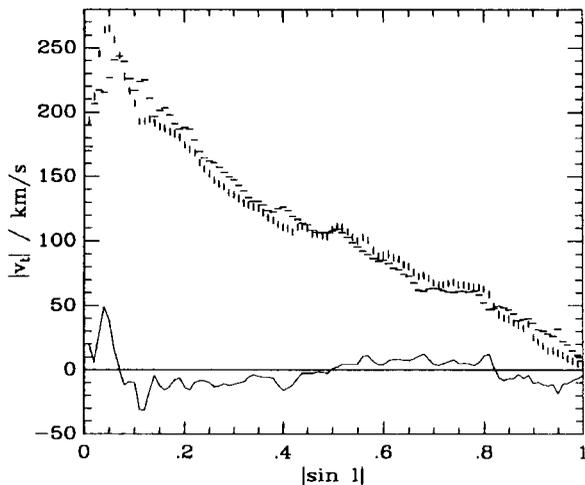
Fig. 13.5 Terminal velocity as function of $|\sin l|$. "|" for 1st galactic quadrant $0° < l < 90°$; "–" for 4th galactic quadrant $270° < l < 360°$, the continuous line shows the difference of the two

Let

$$v_r(s) = v_c + \frac{dv}{ds}\bigg|_c (s - s_c) + \frac{1}{2}\frac{d^2v}{ds^2}\bigg|_c (s - s_c)^2 + \dots ;$$

then from (13.48), changing the suffix 0 to c we obtain,

$$\frac{d^2v}{ds^2} = -\frac{2r_c}{r^2}A(r) - (s - s_c)\frac{d}{ds}\left(\frac{2r_c}{r^2}A(r)\right).$$

Substituting this into the Taylor series we find for terms up to and including $(s - s_c)^2$:

$$\boxed{v_r(s) = v_c - \frac{A(r_c)}{r_c}(s - s_c)^2} \qquad . \tag{13.53}$$

The large-scale velocity field of galactic rotation thus appears as a linear velocity field with a $\sin 2l$ longitudinal dependence (13.46) in the solar neighborhood, while it shows up as a quadratic velocity field (13.53) in the vicinity of the subcentral point. Not only will such a velocity field influence the observed frequency of the line, but it will also have an effect on the line shape. The peak optical depth for the local gas should have its lowest value in longitudes around $l \cong 45°$ with

$$\tau_{max} \lesseqgtr w\frac{N_H}{T_s}\frac{1}{A}, \tag{13.54}$$

while $\tau_{max} \to \infty$ near $l \cong 90°$. Indeed the largest brightness temperatures in the local gas are observed near $l = 70°$ to $75°$ where values of $T_b \cong 125$ K are reached. If we therefore observe a local value of $T_b \cong 45$ K near $l = 45°$ the optical depth should be less than one in these regions resulting in

$$T_b = T_s\,\tau = w\frac{N_H}{A}. \tag{13.55}$$

With $A = 15\,\mathrm{km\,s^{-1}\,kpc^{-1}}$ this results in $N_H = 0.4\,\mathrm{cm^{-3}}$.

Because we can assume such an optically thin situation, the results thus obtained remain correct even if, in fact, the interstellar gas consists of several different phases, each with its own density and spin temperature. For the local gas the appropriate average values are $N_H = 0.4\,\mathrm{cm^{-3}}, T_s = 125$ K.

13.8   Atomic Lines in External Galaxies

The study of the distribution and motion of interstellar gas in external galaxies is mostly based on H I, although mapping of the rotational lines of the CO molecule will complement these data. For a galaxy at a distance D in Mpc, the total mass of H I in a beam of θ arcsec, if the line is optically thin, is (from (13.17):

$$M_{HI} = 0.39 \times 10^2 \, D^2 \, \theta^2 \int T_{MB} \, dv \qquad (13.56)$$

or

$$M_{HI} = 2.36 \times 10^5 \, D^2 \int S_v \, dv, \qquad (13.57)$$

where the line is integrated over velocity in $km\,s^{-1}$ and S_v is in Jy. For rotation curves, one must take into account the inclination and the motions in the galaxy. We show a sketch in Figs. 13.6 and 13.7. The galaxy is assumed to have a systemic velocity v_0 and non circular velocities z and ω, and the disk of this (planar) galaxy is inclined at an angle i. Then the observed radial velocity of the galaxy, v_0, is

$$v_0(r, \phi) = z \cos i + \omega \sin i \sin \phi + \theta(r, \phi) \sin i \cos \phi. \qquad (13.58)$$

From the (model) rotation curve in Fig. 13.7, one obtains the result, which would be observed with a very small beam and a noiseless receiver.

Mapping results are restricted to relatively close galaxies. A short summary of mapping results shows that: (1) the H I is not centrally concentrated, (2) the H I extent in isolated spiral galaxies is larger than the optical extent, (3) many spiral galaxies show distortions in the distribution of the H I in their outer regions, (4) H I links and bridges have been found between galaxies whose separations are a few tens of kiloparsecs, (5) in most spiral galaxies, the rotation curve appears to be flat to large distances, from the center, while the optical light decreases exponentially (in interacting systems the rotation curves appear to decline, however), and (6) galaxies in groups (for example, the Virgo cluster or even in small groups) show deficiencies of H I as compared to galaxies of the same type outside groups.
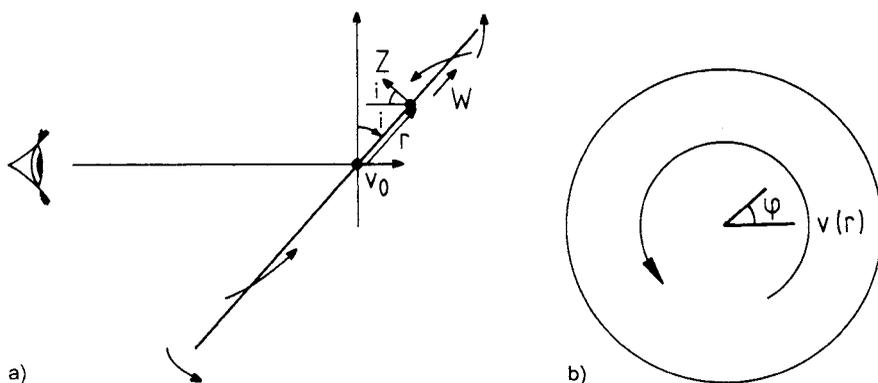
Fig. 13.6 (**a**) A side-view sketch to illustrate the geometry used in obtaining the standard relation for rotation curves of galaxies, (**b**) a face-on view of the same geometry

Fig. 13.7 The differential
velocity field of a flat, inclined
model galaxy with solid body
rotation for the inner parts
and a flat rotation curve in
the outer regions. The solid
body rotation velocity field
is shown as equal spaced
parallel contour lines
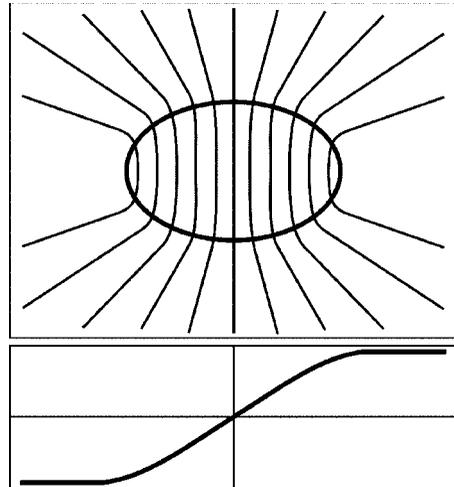
13.8.1 Virial Masses

The simplest and in many ways the most reliable estimate of mass is obtained by applying the virial theorem. The fundamental assumption is that the objects are stable entities whose motions are gravitationally bound. In addition, one assumes that the linewidths reflect *only* the effect of gravitation, and that the lines used are not significantly broadened by high optical depth. The derivation of this result is contained in most classical mechanics textbooks (see, e.g., Binney and Tremaine 1987). We begin by examining the interactions between all constituents, treating each as a classical particle. The momentum of each is $\boldsymbol{p}_i = m_i\boldsymbol{v}_i$. Then

$$\frac{d}{dt}(\boldsymbol{p}_i \cdot s_i) = \dot{\boldsymbol{p}}_i \cdot s_i + \boldsymbol{p}_i \cdot \dot{s}_i . \tag{13.59}$$

Averaging over time, we have

$$\frac{1}{\tau}\int_0^\tau d(\boldsymbol{p}_i \cdot s_i) = \frac{1}{\tau}\int_0^\tau \dot{\boldsymbol{p}}_i \cdot s_i\, dt + \frac{1}{\tau}\int_0^\tau \boldsymbol{p}_i \cdot \dot{s}_i\, dt . \tag{13.60}$$

The term on the left side is simply

$$\frac{1}{\tau}\left[\boldsymbol{p}(\tau)_i \cdot s(\tau)_i - \boldsymbol{p}(0)_i \cdot s(0)_i\right] . \tag{13.61}$$

On the right-hand side,

$$\boldsymbol{p}_i \cdot \dot{s}_i = m_i\boldsymbol{v}_i \cdot \boldsymbol{v}_i = 2T_i . \tag{13.62}$$

here T_i is the kinetic energy. Also, we note that $\dot{\boldsymbol{p}}_i = \boldsymbol{F}_i$. From this, we have:

$$\frac{1}{\tau}\left[\boldsymbol{p}(\tau)_i \cdot \boldsymbol{s}(\tau)_i - \boldsymbol{p}(0)_i \cdot \boldsymbol{s}(0)_i\right] = 2\overline{T} + \sum_i \overline{\boldsymbol{F}_i \cdot \boldsymbol{s}_i}. \tag{13.63}$$

Allowing time to increase without bound, the term on the left will tend to zero for *bounded* motion. This is the fundamental assumption. The first term on the right side is the total energy of motion, kinetic and turbulent, the second term is the gravitational energy of all constituents of the cloud.

The gravitational force between points i and j is:

$$\boldsymbol{F}_{ij} = \frac{Gm_i m_j}{s_{ij}^3}(\boldsymbol{s}_j - \boldsymbol{s}_i). \tag{13.64}$$

The total gravitational force from objects j acting on object i is then:

$$\boldsymbol{F}_i = \sum_{j \neq i} \boldsymbol{F}_{ij} \tag{13.65}$$

then,

$$\sum_i \boldsymbol{F}_i \cdot \boldsymbol{s}_i = \sum_i \sum_{j \neq i} \frac{Gm_i m_j}{s_{ij}^3}(\boldsymbol{s}_j - \boldsymbol{s}_i) \cdot \boldsymbol{s}_i. \tag{13.66}$$

If the double sum in (13.66) were written out, we would find that any pair of objects, say k and l, would contribute twice; the first arising when the first summation index equaled k and the second l, and again when the first summation index equaled l and the second k. This allows us to reduce the double sum to a single sum, by combining the terms of the form $[(\boldsymbol{s}_l - \boldsymbol{s}_k) \cdot \boldsymbol{s}_k + (\boldsymbol{s}_k - \boldsymbol{s}_l) \cdot \boldsymbol{s}_l]$. If this is done, terms in (13.65) of the form

$$\frac{Gm_i m_j}{s_{kl}^3}(\boldsymbol{s}_k - \boldsymbol{s}_l) \cdot (\boldsymbol{s}_l - \boldsymbol{s}_k) = \frac{Gm_i m_j}{s_{kl}} \tag{13.67}$$

can be simplified. Then using this relation, (13.60) reduces to

$$2\overline{T} = -\sum_{\text{all pairs } ij} \frac{Gm_i m_j}{s_{ij}}. \tag{13.68}$$

If we assume that all particles have the same mass, then

$$2\overline{T} = \overline{\sum m_i v_i^2} = nm\overline{\langle v^2 \rangle}. \tag{13.69}$$

For the gravitational part of the expression,

$$\sum_{\text{all pairs } ij} \frac{Gm_i m_j}{s_{ij}} = \frac{Gm^2}{\langle R \rangle} \sum_{\text{all pairs}}. \tag{13.70}$$

The sum over all pairs is $n(n-1)/2$. For large n, $n-1 \approx n$. Furthermore, $nm = M$, the total mass. Then the expression becomes,

$$\overline{v^2} = \frac{GM}{2R} \tag{13.71}$$

The velocity in (13.71) is the three-dimensional root mean square, RMS, velocity. Since we observe the one-dimensional velocity, we must multiply this measured value by a factor 3; to convert to a FWHP velocity, we must divide by $8\ln 2$. That is, $\overline{v^2} = (3/8\ln 2)\Delta v_{1/2}^2$. Then, the virial relation becomes:

$$\boxed{\frac{M}{M_\odot} = 250 \left(\frac{\Delta v_{1/2}}{\mathrm{km\,s^{-1}}}\right)^2 \left(\frac{R}{\mathrm{pc}}\right)} \quad . \tag{13.72}$$

13.8.2 The Tully-Fisher Relation

One can also determine a distance scale by comparing the luminosity and maximum rotational velocity, V, of a galaxy. The *Tully -Fisher* (T-F) relation, $V^4 \sim L$, was shown to exist by Tully and Fisher (1976). Basically, the approach is to measure the maximum velocity from H I data, and then to compare this with luminosity. There is an observational problem in that the best estimate of the velocity range is measured for an edge-on galaxy. However, because of extinction, the luminosity will be underestimated. For face-on galaxies, on the other hand, the luminosity will be unbiased, but the velocity range is decreased because of projection effects.

This problem can be best dealt with if one uses near infrared luminosities, because then absorption effects are reduced. The basis of the near IR correlation for normal, quiescent spiral galaxies is that the old disk population dominates the near IR luminosity as well as the total mass. When corrected for inclination, the mass is proportional to the total extent of velocity to some power (i.e. the virial theorem). The T-F relation is calibrated using a set of galaxies with known mass and luminosities. If one fits a dependence of the near infrared H (i.e. $1.6\,\mu m$) magnitude on velocity width, one obtains the relation

$$H_{-0.5} = -21.23 - 10.0 \left[\log(V_m) - 2.5\right] \tag{13.73}$$

where V_m is the velocity width and $H_{-0.5}$ is the integrated IR luminosity out to -0.5^m from the maximum luminosity of the galaxy in question. More recent studies (see, e.g., the study of the Coma cluster by Bernstein et al. 1994) have found the IR intensity in magnitudes to be 5.6 times the logarithm of the linewidth.

Clearly this method will not be very accurate for face-on galaxies. One notes that magnitude is proportional to 10 times the logarithm of line width. The dependence is a general one if all galaxies have the same mass profile, central mass density, and average mass to light ratio. This can be proven as follows:

Take the mass distribution as

$$\rho(\mathbf{r}) = \rho_0 f(\mathbf{x}) = \rho_0 f(\mathbf{r}/r_e), \tag{13.74}$$

the corresponding surface density on the plane of the sky is

$$\mu(r) = \mu_0 g(r/r_e) = \mu_0 g(x). \tag{13.75}$$

The functions f and g are dimensionless relations and r_e is a characteristic scale length. The rotation curve is similarly written

$$V(r) = V_m h(x). \tag{13.76}$$

Then the kinetic energy is

$$T = \tfrac{1}{2} \int \rho(r) V^2(r) \, d\tau = \tfrac{1}{2} \rho_0 V_m^2 r_e^3 \int f(x) h^2(x) \, d\tau = \tfrac{1}{2} \rho_0 V_m^2 r_e^3 a, \tag{13.77}$$

where a is a numerical constant computed from the potential energy

$$\Omega = \tfrac{1}{2} \int \Phi(r) \rho(r) \, d\tau. \tag{13.78}$$

$\Phi(r)$ can be expressed as a Green's function solution of the Poisson equation:

$$\Phi(r) = G\rho_0 \int \frac{f(r')}{|r - r'|} \, d\tau. \tag{13.79}$$

Then

$$\Omega = \frac{1}{2} G\rho_0^2 \int f(r') \int \frac{f(r')}{|r - r'|} \, d\tau \, d\tau'$$

$$= \frac{1}{2} G\rho_0^2 r_e^5 \iint \frac{f^2(x)}{|x - x'|} \, d\tau \, d\tau' = \frac{1}{2} G\rho_0^2 r_e^5 b. \tag{13.80}$$

Adopting $T = \Omega$, from the virial theorem we have

$$V_m^2 = G\rho_0 r_e^2 \frac{b}{a}. \tag{13.81}$$

But

$$M = \int \mu(x) \, d\sigma = \mu_0 r_e^2 \int g(x) \, d\sigma$$

$$= \rho_0 r_e^3 \int f(x) \, d\tau = \rho_0 r_e^3 d = \frac{1}{G} \frac{V_m^2}{R}, \tag{13.82}$$

where the right-most relation is obtained from the virial theorem. The integrals over density and the parameter r_e must be eliminated, since these are basically not measurable – they are, at least, very uncertain. Doing this we finally obtain

$$V_m{}^4 = \frac{b^2 c}{4 a^2 d^2} \mu_0 G^2 M, \tag{13.83}$$

that is, a relation between V_m and the total mass M of the galaxy. Adopting $M \sim L$, where L is a conveniently chosen total luminosity then

$$2.5 \log L \propto 2.5 \log V_m^4 = 10 \log V_m. \tag{13.84}$$

This is the Tully-Fisher relation.

Problems

1. The ratio of the populations in the upper, N_u and lower, N_l levels of the ground state of HI is given by the Boltzmann relation, where the statistical weights in these levels are 3 and 1, respectively: $N_u/N_l = 3\exp(-0.0682/T_s)$. Assume that the spin temperature, T_s, equals the kinetic temperature, T_K. Calculate the population ratio for a temperature of 100 K. Repeat for a temperature of 3 K (the lowest temperature possible under local thermodynamic equilibrium), for 10^4 K (the warm interstellar medium) and 10^6 K. Compare the differences in populations.

2. In this problem, we determine the value of the FWHP linewidth in terms of T_K, the kinetic temperature, and the mass of the emitter, m. We assume that the thermal motion of atoms in three dimensions is described by the Boltzmann relation for velocities v between $\pm\infty$: $f(v) = (m/2\pi kT_K)^{3/2} \exp\left(-\frac{mv^2}{2kT_K}\right)$.
(a) Show that the requirement $\int f(v)dV = 1$ is fulfilled.
(b) Use the distribution from part **(a)**, with the definition $V_{RMS} = \sqrt{\int V^2 f(v)dV}$. The above relation between the line-of-sight FWHP $\Delta V_{1/2}$ (which is measured) and the three-dimensional V_{rms} is $\Delta V_{1/2} = \sqrt{8\ln2/3}\,V_{RMS}$. Use this to relate the measured linewidth to the emitter mass and kinetic temperature for hydrogen. Show that $1\,\mathrm{km\,s^{-1}}$ is equivalent to motion in a gas of $T_K = 21.2$ K.
(c) Show that the general result is $T_K = 21.2\,(m/m_H)(\Delta V_{1/2})^2$, where m_H is the mass of a hydrogen atom.
(d) Compare this value with the speed of sound in an isothermal gas, $c_0 = \sqrt{P/\rho}$, where P is pressure in dyne cm^{-2} and ρ is density in g cm^{-3}. Check that this is dimensionally correct, then evaluate c_0 in terms of kinetic temperature and density (in cm^{-3}) for a perfect gas consisting of hydrogen atoms.

3. In Fig. 13.8, we show the (idealized) cross section for a neutral–neutral collision to excite the population of a two-level system (levels are separated by an energy E_0). On the basis of this description, explain the behavior of the cross section with particle energy.

4. The Zeeman effect allows an estimate of the magnetic field without assumptions. If the sensitivity of the Zeeman splitting of the HI line to **B** field strength is 2.8 Hz per μG for HI, estimate the splitting on earth ($B \sim 1$ G), the Sun ($B \sim 10^3$ G) and in the interstellar medium ($B \sim 1\,\mu$G). Assume that $\Delta V_{1/2}$ in the interstellar medium is
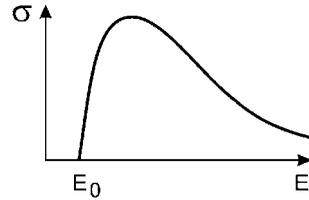
$1 \, \mathrm{km \, s^{-1}}$. As a fraction of this linewidth, what is the shift in the line center caused by the Zeeman effect?

5. (a) Show that the classical expression for the magnetic field at r of a nucleus at position s with a magnetic moment $\boldsymbol{\mu_n}$ is $\boldsymbol{B} = \boldsymbol{\mu_n}(3\frac{r(r \times s)}{r^5} - \frac{s}{r^3})$. From this, the potential energy of an electron with magnetic moment $-\boldsymbol{\mu_e}$ is

$$W = -\boldsymbol{\mu_e} \times \boldsymbol{B} = \frac{\boldsymbol{\mu_n} \times \boldsymbol{\mu_e}}{r^3} - 3\frac{(\boldsymbol{\mu_n} \times r)(\boldsymbol{\mu_e} \times r)}{r^5} . \tag{13.85}$$

(b) Find the size of the Bohr magneton, $\mu_B = he/4\pi mc$, where m is the mass of the electron. Find the size of the nuclear magneton, μ_N, given by the same relation except that the mass of the electron is replaced by the mass of a proton.

(c) Make a qualitative estimate of the energy of interaction as $W = (\boldsymbol{\mu_n} \times \boldsymbol{\mu_e})/r^3$, taking r as one-half the radius of the lowest Bohr orbit, and the μ of the HI nucleus as 1 nuclear magnetron and that of the electron as 1 Bohr magneton. If $W = h\nu$, what is the frequency of such a transition?

(d) In the vector model of the atom, the total angular momentum, $\boldsymbol{F}$, is the vector sum of the electron orbital angular momentum, the electron spin angular momentum and the nuclear spin angular momentum: $\boldsymbol{F} = \boldsymbol{L} + \boldsymbol{S} + \boldsymbol{I} = \boldsymbol{J} + \boldsymbol{I}$. For each of the vectors, one has a quantum-mechanical relation, for example, $\boldsymbol{F} \times \boldsymbol{F} = F(F+1)$ where F is the eigenvalue. Obtain similar expressions for, and then obtain the expression for the energy level spacings in terms of these angular momentum quantum numbers using the relations above. The hyperfine interaction energy follows the conditions that $\boldsymbol{\mu_n} = \mu_n \boldsymbol{I}$ and $\boldsymbol{\mu_e} = \mu_e \boldsymbol{J}$. Use the relation in part 5(c) to obtain similar relations for $\boldsymbol{L}, \boldsymbol{S}, \boldsymbol{I}$ and $\boldsymbol{J}$. Show that the interaction energy of the magnetic moments is $\boldsymbol{\mu_n} \times \boldsymbol{\mu_e} = \boldsymbol{I} \times \boldsymbol{J} \sim [F(F+1) - I(I+1) - J(J+1)]$.

6. The ground states (orbital angular momentum vector $L = 0, J = 1/2$) of HI, DI and $^3\mathrm{He}^+$ have spherically symmetric electron distributions. From this, use a qualitative approach to show that the semi-classical model involving circular orbits and electron spin perpendicular to the plane of the orbit will *not* produce hyperfine energy level splitting, since the orbital plane can take on different angles with respect to the direction of the nuclear spin. To produce the hyperfine level splitting for $L = 0$ states, the electron wavefunction must be evaluated *at the location of the nucleus*. Only with this concept can one arrive at a non-zero value for the frequency of the hyperfine transitions. Thus, the very existence of the HI line relies *completely* on the non-classical concept that the electron has a finite probability of being located at the nucleus. This is *only* possible in a wave mechanics picture where the electron is

treated as a probability density. The wave mechanical treatment is given in the next problem.

7. For the electronic ground state ($L = 0$) of HI, DI and ^{3}He$^+$, the energy of interaction of the electron and nuclear magnetic moments is given by

$$W = \frac{4}{3}\boldsymbol{\mu_e} \times \boldsymbol{\mu_n} \left(\frac{4Z^3}{In^3}\right), \tag{13.86}$$

where Z is the nuclear charge, I is the angular momentum quantum number of the nucleus, and n is the principal quantum number. For HI, the energy levels are designated by the quantum numbers $F_u = 1$, $F_l = 0$ and $I = 1/2$. For DI, $F_u = 3/2$, $F_l = 1/2$ and $I = 1$. In both cases, $J = 1/2$. The magnetic moment of the HI nucleus is 2.79 nuclear magnetons. For the DI nucleus this is 0.857 nuclear magnetons. Use (Eq. 12.43) to scale the DI frequency from the HI frequency.

8. Repeat the previous problem for the hyperfine interaction for ^{3}He$^+$. For this ion, $Z = 2$, and the upper and lower energy levels have quantum numbers $F_u = 0$ and $F_l = 1$. The magnetic moment of the ^{3}He$^+$ nucleus is -2.1274 nuclear magnetons.

9. An estimate of $|\mu_{ul}|$ for hyperfine transitions is given in, e.g., Field (1958): $|\mu_{ul}^2| = \beta^2 \mu_e^2$ where β^2 is $4/3$ for a spin -1 system (D nucleus), and 1 for a spin $-1/2$ system (H nucleus). For HI, $A_{ul} = 2.869 \times 10^{-15}$ s^{-1}. Given this result, the relations above, and the result of Problem 5, Chap. 11, use the line rest frequencies of HI, DI and ^{3}He, 1420.406 MHz, 327.384 MHz and 8665.65 MHz, respectively, and scaling arguments to obtain the A coefficients for DI and ^{3}He$^+$. Compare the results with the compilation in Table 13.1.

10*. This problem outlines an estimate of the amount of telescope integration time needed to detect the 92 cm DI line toward an intense background continuum source. When the 100 m telescope, FWHP beam size $9'$ at 21 cm, and beam efficiency, $\eta_B = 0.75$ is used to measure absorption of the 21 cm line of HI toward the supernova remnant Cassiopeia A, one finds an apparent optical depth, $\tau_{app} = -ln(1 - (T_{line})/(T_{cont}))$, of 2.5. The total continuum flux density of Cas A at 21 cm is $S = 3000$ Jy; this varies with wavelength as $S \sim \lambda^{0.7}$. Take the FWHP size of Cas A as $5.5'$ and assume that the source and beam are Gaussian shaped. We want to search for deuterium using the hyperfine transition.
(a) Use the source and telescope parameters to estimate the FWHP beamwidth of the 100 m telescope, and the peak continuum antenna temperature at the DI wavelength, 92 cm.
(b) If the HI and DI lines arise from the same region, the linewidths in km s^{-1} are equal. The linewidth in frequency units, $\Delta\nu_l$, will follow the Doppler relation ($\Delta V/c = \Delta\nu_l/\nu_l$). The DI profile is assumed to have a FWHP of 2 km s^{-1}; estimate the FWHP of the DI line profile in kHz.
(c) Use (Eq. 12.17) and set $\varphi(\nu)$ equal to dV obtain an expression for the *total* column density of DI. The *completely general* relation between temperature and column density of *any* two-level system, namely

$$N_l = 93.5 \frac{g_l \, \nu^3}{g_u \, A_{ul}} \frac{1}{[1 - \exp(-4.80 \times 10^{-2} \nu / T_{ex})]} \int \tau dV , \qquad (13.87)$$

where N_l is the column density in the lower level.

(d) Derive this relation.

(e) For DI, as for HI, $T_{ex} = T_s = T_{kin}$. Assume that $h\nu \ll kT_{ex}$ to simplify this relation. For the 92 cm line, $A_{ul} = 4.63 \times 10^{-17}$ s^{-1}, g_u=4 and g_l=2, what is the relation for DI?

Assume that the spin temperatures, T_s, of DI and HI are equal that $\tau(\mathrm{HI}) = 2.5$ and that the D/H ratio is 1.5×10^{-5}. What is the antenna temperature of the DI line at 92 cm if $T_{line} = T_{cont} \tau$? From the DI line antenna temperature and the system noise (= receiver noise of 100 K plus the source noise), determine the integration time needed to detect a DI line, for a spectral resolution which is 1/2 of the FWHP linewidth. Compare to the results in Heiles et al. 1993 ApJ Suppl **89**, 271.

11. (a) Suppose a uniform, extended HI cloud has a physical temperature of $T_K = 2.73$ K. If the only background source is the 2.73 K microwave background, would you expect to observe the HI line in emission or absorption or no line radiation at all?
(b) Repeat if there is a background source with main beam brightness temperature, $T_{MB} = 3$ K, that is, $T_{MB} > T_K$. What would be the temperature of the absorption, ΔT_L, in K if $\tau = 1$?
(c) Repeat for $T_K = 3.5$ K.

12. Suppose that there is a layer of neutral gas in thermal equilibrium at a temperature T_K, next to a layer of ionized gas, also at an actual temperature $T_{MB} = T_K = T \gg 2.73$ K. Assume that all layers are much larger than the telescope beam and assume that there is no absorption of the HI line radiation in the ionized gas. Calculate the line intensity in two cases:
(a) The ionized layer is behind the neutral gas layer, and
(b) the ionized layer is in front of the neutral gas layer.

13. In the next three problems we investigate the details of geometry (Fig. 13.9). Assume that all regions are Gaussian shaped. There is a continuum source behind
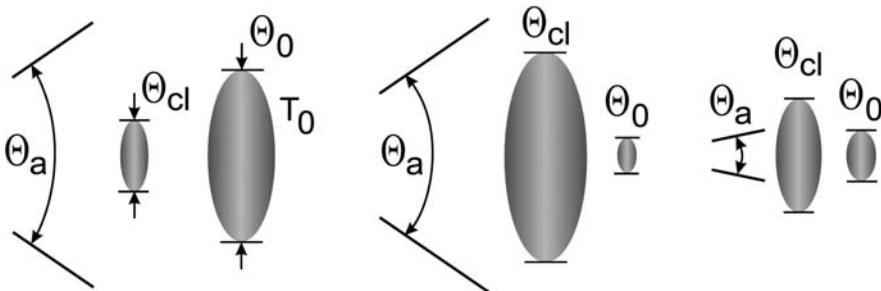
Fig. 13.9 Three sketches for Problems 12, 13 and 14. These deal with the geometry of a neutral gas cloud in front of a continuum source where the relative sizes of antenna beam θ_a, the cloud θ_{cl}, and the continuum source θ_0 differ

the cloud of neutral gas containing HI. The HI in this cloud has an excitation temperature T_{cl} and an angular size θ_{cl}. The continuum source has an *actual* brightness temperature T_0 and angular size θ_0. Assume that θ_a, the beam size of the antenna, is much larger than all other sizes. The cloud covers a fraction f of the background source, that is $\theta_{cl} = f\theta_0$. Specify the conditions under which there will be line absorption against the continuum source. Obtain the expression for the main beam brightness temperature of the line, $\Delta T_L = T_L - T_0$. If $\tau \ll 1$, show that $|\Delta T_L / T_0| = f\tau$

14. Repeat the last problem for the situation in which $\theta_{cl} \gg \theta_0$, but *both* are much smaller than the antenna beam, θ_a. Obtain the expression for the main beam brightness temperature. Under what conditions does one find line absorption?

15. Repeat for the case in which the antenna beam is much smaller than either θ_{cl} or θ_0. Under what conditions does one find absorption?

Chapter 14
Recombination Lines

14.1 Emission Nebulae

The physical state of the interstellar medium varies greatly from one region to the next because the gas temperature depends on the local energy input. There exist large, cool cloud complexes in which both dust grains and many different molecular species are abundant. Often new stars are born in these dense clouds, and since they are sources of thermal energy the stars will heat the gas surrounding them. If the stellar surface temperature is sufficiently high, most of the energy will be emitted as photons with $\lambda < 912\,\text{Å}$. This radiation has sufficient energy to ionize hydrogen. Thus young, luminous stars embedded in gas clouds will be surrounded by emission regions in which the gas temperature and consequently the pressure will be much higher than in cooler clouds. The emission nebulae therefore will expand; this expansion is probably aided by strong stellar winds.

Occasionally an ion will recombine with a free electron. Since the ionization rate is rather low, the time interval between two subsequent ionizations of the same atom will generally be much longer than the time for the electron to cascade to the ground state, and the cascading atom will emit recombination lines. For large H II regions, dynamic time scales are $2L/\Delta v \sim 10\,\text{pc}/10\,\text{km s}^{-1} \sim 10^6$ years, while the ionized gas recombines in $1/\alpha n_i \sim 10^4$ years. H II regions are thus dynamical features with an evolutionary time scale comparable only to that of other extremely young objects.

Planetary nebulae (PN) are another class of ionized nebula. They are objects of much greater age than classical H II regions. Stars in a fairly advanced stage of evolution produce extended atmospheres which are only loosely bound and which have such large dimensions that they appear as faintly luminous greenish disks resembling planets in visual observations with small telescopes – hence their name.

In this chapter the physics of the radio line emissions of these different objects will be discussed, and we will describe how the physical parameters of these nebulae can be derived from these observations.

14.2   Photoionization Structure of Gaseous Nebulae

14.2.1   Pure Hydrogen Nebulae

Stellar photons with $\lambda < 912$ Å can ionize neutral hydrogen. After some time a stationary situation will form in the gas in which the recombinations will just balance the ionizations. The ionization rate by stellar radiation is fairly low in an average H II region, resulting in a typical ionization time scale of 10^8 s; this is much longer than the time it takes an excited atom to cascade down to the ground state by radiative transitions. If ionization by starlight is the dominant energy source of the gas, virtually all hydrogen atoms are either ionized or they are in the $1^2S_{1/2}$ ground level, and we need only to consider this photoionization cross section (Fig. 14.1). The detailed computations are complicated [see, e.g. Rybicki and Lightman (1979), p. 282, or Spitzer (1978), p. 105, for further details]. Figure 14.1 shows the general frequency dependence of the ionization with $\sigma_\nu = 6.30 \times 10^{-18}$ cm^{-2} per atom for $\nu \cong \nu_1 = 3.3 \times 10^{15}$ Hz decreasing as ν^{-3} for $\nu > \nu_1$ and $\sigma_\nu \approx 0$ for $\nu < \nu_1$. The optical depth $\tau_\nu = \sigma_\nu N_{\mathrm{HI}}\, s$ therefore is large for $\nu \gtrsim \nu_1$, with a mean free path (path s_0 for which $\tau = 1$) of

$$\left(\frac{N_{\mathrm{HI}}}{\mathrm{cm}^{-3}}\right)\left(\frac{s_0}{\mathrm{pc}}\right) = \frac{1}{20}. \tag{14.1}$$

Therefore, all neutral hydrogen atoms within s_0 will be ionized and N_{HI} will be absent; at some s_1 there will be a transition from H$^+$ to H. The width of this transition region is governed by the recombination rate of hydrogen. When a proton recombines with a free electron the resulting hydrogen atom could be in any excited state. If α_i is the probability for recombination into the quantum state i, then the total recombination coefficient is

$$\alpha_{\mathrm{t}} = \sum_{i=1}^{\infty} \alpha_i.$$

Since on recombination, the excess energy of the free electron is radiated away as a photon, this photon energy is usually less than the ionization energy of hydrogen in the ground level; only if the atom recombines to the ground state, will $h\nu_{\mathrm{c}} > 13.6$ eV. Due to the large absorption coefficient this radiation will be quickly re-absorbed. It is therefore scattered throughout the nebula until a recombination occurs at $i > 1$. Therefore the effective recombination probability is given by

$$\alpha_{\mathrm{t}} = \sum_{i=2}^{\infty} \alpha_i.$$

Surrounding a young, high-temperature star we will therefore find an ionized region. For the details we would need to consider the detailed balance of the ionization equation; see Spitzer (1978) or Osterbrock (1989).
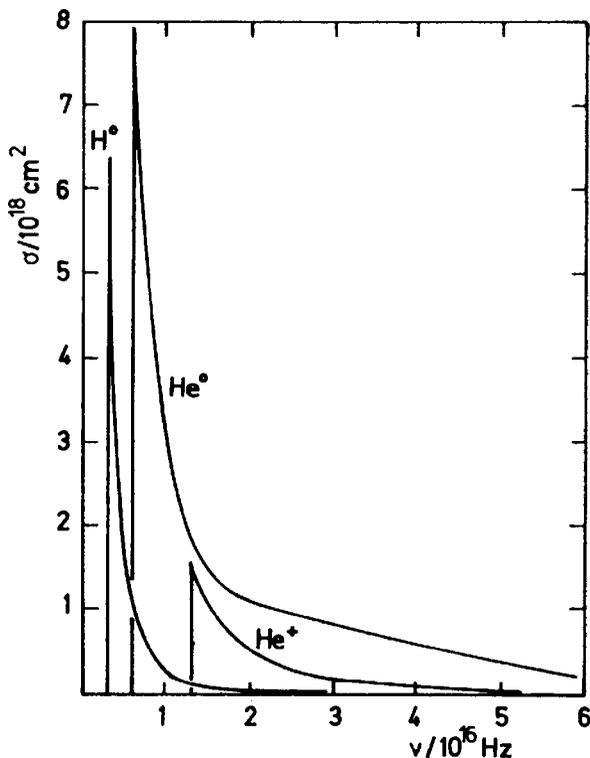
Fig. 14.1   The photoionization cross sections for H^0, He^0 and He^+ (Osterbrock 1989)

Let the ionizing star emit N_{L_c} Lyman continuum quanta per unit surface area, and let its radius be R_*, then the effective radius of the ionized region will be such that the total ionization rate caused by the star should just be equal to the total recombination rate of the surrounding H II region; that is,

$$4\pi R_*^2 N_{L_c} = \tfrac{4}{3}\pi N_e N_p \,\alpha_t \, s_0^3 . \tag{14.2}$$

If the nebula consists mainly of hydrogen, then $N_e \approx N_p$ and

$$s_0 N_e^{2/3} = \left[\frac{3 R_*^2 N_{L_c}}{\alpha_t} \right]^{1/3} = U \tag{14.3}$$

where U is a function of the spectral type of the star. Except for some atomic properties the quantity in square brackets depends only on properties of the exciting stars; it does *not* depend on properties of the nebula. N_{L_c} depends mainly on the surface temperature of the star, and this again is measured by its spectral type, so that the quantity on the right-hand side of (14.3), the excitation parameter U, will be a function of

Table 14.1 The flux of Lyman continuum photons N_{L_c} and the excitation parameter U for stars of spectral type O4–B1 [after Panagia (1973)]

Sp T	T_{eff}/K	log (L_c / photons s^{-1})	$U/(pc\,cm^{2/3})$
O 4	52 000	50.01	148.0
O 5	50 200	49.76	122.0
O 6	48 000	49.37	90.0
O 7	45 200	48.99	68.0
O 8	41 600	48.69	54.0
O 9	37 200	48.35	41.0
O 9.5	34 800	48.10	34.0
B 0	32 200	47.62	24.0
B 0.5	28 600	46.65	11.0
B 1	22 600	45.18	3.5

the spectral type only (see Table 14.1). The more recent stellar atmosphere models of Kurucz (1979) indicate that the number of He ionizing photons are reduced.

The average degree of ionization in the H II region can be estimated in the following way. Defining

$$x = \frac{N_p}{N_{HI} + N_p} \tag{14.4}$$

where N_{HI} is the neutral hydrogen gas density and N_p the number density of the protons, then using (14.1)

$$\frac{x}{1-x} = \frac{N_e}{N_{HI}} \cong 20\left(\frac{s_0}{pc}\right)\left(\frac{N_e}{cm^{-3}}\right)$$

or, with (14.3),

$$\boxed{\frac{x}{1-x} = 20\left(\frac{N_e}{cm^{-3}}\right)^{1/3} U(SpT)} \quad . \tag{14.5}$$

From (14.5) we find that $x/(1-x)$ is of the order of 10^3 in almost the whole of the H II region. The value of x changes abruptly only at the boundary between the H II and the H I region.

We can estimate the thickness of this transition layer by assuming a fractional ionization $x = 1/2$ in it, i.e. by putting $N_p \approx N_{HI}$. From (14.1) the thickness of this transition zone will be

$$\frac{\Delta s}{pc} = \frac{1}{20}\left(\frac{N_{HI}}{cm^{-3}}\right)^{-1},$$

a value that is usually quite small compared to that of s_0 in (14.3). The gas is therefore divided into two sharply separated regions: the almost completely ionized H II region surrounded by the almost completely non ionized H I region. The

dependence of x on s can be determined explicitly using the ionization equation as first done by Strömgren (1939); for the following discussion this accuracy should suffice.

14.2.2 Hydrogen and Helium Nebulae

So far we have not accounted for the fact that about 10% of the atoms in the interstellar gas are helium. These will be ionized also, but, due to their ionization potential of 24.6 eV, the wavelength of the ionizing radiation must be shorter than 504 Å. Provided that the excitation of He radio recombination lines is like that of H recombination lines, one could hope to determine the He abundance relative to hydrogen in interstellar space in a way that is reasonably free from many of the problems met when the He abundance is determined from optical data; due to the large difference in the excitation energies needed for H and He lines in the optical range, it is very difficult to separate small abundance differences from differences in temperature and pressure. As will be shown for Rydberg atoms, in an environment with small optical depth for the lines we obtain from (14.24)

$$\frac{T_L(\mathrm{H\,II})}{T_L(\mathrm{He\,II})} = \frac{\int_{s_1}^{s_2} N_i(\mathrm{H\,II})\,ds}{\int_{s_1'}^{s_2'} N_i(\mathrm{He\,II})\,ds} \propto \frac{\int_{s_1}^{s_2} N_e^2(s)\,ds}{\int_{s_1'}^{s_2'} N_e^2(s)\,ds}. \tag{14.6}$$

Then we obtain

$$\frac{T_L(\mathrm{H\,II})}{T_L(\mathrm{He\,II})} = \left\langle \frac{N(\mathrm{H})}{N(\mathrm{He})} \right\rangle \tag{14.7}$$

only if:

1) The H II and He II regions have the same extent. This will be the case if the number of UV photons emitted by the exciting stars which can ionize He is similar to those that can do this for H. As shown in Sect. 14.2 this is so if $T^* > 37\,000$ K.
2) $N(\mathrm{He}^{++})$ is negligible compared to $N(\mathrm{He}^+)$. Since the ionization potential for He^{++} is $\chi_2 = 54.4$ eV compared to $\chi_1 = 24.6$ eV this can safely be adopted.
3) $N(\mathrm{H\,I}) \cong 0 \cong N(\mathrm{He\,III})$ in H II regions.
4) The H and He recombination lines have the same excitation.

Each photon that can ionize He is able to ionize H, but the reverse may not be true. For a precise description of the situation we would need a system of coupled differential equations. However, we can obtain quite a good estimate by a qualitative calculation.

The number of UV photons that can ionize H I is given by

$$\int_{\nu_1}^{\infty} \frac{L_\nu}{h\,\nu}\,d\nu = Q(\mathrm{H\,I})$$

and in a stationary situation, this number should equal the total number of H recombinations. Therefore

$$Q(\text{H\,I}) = \tfrac{4}{3}\pi s_0^3 N_e N_p \alpha_t(\text{H\,I}),\tag{14.8}$$

where s_0 is the radius of the H II region and $\alpha_t(\text{H\,I})$ the total effective recombination coefficient for hydrogen. For He we obtain a similar expression

$$Q(\text{He}) = \tfrac{4}{3}\pi s_1^3 N_e N_{\text{He}} \alpha_t(\text{He})\tag{14.9}$$

and, due to the different ionization potentials of H and He, $s_1 < s_0$. In the He II region the electrons arise from both H and He, so that $N_e(\text{He\,ii}) = N_H + N_{\text{He}}$, while in the H I region outside s_1, only H contributes, so that $N_e(\text{H\,I}) = N_H$. Dividing (14.8) by (14.9), we obtain

$$\boxed{\left(\frac{s_0}{s_1}\right)^3 = \frac{Q(\text{H\,I})}{Q(\text{He\,II})}\frac{N_{\text{He}}}{N_H}\left(1+\frac{N_{\text{He}}}{N_H}\right)\frac{\alpha_t(\text{He\,II})}{\alpha_t(\text{H\,I})}}\tag{14.10}$$

In this, only the Q factors depend on the exciting star; the remaining factors depend only on the properties of the interstellar medium.

The most realistic stellar atmosphere models are from Kurucz (1979). For an accurate measurement of the He/H ratio, if the true He/H ratio is 0.10, one requires that the ionization is provided by a star of O7 (luminosity $10^5\,\text{L}_\odot$) or earlier (Mezger 1980).

Observationally, there is a problem. The large beams of single radio telescopes include many H II regions with very different excitation parameters. A prime example is the Sgr B2 region; with a 2.6′ beam, He^+/H^+ number ratios ≤ 0.014 were measured, while with high-resolution interferometry, maximum ratios of ≈ 0.08 are reached (Mehringer et al. 1993). These differences show that the angular resolution used has a large effect on the observed He^+/H^+ ratio, and that only the highest angular resolution data can provide reliable results.

14.2.3 Actual H$_{II}$ Regions

Actual H II regions contain H, He and "metals", that is, elements heavier than He. Mostly, one assumes that "metals" consist of the elements C, N, and O. For solar metalicity, the combined abundance of C, N, O relative to H is $\sim 10^{-3}$ by number. Since C, N and O are produced in stars, there can be considerable variation in this abundance. Although these elements make up only a small part of the mass of H II regions, they have a large effect on the electron temperature. This is because of the following considerations. In a time independent situation the value of T_e is obtained by setting the heating rate equal to the cooling rate. The heating rate is given by the number of photoelectrons ejected from atoms:

$$G = \Gamma N_a \Delta E\tag{14.11}$$

where G is in erg cm^{-3} s^{-1} and ΔE is a few eV. The loss Λ is

$$\Lambda = N_e N_t \beta E_{ex} e^{-\frac{E_{ex}}{kT_e}} \tag{14.12}$$

where N_t is the number density of target ions, β is a temperature dependent, slowly varying rate constant. The exponential is the Boltzmann factor. The equilibrium temperature is obtained by setting $G = \Lambda$, and assuming the equality of ionizations and recombinations

$$\Gamma N_a = \alpha N_e N_i, \tag{14.13}$$

so

$$\alpha N_e N_i \Gamma \Delta E = N_e N_t \beta E\, e^{E_{ex}/kT_e} \tag{14.14}$$

or

$$T_e = \frac{E_{ex}}{k \ln \left(\dfrac{N_t E_{ex} \beta}{N_i \alpha \Delta E} \right)}. \tag{14.15}$$

Then the value of T_e depends on the abundance of coolants, in this case C, N, O, relative to the number of ions, but *not* on density. This qualitative treatment is confirmed by detailed calculations carried out by Rubin (1985).

14.3 Rydberg Atoms

When ionized hydrogen recombines at some level with the principal quantum number $n > 1$, the atom will emit recombination line emission on cascading down to the ground state. The radius of the nth Bohr orbit is

$$a_n = \frac{\hbar^2}{Z^2 m e^2} n^2, \tag{14.16}$$

and so for large principle quantum number n, the effective radius of the atom becomes exceedingly large. Systems in such states are generally called Rydberg atoms. Energy levels in these are quite closely spaced, and since pressure effects at large n caused by atomic collision may become important, the different lines eventually will merge. The Inglis-Teller formula gives a semi-empirical relation between the maximum number of resolvable lines n_{max} and the electron density

$$\log \left(\frac{N_e}{cm^{-3}} \right) = 23.26 - 7.5 \log n_{max}. \tag{14.17}$$

For $N_e < 10^6$ cm^{-3} this gives $n_{max} > 200$ so that lines with very large quantum numbers should be observable. The frequency of the atomic lines of hydrogen-like atoms are given by the Rydberg formula

$$v_{ki} = Z^2 R_M \left(\frac{1}{i^2} - \frac{1}{k^2} \right), \quad i < k \tag{14.18}$$

where

$$R_M = \frac{R_\infty}{1 + \dfrac{m}{M}} \tag{14.19}$$

if m is the mass of the electron, M that of the nucleus and Z is the effective charge of the nucleus in units of the proton charge. For $n > 100$ we always have $Z \approx 1$ and the spectra of all atoms are quite hydrogen-like, the only difference being a slightly changed value of the Rydberg constant (Table 14.2).

Lines corresponding to the transitions $n + 1 \rightarrow n$ are most intense and are called α lines. Those for transitions $n + 2 \rightarrow n$ are β lines; $n + 3 \rightarrow n$ transitions are γ lines; etc. In the identification of a line both the element and the principle quantum number of the lower state are given: so H 109 α is the line corresponding to the transition $110 \rightarrow 109$ of H while He 137 β corresponds to $139 \rightarrow 137$ of He (see Fig. 14.2). Transitions with $\Delta n = 1$, that is, α transitions with $n > 60$, produce lines with $\lambda > 1$ cm in the radio wavelength range. Kardashev (1959) first suggested that such lines might be observable; they were first positively detected by Höglund and Mezger (1965). All radio recombination lines are fairly weak, even in bright, nearby H II regions such as the Orion nebula M 42.

All atoms with a single electron in a highly excited state are hydrogen-like. The radiative properties of these Rydberg atoms differ only by their different nuclear masses. The Einstein coefficients A_{ik}, the statistical weights g_i and the departure co-efficients b_i are identical for all Rydberg atoms, if the electrons in the inner atomic shells are not involved. The frequencies of the recombination lines are slightly shifted by the reduced mass of the atom. If this frequency difference is expressed in terms of radial velocities (see Table 14.2), this difference is independent of the quantum number for a given element.

The line width of interstellar radio recombination lines is governed by external effects; neither the intrinsic line width nor the fine structure of the atomic levels has observable consequences. In normal H II regions, evidence for broadening of the lines by inelastic collisions is found for $N \geq 130$, from the broad line wings. For $N < 60$ the observed linewidth is fully explainable by Doppler broadening. One part of this is thermal Doppler broadening since for H,

Table 14.2 The Rydberg constant for the most abundant atoms

Atom	Atomic mass (a.m.u.)	$R_A/(\text{Hz})^{(a)}$	$\Delta V/(\text{km s}^{-1})^{(b)}$
^{1}H	1.007825	$3.288\,051\,29\,(25) \times 10^{15}$	0.000
^{4}He	4.002603	$3.289\,391\,18$	-122.166
^{12}C	12.000000	$3.289\,691\,63$	-149.560
^{14}N	14.003074	$3.289\,713\,14$	-151.521
^{16}O	15.994915	$3.289\,729\,19$	-152.985
	∞	$3.289\,842\,02$	-163.272

(a) Rydberg constant for the atom.
(b) velocity offset from Hydrogen.
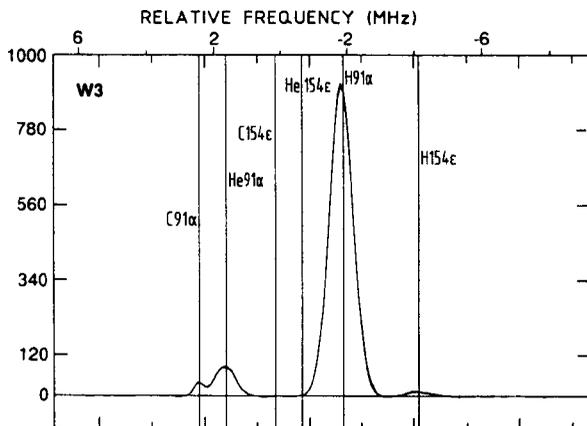
Fig. 14.2 Recombination lines in the H II region W 3 at 8.5 GHz. The most intense lines are H 91 α, He 91 α, C 91 α, the total integration time, t, for this spectrum is 75 hours; even after this time, the RMS noise follows a theoretical dependence of $1/\sqrt{t}$ [from Balser et al. (1994)]

$$\Delta V_{\frac{1}{2}} = 0.21 \sqrt{T_K}.$$

The electrons have a velocity distribution that is described very closely by a Maxwellian velocity distribution; long range Coulomb forces eliminate any deviations with a relaxation time that is exceedingly short. This distribution is characterized by an electron temperature T_e and, due to the electrostatic forces, the protons should have a similar distribution with the same temperature. The spectral lines are observed to have Gaussian shapes. Thermal Doppler motions for $T_e \cong 10^4$ K produce a line width of only 21.4 km s^{-1}, however a width of ~ 25 km s^{-1} to ~ 30 km s^{-1} is observed. Therefore it is likely that nonthermal motions in the gas contribute to the broadening. These motions are usually referred to as *micro turbulence*. If we include this effect, the half width of hydrogen is generalized to

$$\Delta V_{\frac{1}{2}} = \sqrt{0.04576\, T_e + v_t^2}. \tag{14.20}$$

14.4 Line Intensities Under LTE Conditions

To compute the absorption coefficient as given by (12.17) for a given recombination line of hydrogen in local thermodynamic equilibrium, several further parameters for this transition must be specified. For the statistical weight g_n of the level with the principal quantum number n, quantum theory gives

$$g_n = 2n^2. \tag{14.21}$$

This is assumed to hold in high n states of hydrogen. The transition probability A_{ki} for hydrogen has been determined by many different authors. From the correspondence principle, the data for high quantum numbers can be computed by using classical methods, and therefore we will use for A_{ki} the expression (12.24) for the electric dipole. For the dipole moment in the transition $n+1 \rightarrow n$ we put

$$\mu_{n+1,n} = \frac{e\,a_n}{2} = \frac{h^2}{8\pi^2 m e}\,n^2 ,$$

where $a_n = a_0\,n^2$ is the Bohr radius of hydrogen, and correspondingly

$$\nu_{n+1,n} = \frac{m e^4}{4\pi^2 h^3 n^3} .$$

Substituting this expression into (12.24) we obtain, for the limit of large n

$$A_{n+1,n} = \frac{64\pi^6 m e^{10}}{3 h^6 c^3}\frac{1}{n^5} = \frac{5.36 \times 10^9}{n^5}\,\text{s}^{-1} . \tag{14.22}$$

We adopt a Gaussian line shape, $\varphi(\nu)$. Introducing the full line width $\Delta\nu$ at half intensity points, we obtain for the value of φ at the line center

$$\varphi(0) = \left(\frac{\ln 2}{\pi}\right)^{1/2}\frac{2}{\Delta\nu} . \tag{14.23}$$

Another factor in (12.17) is N_n, the density of atoms in the principal quantum state n. This is given by the Saha-Boltzmann equation [cf. Osterbrock (1989), p. 61, Spitzer (1978), Sect. 2.4, or Rybicki and Lightman (1979), (9.45)]

$$N_n = n^2\left(\frac{h^2}{2\pi m k T_e}\right)^{3/2} e^{X_n/k T_e} N_p N_e , \tag{14.24}$$

where

$$X_n = h\nu_0 - \chi_n = \frac{h\nu_0}{n^2} \tag{14.25}$$

is the ionization potential of the level n.

Substituting (14.21, 14.22, 14.23, 14.24) into (12.17) and using $X_n \ll k T_e$ for lines in the radio range so that $\exp(X_n/k T_e) \cong 1$ and $1 - \exp(-h\nu_{n+1,n}/k T_e) \cong h\nu_{n+1,n}/k T_e$, we obtain for the optical depth in the center of a line emitted in a region with the emission measure for an α line

$$\text{EM} = \int N_e(s)\,N_p(s)\,\mathrm{d}s = \int\left(\frac{N_e(s)}{\text{cm}^{-3}}\right)^2 \mathrm{d}\left(\frac{s}{\text{pc}}\right) \tag{14.26}$$

$$\tau_L = 1.92 \times 10^3\left(\frac{T_e}{\text{K}}\right)^{-5/2}\left(\frac{\text{EM}}{\text{cm}^{-6}\,\text{pc}}\right)\left(\frac{\Delta\nu}{\text{kHz}}\right)^{-1} . \tag{14.27}$$

Here we have assumed that $N_p(s) \approx N_e(s)$ which should be reasonable due to the large abundance of H and He (= 0.1 H). We always find that $\tau_L \ll 1$, and therefore that $T_L = T_e \tau_L$, or

$$
T_L = 1.92 \times 10^3 \left(\frac{T_e}{K}\right)^{-3/2} \left(\frac{EM}{cm^{-6}\,pc}\right) \left(\frac{\Delta v}{kHz}\right)^{-1} \qquad . \tag{14.28}
$$

For $v > 1$ GHz we find that $\tau_c < 1$ for the continuum, so that we obtain on dividing (14.27) by (10.35) and using the Doppler relation

$$
\frac{T_L}{T_c}\left(\frac{\Delta v}{km\,s^{-1}}\right) = \frac{6.985 \times 10^3}{a(v,T_e)} \left[\frac{v}{GHz}\right]^{1.1} \left[\frac{T_e}{K}\right]^{-1.15} \frac{1}{1 + N(He^+)/N(H^+)} \quad . \tag{14.29}
$$

The last factor is due to the fact that both N_{H^+} and N_{He^+} contribute to $N_e = N_{H^+} + N_{He^+}$ while N_p in (14.24) is due to N_H only. Typical values for the H II region Orion A are $N(H_e^+)/N(H^+) = 0.08$, $T_L/T_C = 0.245$ and $\Delta V_{1/2} = 25.7\,km\,s^{-1}$ at 22.364 GHz ($\lambda = 1.3$ cm); the value of T_e^* is 8200 K. Equation (14.29) is valid only if both the line radiation and the continuous radiation are optically thin, but the effect of a finite optical depth is easily estimated.

We assume that both the line and the continuum radiation are emitted by the same cloud region with an electron temperature T_e. At the line center, the brightness temperature is

$$
T_{bL} = T_e \left(1 - e^{-(\tau_L + \tau_c)}\right).
$$

At frequencies adjacent to the line, we obtain

$$
T_{bc} = T_e \left(1 - e^{-\tau_c}\right)
$$

so that, for the brightness temperature of the line alone, we find

$$
T_L = T_{bL} - T_{bc} = T_e\, e^{-\tau_c}\left(1 - e^{-\tau_L}\right). \tag{14.30}
$$

Therefore, if $\tau_c \gg 1$, no recombination lines are visible (Fig. 14.3); they are observed only if τ_c is small! This is simply another version of the general principle that optically thick thermal radiation approaches black body radiation and, in black body radiation, there are no lines!

Recombination lines occur over an extremely wide frequency range, from the decimeter range down to the ultraviolet; just to the long wavelength side of the Lyman limit at $\lambda = 912$ Å. Low frequencies, for which the continuous radiation becomes optically thick, must be avoided. But there are also limits for the highest frequency that can be usefully employed. In Sect. 14.3 we showed that $T_L \propto v^{-1}$; so the amplitude of the recombination line decreases with the frequency. Therefore recombination line radiation of extended diffuse objects is best observed at low
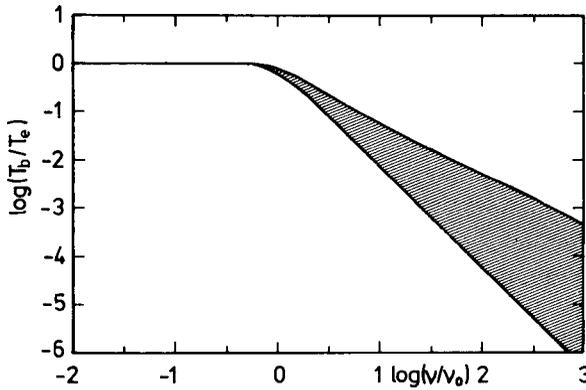
Fig. 14.3 A sketch of the intensities for thermal continuum and recombination lines. Radio recombination lines are visible only in the hatched region. The turnover frequency ν_0 is that frequency where $\tau \approx 1$

frequencies in order to maximize T_L, staying however well above the limiting frequency (10.37) where the source becomes optically thick. Compact H II regions are frequently optically thick even at frequencies of a few GHz and have diameters less than the angular resolution of the telescope. They are, in contrast, best observed at the shortest possible wavelength permitted by both telescope, radiometer, and atmosphere in order to maximize T_L leaving T_L/T_c unchanged.

14.5 Line Intensities when LTE Conditions do not Apply

The diameter of hydrogen atoms is strongly dependent on the principle quantum number [cf. (14.16)].

Atoms in states with high principal quantum numbers are large (cf. Table 14.3), so collisions will have a large effect. A test for the influence of collisions which is independent of instrumental influences therefore should be useful. Let us consider the ratio of the optical depth for two recombination lines which have approximately the same frequency but correspond to different upper (k, k') and lower (i, i') quantum levels. From (12.17) we have

$$\frac{\Delta \nu \, T_L}{\Delta \nu' \, T_L'} = \frac{g_i' \, g_k \, N_i \, A_{ki}}{g_k' \, g_i \, N_i' \, A_{ki}'}, \tag{14.31}$$

but in LTE the density of the different states is governed by the Boltzmann distribution (12.16)

$$\frac{N_i}{N_i'} = \frac{g_i}{g_i'} \exp\left(-\frac{X_i - X_i'}{kT_e} \right).$$

Table 14.3 Diameter and relative density of the hydrogen atom as a function of the principle quantum number

n	Diameter (cm)	ϱ/ϱ_1
1	1.06×10^{-8}	1
10	1.06×10^{-6}	10^{-6}
100	1.06×10^{-4}	10^{-12}
200	4.24×10^{-4}	1.6×10^{-14}
300	9.52×10^{-4}	1.4×10^{-15}

Since

$$X_i - X_i' \ll kT_e,$$

the exponential term will be $= 1$, so we arrive at

$$\frac{\Delta v\, T_L}{\Delta v'\, T_L'} = \frac{g_k A_{ki}}{g_k' A_{ki}'}. \tag{14.32}$$

The right-hand side of this equation contains only atomic quantities and can thus be computed theoretically, while the left-hand side contains observable quantities. There are many possible candidates for this test; the transitions H 110 α, H 138 β, H 158 γ, H 173 δ and H 186 ε all have frequencies close to 4.9 GHz, but numerous other combinations are possible.

The simplification from (14.31) to (14.32) can be made only if LTE conditions hold. If the observations give a result that differs from that theoretically expected, we can be certain that LTE does not apply. Unfortunately even if the observations give a result that agrees with (14.32) we still cannot be certain that LTE applies since the different NLTE effects sometimes seem to "conspire" to fulfill (14.32). This has been discussed by Seaton (1980). We will now investigate how such deviations from LTE will affect the recombination line radiation.

By reversing the arguments leading to (12.16), it is clear that NLTE is equivalent to deviations from the Boltzmann equation. We have already met an elementary version of this situation in low-temperature gas for a two-level system, where the brightness temperature T_b of the ambient radiation field and the kinetic temperature of the gas T_K differed. We then described the relative population of the two lowest states by introducing an excitation temperature T_{ex}, that is some appropriate mean of the brightness temperature T_b and the kinetic temperature T_K. But with a many level system we would find a different excitation temperature for each transition. Therefore the procedure introduced by Menzel (1937) has been applied: *departure coefficients* b_n relate the true population of level, N_n, to the population under LTE conditions, N_n^*, by

$$N_n = b_n N_n^*. \tag{14.33}$$

The b_n factors are always < 1, since the A coefficient for the lower state is larger and the atom is smaller so collisions are less effective. For states i and k, with $k > i$ we have therefore $b_n \to 1$ for LTE, and

$$\frac{N_k}{N_i} = \frac{b_k}{b_i}\frac{g_k}{g_i}\exp\left(-\frac{X_k - X_i}{kT_e}\right) = \frac{b_k}{b_i}\frac{g_k}{g_i}\,e^{-h\nu_{ki}/kT_e}. \tag{14.34}$$

Using (12.17) we obtain

$$\kappa_\nu = \frac{c^2}{8\pi}\frac{1}{\nu_{ki}^2}\frac{g_k}{g_i}N_i A_{ki}\left(1 - \frac{b_k}{b_i}e^{-h\nu_{ki}/kT_e}\right)\varphi(\nu) \tag{14.35}$$

so that we can write (Goldberg 1968)

$$\boxed{\kappa_\nu = \kappa_\nu^* b_i \beta_{ik}} \quad , \tag{14.36}$$

where

$$\boxed{\beta_{ik} = \frac{1 - \dfrac{b_k}{b_i}\exp\left(-\dfrac{h\nu_{ki}}{kT_e}\right)}{1 - \exp\left(-\dfrac{h\nu_{ki}}{kT_e}\right)}} \quad . \tag{14.37}$$

κ_ν^* is the absorption coefficient as given by (12.17) for local thermodynamic equilibrium. Since $h\nu_{ki} \ll kT_e$, (14.37) can be simplified to

$$\beta_{ik} = \left[1 - \frac{b_k}{b_i}\left(1 - \frac{h\nu_{ki}}{kT_e}\right)\right]\frac{kT_e}{h\nu_{ki}}$$

or

$$\boxed{\beta_{ik} = \frac{b_k}{b_i}\left[1 - \frac{kT_e}{h\nu_{ik}}\frac{b_k - b_i}{b_k}\right]} \quad . \tag{14.38}$$

For $0 < b_k - b_i \ll b_k$, $\dfrac{kT_e}{h\nu_{ik}} \gg 1$ and $k - i = \Delta n$, this is equal to

$$\boxed{\beta_{ik} = \frac{b_k}{b_i}\left[1 - \frac{kT_e}{h\nu_{ik}}\frac{d\ln b_n}{dn}\bigg|_i \Delta n\right]} \quad , \tag{14.39}$$

where the differential is evaluated for level i. Substituting numerical values for the physical constants we obtain for $\beta = (b_i/b_k)\beta_{ik}$

$$\beta = 1 - 20.836 \left(\frac{T_e}{K} \right) \left(\frac{\nu}{GHz} \right)^{-1} \frac{d \ln b_n}{dn} \Delta n \qquad (14.40)$$

and

$$\kappa_\nu = \kappa_\nu^* b_n \beta \qquad . \qquad (14.41)$$

Although $0 < b_n < 1$, β can differ considerably from 1 and may become negative. This means that $\kappa_\nu < 0$; that is, we have maser amplification. In order to obtain an indication of how the line intensities are affected the equation of transfer has to be solved. From the definition (12.15) for the emissivity ε_ν, we must have

$$\varepsilon_L = \varepsilon_L^* b_k \qquad (14.42)$$

where ε_L^* is the appropriate value for LTE. According to Kirchhoff's law, we must have

$$\frac{\varepsilon_L^*}{\kappa_L^*} = B_\nu(T). \qquad (14.43)$$

The equation of transfer (1.9) then becomes

$$-\frac{dI_\nu}{d\tau_\nu} = S_\nu - I_\nu \qquad (14.44)$$

with the source function

$$S_\nu = \frac{\varepsilon_\nu}{\kappa_\nu} = \frac{\varepsilon_L^* b_n + \varepsilon_c}{\kappa_L^* b_i \beta_{in} + \kappa_c}. \qquad (14.45)$$

Using (14.43) this can be written as

$$S_\nu = \eta_\nu B_\nu(T) \qquad (14.46)$$

where

$$\eta_\nu = \frac{\kappa_L^* b_n + \kappa_c}{\kappa_L^* b_i \beta_{in} + \kappa_c}. \qquad (14.47)$$

Kirchhoff's law is therefore not valid under NLTE conditions. For an isothermal slab of material with constant density, the brightness temperature at the line center is

$$T_L + T_c = \eta_\nu T_e \left(1 - e^{-\tau_L - \tau_c} \right)$$

or

$$r = \frac{T_L}{T_c} = \eta_\nu \frac{1 - e^{-\tau_L - \tau_c}}{1 - e^{-\tau_c}} - 1. \qquad (14.48)$$

Under conditions of LTE, $b_i = 1$ and $\beta_{ik} = 1$ so that $\eta_v = 1$ and

$$r^* = \frac{T_L^*}{T_c} = \frac{1 - e^{-\tau_L^* - \tau_c}}{1 - e^{-\tau_c}} - 1 .$$ (14.49)

Dividing (14.48) by (14.49) we have

$$\frac{T_L}{T_L^*} = \frac{r}{r^*}$$

so that this ratio describes the influence of NLTE effects on the line intensity. Expanding the exponentials in (14.48) and retaining the quadratic terms in τ_L and τ_c, we obtain

$$r = \eta_v \frac{\tau_L \left(1 - \frac{1}{2} \tau_L - \frac{1}{2} \tau_c \right)}{\tau_c \left(1 - \frac{1}{2} \tau_c \right)} - 1 .$$ (14.50)

Substituting (14.47) for η_v and (14.35) and (14.36) for κ_v with $\tau_L = -\kappa_L s$ and $\tau_c = -\kappa_c s$ we find that

$$\frac{r}{r^*} = b_k - \frac{1}{2} \tau_c b_k \left[1 + \frac{b_i}{b_k} \beta_i (1 + r^* b_k) \right] .$$ (14.51)

In most cases of interest $|\beta_{ik}| \gg 1$ and $r^* b_k \ll 1$; hence

$$\boxed{\frac{r}{r^*} = \frac{T_L}{T_L^*} = b \left(1 - \frac{1}{2} \tau_c \beta \right)} .$$ (14.52)

The first term of (14.52) accounts for NLTE line *formation* effects while the second describes NLTE *transfer* effects, that is, Maser amplification of the line radiation.

One may wonder why the *continuum optical depth* multiplies β. This is because the atoms do not distinguish between line and continuum photons, but amplify a background. Actually (14.52) is an approximation to a uniform region. In actual situation the equation of transfer must be evaluated at each position through the nebula. In our direction the largest amplification will occur in the layer nearest us, since the line and continuum intensities increase from back to front.

In order to apply these concepts, the departure coefficients b_n for a given H II region must be determined. But this requires that all important processes affecting the level population are taken into account. These are

1) level population and transfer effects;
2) collisional excitation and de-excitation by electrons and protons;
3) collisional ionization and three body recombination;
4) redistribution of angular momentum by collisions.

We will consider these processes one by one. First, to obtain the b and β terms in (14.52), the level populations must be calculated using the rate equation:

$$\sum_{r\neq s}(N_rC_{rs}+N_rB_{rs}U_{rs})+\sum_{r>s}N_rA_{rs}+N_eN_i\,(\alpha_{is}+C_{is})$$

$$= N_s\sum_{l<s}A_{sl}+N_s\sum_{l\neq s}(C_{sl}+B_{sl}U_{sl})+N_sC_{si}\,, \tag{14.53}$$

where N_e is the electron density, N_i the ion density, A_{rs} the spontaneous Einstein coefficient for the transition $r\rightarrow s$ and B_{rs} and B_{sr} the corresponding coefficients for stimulated emission and absorption. $U_{rs}=4\pi I_v/c$ is the radiation density at the frequency v_{rs} corresponding to the transition $r\rightarrow s$, and α_{is} is the radiative recombination coefficient for transitions to the level s. Finally, C_{si} represents the collisional ionization rate for transitions for the level s, and C_{is} the corresponding three body recombination rate.

Using a procedure described by Dupree (1969) this can be rewritten to arrive at the equation

$$\sum_{r=s-s_0}^{r=s+s_0} R_{rs}b_s = S_s \tag{14.54}$$

where

$$R_{rs} = \begin{cases} -\dfrac{g_r}{g_s}\,\mathrm{e}^{X_r-X_s}(C_{rs}+B_{rs}U_{rs}+A_{rs}); & r<s \\[2ex] \displaystyle\sum_{l<s}A_{sl}+\sum_{\substack{l\neq s \\ s-s_0}}^{s+s_0}(C_{sl}+B_{sl}U_{sl})+C_{si}; & r=s \\[2ex] -\dfrac{g_r}{g_s}\,\mathrm{e}^{X_r-X_s}(C_{rs}+B_{rs}U_{rs}); & r>s \end{cases} \tag{14.55}$$

and

$$S_n = \sum_{r=n+n_0+1}^{\infty} b_r\frac{g_r}{g_n}\,\mathrm{e}^{X_r-X_n}A_{rn}+C_{ni}+\frac{(2\pi m k T_e)^{3/2}}{h^3}\,\frac{2g_i}{g_n}\,\mathrm{e}^{-X_n}\,\alpha_{in}\,. \tag{14.56}$$

Here g_n and g_i are the statistical weights of an electron at the bound state n and in the continuum,

$$X_n = 1.58\times 10^5\left(\frac{T_e}{\mathrm{K}}\right)^{-1}\frac{1}{n^5}$$

and n_0 is the maximum value of $\Delta r = |r-n|$ for which collisions and stimulated radiative transitions are considered.

This is mainly a numerical problem once the transition rates A_{rs},B_{rs},I_v and the collision rates are specified. Approximate numerical solutions were first given by Menzel and Pekeris (1935); in the meantime much larger systems, with the rate equations above, have been solved, in particular by Brocklehurst (1970).

Usually two major cases are considered.

Case A: The nebula is optically thin in all lines: each photon leaves the nebula.
Case B: The Lyman photons never escape; after many scatterings they are broken down into two photons or are absorbed by dust.

In most nebulae, case B is likely to apply and the collision rates C_{rs} in (14.54, 14.55, 14.56) have to be computed by quantum mechanical methods, but according to (12.36) they will depend on the local electron density. In low-density regions the b_n factors will therefore be determined mainly by the radiation field. Observationally, the relative importance of collisions and the radiation field are strongly dependent on the principal quantum number of the level considered. In a diagram of b_n values as functions of the quantum number n (Fig. 14.4) we can therefore distinguish two main regions: a *radiative domain* for low n independent of the ambient density and a *collisional domain* where the b_n values depend strongly on the electron density N_e. As $n \to \infty$, $b_n \to 1$ and $\beta \to 0$.

14.5.1   Collisional Broadening

The intensity of the recombination lines as found from the arguments presented in the preceding sections seems to be remarkably insensitive to gas density. Reasons for this become clear if line broadening is explicitly considered. The broadening is not so much caused by the quasi static electric fields of the colliding ions (Stark effect) as by random phase perturbation of the emitted line (impact effects). These have been investigated by Griem (1967) and Brocklehurst and Leeman (1971). The line shape is a Lorentz profile but, since this will be convolved by the Doppler profile caused by the random velocities of the atoms, the observed line profile eventually will be a Voigt function [see Rybicki and Lightman (1979), (10.76)].

Brocklehurst and Seaton (1972) find

$$\frac{\Delta v_{\rm I}}{\Delta v_{\rm D}} = 0.142 \left(\frac{n}{100}\right)^{7.4} \left(\frac{N_e}{10^4\,{\rm cm}^{-3}}\right) \left(\frac{T_e}{10^4\,{\rm K}}\right)^{-0.1} \left(\frac{T_{\rm D}}{2 \times 10^4\,{\rm K}}\right)^{-1/2} \quad (14.57)$$

for the ratio of the dispersion profile for the H recombination lines and the half power line width of the Doppler broadening in a medium, with $T_e = 1 \times 10^4$ K and an equivalent Doppler temperature $T_{\rm D} = 2 \times 10^4$ K.

The impact line width thus depends very strongly on the principal quantum number of the line. For $N_e = 10^4\,{\rm cm}^{-3}$ we find $\Delta v_{\rm I}/\Delta v_{\rm D} = 0.14$ if $n = 100$, while for $n = 150$, $\Delta v_{\rm I}/\Delta_{\rm D} = 20$. Lines with such large impact widths are not easily detectable with presently used techniques, since instrumental baselines blend with the broadened profiles. In removing instrumental baseline ripple, one may also remove wide real line wings. Therefore there is a rather sharply defined maximum principal quantum number $n_{\max}$ for which the recombination lines of gas of given density can be detected. In this sense the Inglis-Teller criterion (14.17) can be derived from (14.57).

If we observe the recombination line radiation of a strongly clumped cloud, the line radiation of the high density parts will not be collisionally broadened if $n > n_{\max}$ for this density. Lines of a given n can therefore only be detected for a gas with a density below a critical density N_n. High n lines are indicative of a gas of low density while high density gas can only be detected by low n lines.
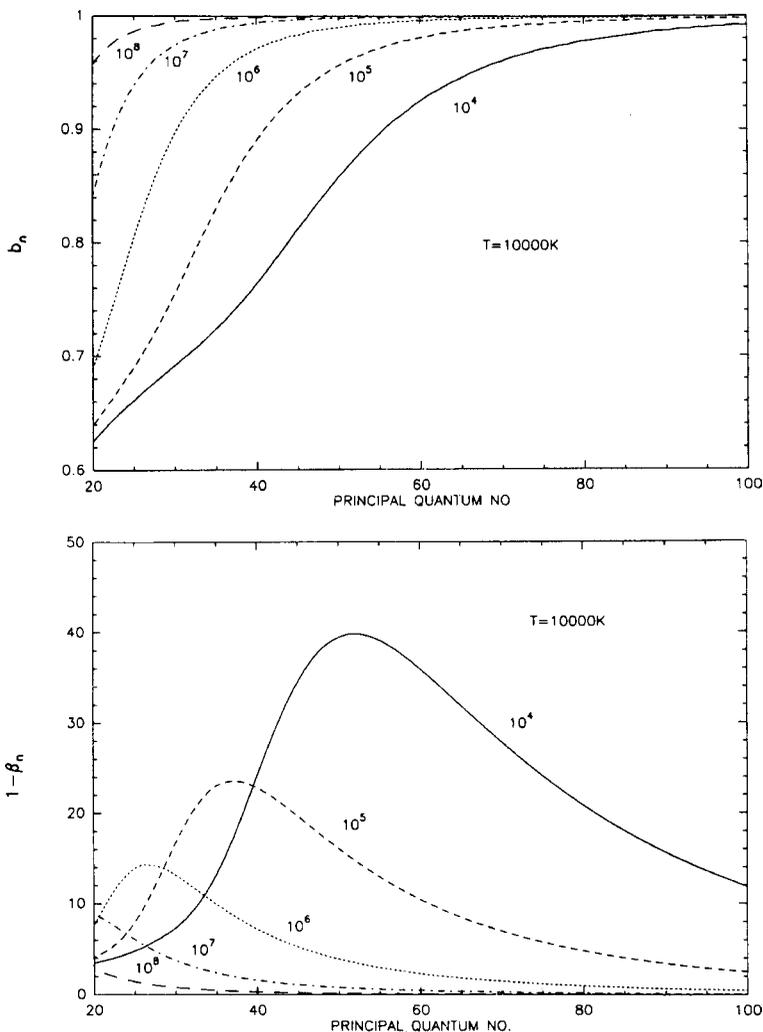
Fig. 14.4 Population departure coefficients b_n and their differential variation $\mathrm{d}\ln b_n/\mathrm{d}n$ plotted against the principal quantum number for $T_e = 10^4$ K and different N_e (Walmsley 1990)

In regard to condition 3, the relatively low densities insure that three body processes are rare. With respect to condition 4, if $N(n)$ is the total population in level n, Pengelly and Seaton (1964) showed that

$$N(n,l) = \frac{2l+1}{n^2} N(n)$$

holds for $n > 47$ at an electron density of $1\,\mathrm{cm}^{-3}$. At higher densities this holds at even lower n levels.

14.6   The Interpretation of Radio Recombination Line
  Observations

Recombination line measurements provide a reasonably simple way to determine
the radial velocities and electron temperatures, T_e, of H II regions, even those which
are optically obscured. As shown, T_e depends in a rather complicated manner on
T_L/T_e, $\langle N_e \rangle$, EM and ΔV; that is, on the details of the structure of the H II region.
However, insight into the physics is gained by the following approximate method.
If we have an H II region in which

1) the structure is plane-parallel, homogeneous and isothermal;
2) all optical depths are small: $|\tau_L + \tau_c| \ll 1$ and $\tau_c \ll 1$;
3) the lines can be treated as if they were formed and transferred in LTE: $b_n = 1$,
   $\beta = 1$, then

(14.29) is valid and can be solved for T_e which we will call T_e^* here:

$$\frac{T_e^*}{K} = \left[\frac{6.985 \times 10^3}{a(v,T)} \left(\frac{v}{\text{GHz}} \right)^{1.1} \frac{1}{1 + \frac{N(\text{He})}{N(\text{H})}} \left(\frac{\Delta v_{\frac{1}{2}}}{\text{km s}^{-1}} \right)^{-1} \left(\frac{T_c}{T_L} \right) \right]^{0.87} \qquad (14.58)$$

For most galactic H II regions this is a good approximation of the actual electron
temperature T_e for $v \approx 10$ to 40 GHz. The exact precision is not clear, but is prob-
ably better than 20%; for higher precision, detailed numerical models are needed.
Even if assumptions 2 and 3 are valid, a different geometry can cause significant de-
viations of T_e from T_e^*. Non-LTE effects can be taken into account by using (14.52)
resulting in

$$T_e = T_e^* \left[b \left(1 - \tfrac{1}{2} \beta \, \tau_c \right) \right]^{0.87} . \qquad (14.59)$$

It has been remarked by several observers that the radio recombination lines of
H seem to be emitted close to LTE, or are in LTE. This is certainly *not* a correct
conclusion, since a number of factors conspire to produce a selection which *appears*
to be close to LTE. Brown et al. (1978) had presented models which predicted a
large maser effect in HII regions, but these were based on an unrealistic choice
of source parameters. Basically, at very low densities, large quiescent H II regions
would produce radio recombination lines which are dominated by line masering.
However, such regions are very rare (see Sect.14.6.1).

A consideration of numerical values for the right side of (14.59) is useful. For the
2.5′ diameter dense core of Orion A, one has an RMS electron density of $\langle N_e \rangle = 5 \times 10^3 \, \text{cm}^{-3}$, and T_e^* of 8200 K. At 23 GHz, the continuum optical depth is $\tau_C = 3.7 \times 10^{-3}$. Then from the tabulations of Brocklehurst (1970), we have for the H 66α line,
$b_{66} = 0.99$ and $\beta_{66} = -42$. Then $T_L = 1.07\,T_L^*$, and the LTE value of T_e represents
the actual value of T_e very well. Because masering effects are more important at
lower frequencies, the best choice of frequency for a survey of many galactic H II
regions will be ~ 10 GHz.

Large scale surveys of H II regions in the disk region of our galaxy show a decrease of T_e^* with distance from the galactic center, D_{GC} (see Fig. 14.6). Studies of the heating and cooling of H II regions indicate that T_e is less dependent on density than on the abundance of heavier elements such as C, N and O. Presumably this trend is caused by an increase in the abundance of C, N and O with D_{GC}. However, in the galactic center itself, the T_e values reach those found near the Sun. These higher T_e values indicate that additional factors may influence T_e.

A comparison of radio and optical determinations of T_e shows that the optical value from the O^{++} forbidden line ratio is

$$\exp\left(-\frac{E_R}{kT_e}\right) = \frac{\int N_e N(O^{++}) \exp\left(-\frac{E_{31}}{kT_e}\right) dV}{\int N_e N(O^{++}) \exp\left(-\frac{E_{21}}{kT_e}\right) dV},\tag{14.60}$$

where $E_{21}/k = 28700$ K and $E_{31}/k = 61700$ K. If T_e varies over the nebula, this ratio is weighted towards the higher values of T_e, while (14.59) is weighted towards the lowest value. In a uniform density region, T_e should be a constant, so variations of T_e indicate density enhancements. Of course, detailed comparisons can be made only for nearby, optically un obscured regions. Up to the last few years it appeared that the T_e values for the dense core of Orion A from the O^{++} forbidden lines and from radio recombination lines agreed fairly well. Newer values, from optical recombination lines, indicate some discrepancies. These T_e values are of great importance since the conversion of optical line intensities to column densities depends critically on T_e. Another determination of N and O abundances can be made using fine structure lines of N and O in the far IR. These have the advantage of being largely temperature independent, but they *are* also somewhat density dependent.

Martin-Hernandez et al. (2002) have presented the distribution of the N^{++}/O^{++} ratio from fine structure lines. These lines are emitted from ultra-compact HII regions. The data were taken using the long and short wavelength spectrometers of the ISO mission. In the disk of our galaxy, this ratio shows a factor of 8 increase in the ratio from 15 kpc to 4 kpc from the galactic center. This increase reflects the production and ejection of processed material, such as nitrogen in lower mass stars in the Asymptotic Giant Branch (AGB). A larger amount of heavier elements is consistent with the decrease in T_e, this is plotted in Fig. 14.6.

14.6.1 Anomalous Cases

The curves for the β factors in Fig. 14.4 show that the high quantum number states for the H atoms are always inverted. This is because decays become faster as n becomes smaller, and the collision rates, which bring the H atom closer to LTE, also become smaller. Then the trend is *always $b_n < b_{n+1}$*. Since $\frac{h\nu}{k} \ll T_e$, and $g_n \approx g_{n+1}$ for large n, this requires a negative value of T_{ex}, which means an inverted population. However, an inverted population alone will not give rise to line masering

(see (1.38)), since from (12.17), $\tau \sim N_m / T_{\text{ex}}$ and then $T_{\text{ex}} \tau \propto N_m$. For this reason, the optical depth of the continuum is important. Observations show that strong maser emission from radio recombination lines is very rare [see e.g. Martin-Pintado et al. (1994)]. The most prominent example of a recombination line maser is the source MWC349. This masering extends even into the infrared (Martin-Pintado et al. 1994, Thum et al. 1998).

14.7   Recombination Lines from Other Elements

Recombination lines of other elements besides H and He have been observed. Unambiguous identifications are more difficult because the radial velocity differences due to the atomic weights of the nuclei become less with increasing reduced masses M_A and converge towards $\Delta V = -163.3 \, \text{km s}^{-1}$ for $M_A \rightarrow \infty$ (see Tab. 14.5). The line seen blended with the He line in the spectrum (Fig. 14.2) has been identified with carbon. If this line originates in the same volume as the H and He lines, the line intensity is a factor of about 60 stronger than expected from the average abundance of carbon.

The following explanation is generally accepted: the line does not originate in the H II region but in the surrounding gas which is only partly ionized and has a much lower electron temperature ($T_e \cong 200 \, \text{K}$) than the H II region. This gives rise to a larger value of T_L, for a given emission measure EM. In addition, the behavior of β and b factors also increases T_L at lower observing frequencies. This also explains why the radial velocity of the C II lines may differ from that of the H II lines by a few km s^{-1}, and usually agrees much better with the velocity of adjacent molecular clouds. To explain why the line should be identified with carbon and not oxygen which is more abundant, one notes that for recombination line radiation the atoms must be ionized. Oxygen has an ionization potential of 13.6 eV while that for carbon is 11.3 eV. Therefore recombination line radiation of oxygen could only be emitted inside H II regions, while carbon can be ionized outside H II regions. All other elements are less abundant than carbon by large factors and their lines would hardly be detectable.

The b and β factors for carbon may be enhanced by the *dielectronic recombination* process. This allows the excess energy and momentum in the recombination process to be shared with another electron, in this case, the electron which gives rise to the fine structure line at 157 μm. A complete set of measurements of carbon recombination lines toward the SNR Cas A has been carried out. These lines arise from the envelopes of clouds in front of the intense continuum source. At low frequencies the populations of ionized carbon are thermalized, so the lines absorb the continuum radiation. At higher frequencies the NLTE effects dominate, and the lines appear in emission, amplifying the continuum source. At even higher frequencies, above a few GHz, the b factors are very small and the background is weaker. In all cases, the optical depths of these lines are $\sim 10^{-3}$. The lack of corresponding lines of hydrogen indicates that the ionization is via photons with $\lambda > 912 \, \text{Å}$; if ionized

by cosmic rays, one would expect to find H lines. Detailed arguments can be used
to set limits on the cosmic ray rates near these regions.

Problems

1. A spherically symmetric, uniform HII region is ionized by an O7 star (mass about
50 $M_\odot$), with an excitation parameter $U = 68 \, \mathrm{pc \, cm^{2/3}}$.
(a) Interpret the meaning of the excitation parameter.
(b) If $N_e = 10^4 \, \mathrm{cm^{-3}}$ what is the radius of this region?
(c) Calculate the *Emission Measure*, $\mathrm{EM} = N_e^2 L$, where L is the diameter of the HII
region.
(d) If this region consists of pure hydrogen, determine the mass.

2. If the ionization is caused by a cluster of B0 stars (each with mass 18 $M_\odot$), with
$U = 24 \, \mathrm{pc \, cm^{2/3}}$, how many of these stars are needed to provide the same excitation
as with one O7 star?

3. (a) Compare the mass of the HII region in Problem 1 to that of the exciting stars
needed to ionize the regions in Problems 1 and 2.
(b) Suppose that the HII region in Problem 1 has an electron density, N_e, of $3 \times$
$10^4 \, \mathrm{cm^{-3}}$, but the same Emission Measure, $\mathrm{EM} = N_e^2 L$, where L is the diameter of
the HII region. Determine the mass of ionized gas in this case.
(c) Now repeat this calculation for the same excitation parameter, but with $N_e =$
$3 \times 10^3 \, \mathrm{cm^{-3}}$.

4. In the core of the HII region Orion A, the diameter is 0.54 pc, the emission
measure, $N_e^2 L = 4 \times 10^6 \, \mathrm{cm^{-6}}$ pc, and the electron density N_e is $10^4 \, \mathrm{cm^{-3}}$. Combine
N_e with the emission measure to obtain the line-of-sight depth. Compare this result
with the RMS electron density obtained by assuming a spherical region with a line-
of-sight depth equal to the diameter. The "clumping factor" is defined as the ratio of
the actual to the RMS electron densities. What is this factor?

5. The assumption in Problem 2 is that all of the exciting stars are of the same
spectral type (and mass). This is not found to be the case. Rather the distribu-
tion of stellar masses follows some distribution. One is the *Salpeter* distribution,
$N(M) = N_0 M^{-1.35}$. Integrate over mass $\int M N(M) dM$ to obtain the total mass of
stars between the limits M_{lower} and M_{upper}. Take M_{lower} as 0.08 $M_\odot$, and M_{upper} as
50 $M_\odot$. Is there more mass in stars of type B0 and larger or in stars with masses
below class B0?

6. (a) In Fig. 14.1 is a sketch of the photoionization cross sections for hydrogen
and two ionization states of helium. Explain why there is a sharp decrease in the
absorption cross section for frequencies lower than ν_0. Calculate the photon energy
corresponding to ν_0.
(b) At frequencies higher than ν_0, the photons are only slowly absorbed. Suppose
that only these (higher energy) photons escape and are absorbed in the outer parts

of an HII region. On this basis, do you expect the electron temperature to be higher or lower than in the center of the H II region? Give an explicit argument.

7. Calculate the Rydberg constant for the nuclei of deuterium (^{2}H) and 3-helium (^{3}He), using (Eq. 14.19). For the electron, the mass is 9.109×10^{-28} g. For D, the nuclear mass is 3.344×10^{-24} g, and for ^{3}He, this is 5.008×10^{-24} g.

8. (a) If the Rydberg constant for ^{4}He is $3.28939118 \times 10^{15}$ Hz, find the separation between the ^{4}He and ^{3}He lines (in $\mathrm{km\,s^{-1}}$). Given that the linewidths of ^{4}He and ^{3}He are $\sim$24 $\mathrm{km\,s^{-1}}$, and that the number ratio ^{3}He/^{4}He $= 10^{-4}$, sketch the shape of each profile and that of the combined profile.
(b) For a ^{4}He line T_A of 2 K, frequency resolution 100 kHz, and system noise temperature of 40 K, how long must one integrate using position switching to detect a ^{3}He recombination line?

9. The exact formula for a transition from the ith to the nth level, where $i < k$ is given in (Eq. 14.18)
(a) If we set $k = n + 1$, show that the approximate Rydberg formula for the transition from the $n + 1^{\mathrm{th}}$ to the nth levels, that is for the $n\alpha$ line, is

$$ \nu = \frac{2Z^2 R_M}{n^3} . $$

(b) Determine the error for this approximation in the case of $n\alpha$ transitions, for $n = 126, 109, 100$ and 166. If a total analyzing bandwidth of 10 MHz is used to search for these recombination lines, show that the line frequencies calculated using the approximate formula do *not* fall in the spectrometer band.

10. Suppose that the recombination lines from the elements ^{4}He and ^{12}C are emitted without turbulence. The ^{4}He arises from a region of electron temperature, T_e, 10^4 K, while the ^{12}C arises from a region with electron temperature 100 K. (Modify (Eq. 14.20), which is valid for H, for these elements using the atomic weights. Assume that the turbulent velocity, v_t, is zero.) The ^{4}He $- ^{12}$C separation is 27.39 $\mathrm{km\,s^{-1}}$. If the intensities are equal, at what level do these lines overlap? Suppose the turbulence of the ^{4}He line is 20 $\mathrm{km\,s^{-1}}$. Now what is the overlap?

11. (a) At larger principal quantum number values, n, the sizes of atoms are larger, so collisional broadening is more important. In Fig. 14.5, we show the behavior of various broadening mechanisms as a function of frequency. Determine the dependence of fractional linewidth on frequency for the most realistic relation, by Brocklehurst and Seaton, given in Eq. (14.57). For $N_e = 10^3$ cm^{-3} and $T_e = 10^4$ K, $T_D = 2 \times 10^4$ K, this is $\Delta v_{\mathrm{coll}}/\Delta v_D = 2.25 \times 10^{-17} n^{7.4}$. Find the limit of detectability for this line-broadening mechanism if the separation between α transitions is equal to twice the line broadening.
(b) The first reliable detection of a radio recombination line was made with a spectrometer which covered a total bandwidth of 2 MHz, at a center frequency of 5 GHz. The Doppler width was 26 $\mathrm{km\,s^{-1}}$. If the line broadening followed the curve marked "Kardashev", could a line be detected?
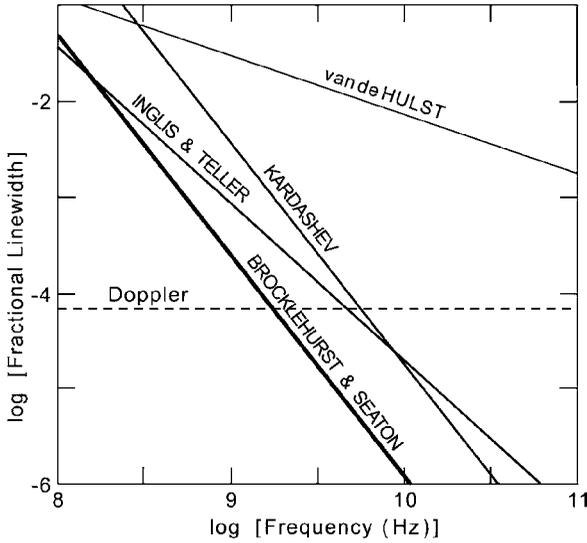
Fig. 14.5 A graph of fractional linewidths predicted by various broadening models, versus frequency. The electron density, N_e, is $10^3 \, cm^{-3}$, the electron temperature, T_e, is 10^4 K (see Problem 11)
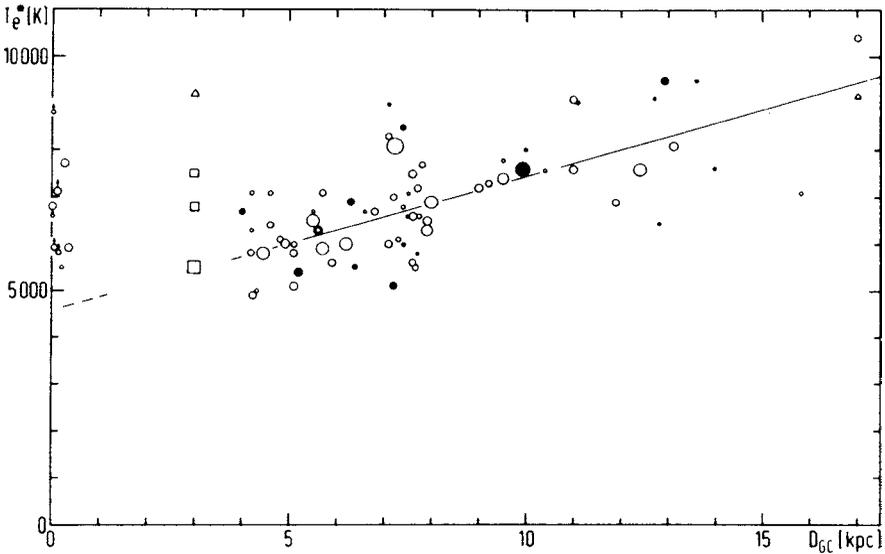
Fig. 14.6 A plot of the LTE electron temperatures, T_e^*, obtained from measurements of the H 76α line and source continuum, plotted versus distance from the galactic center. The different symbols refer to H II regimes with different excitation parameters, U (Table 13.1). The data show a gradient in T_e^* which is believed to represent a gradient in the relative abundance of "metals", mainly C, N and O (Wink et al. 1983)

12. A simple model for line broadening is to use the Bohr radius of the atom, a_n, as the characteristic size, and to then set this geometric cross section, σ, times the product of velocity, v, and free electron density, N_e, to determine the collision rate. By setting this equal to the spontaneous decay rate, one can obtain the critical density, N^*. Evaluate the expression for the Bohr radius (Eq. 14.16), $a_n = (h^2/Z^2 m e^2) n^2$, to obtain a numerical result, $0.529 \times 10^{-8} n^2/Z^2$ cm, where n is the principal quantum number and Z is the nuclear charge. Now use this to estimate $\sigma = \pi a_n^2$. To obtain the critical density, set $A_{n+1,n} = 5.36 \times 10^9 n^{-5}$ s^{-1}, the spontaneous decay rate (this is obtained in the next problem) equal to $\langle N^* \sigma V \rangle$. Compare this result to Fig. 14.5 and (Eq. 14.57). What must be the coefficient of the relation derived in this problem to obtain agreement? This adjustment is identified with the fact that for large principal quantum numbers the disturbance in the upper and lower levels is very similar so that the effect on the linewidth is considerably reduced.

13. Repeat the derivation given in Eq. (14.22, 14.23, 14.24, 14.25, 14.26, 14.27, 14.28, 14.29) for the recombination of the remaining electron of singly ionized helium. This electron will experience the field of the *doubly ionized* nucleus.
(a) Estimate the line frequencies
(b) For measurements of recombination lines of He$^+$ and He at nearly the same frequency, which line is more intense? For this, one must calculate the dipole moment and A coefficient. The dipole moment as obtained from the correspondence principle is $\mu_{n+1,n} = (1/2)ea_n$, where a_n is the Bohr radius for principal quantum number n (cf. (Eq. 12.24)). Refer to the previous problem to obtain a_n for an atom with nuclear charge Z. Use the expression for the dipole moment given above, then use this to show that the A coefficient for high $n\alpha$ lines is $A_{n+1,n} = 5.36 \times 10^9 n^{-5} Z^2$ s^{-1}.

14. Use (Eq. 14.30), with $\tau_L \ll 1$ but $\tau_c \gg 1$ to investigate how T_L is affected by a finite continuum optical depth. Suppose you are unaware of the effect of the continuum optical depth; show that the value of T_L is reduced. Use the fact that T_L is proportional to $1/T_e$ to show that the value of T_e obtained from the measurement of T_L and T_c will be larger than the value one would obtain if τ_c is small.

15. The level populations of hydrogen atoms in an HII region deviate from LTE. We use a specific set of parameters to estimate the size of these quantities. From the Boltzmann relation for $T = 10^4$ K, we find that

$$N(\text{LTE}, 101)/N(\text{LTE}, 100) = 1.00975 .$$

Use the ratio of the b_n factors, $b_{101}/b_{100} = 1.0011$ for $N_e = 10^3$ cm^{-3} and $T_e = 10^4$ K (from Brocklehurst 1970), to determine the excitation temperature between the $n = 100$ and $n = 101$ energy levels. Is T_{ex} greater or less than zero? Determine the ratio of b_n factors for $n = 100$ and 101 which allows *superthermal* populations ($T_{ex} > T_k > 0$) by setting $T_e = \infty$.

16. For the $n = 40$ and 41 levels, for $N_e = 10^3$ cm^{-3} and $T_e = 10^4$ K and $h\nu/k = 4.94$ K, the ratio of $b_{41}/b_{40} = 1.005$. Determine the excitation temperature between these levels.

17. For $N_e = 10^3 \, \text{cm}^{-3}$ and $T_e = 10^4 \, \text{K}$, for the principal quantum number, $n = 100$, the departure coefficient is $b = 0.9692$ and $(\text{d}\ln b_{100}/\text{d}n)\Delta n = 1.368 \times 10^{-3}$. Determine β using (Eq. 14.40). Then calculate T_L/T_L^* using (Eq. 14.52).

18. From the previous problems, as well as an examination of Fig. 14.4, show that the level populations approach LTE with increasing N_e, and also increasing principal quantum number n. Given the HII region parameters in Problem 19 below, will non-LTE effects be larger at smaller values of n?

19. Assume that the carbon 166 α recombination lines, at $v = 1.425 \, \text{GHz}$, are emitted from an isolated region, i.e. without a background source. The parameters of this region are $N_C = N_e = 1 \, \text{cm}^{-3}$, $L = 0.4 \, \text{pc}$ and $T_e = 100 \, \text{K}$. Using the LTE relation, what is T_{line} and $T_{\text{continuum}}$ if $\Delta v = 4.7 \, \text{kHz} = 1 \, \text{km s}^{-1}$? Now use the appropriate non-LTE coefficients, $b_n = 0.75$, $\beta = -7$, and repeat the calculation (Use equations referenced in Problem 17).

20. Modify (Eq. 14.44), with the source function, $S \ll I$, to show that $I = I_0 \, e^{\kappa_v \beta b L}$. This is the situation in which there is an intense background source with $T_{\text{BG}} \gg T_e$. Then repeat Problem 19 for $T_{\text{BG}} = 2500 \, \text{K}$.

21. There are a few neutral clouds along the line of sight to the supernova remnant Cassiopeia A. Assume that these are the only relevant carbon recombination line sources. These clouds are known to have H_2 densities of $\sim 4 \times 10^3 \, \text{cm}^{-3}$, column densities of $\sim 4 \times 10^{21} \, \text{cm}^{-2}$, diameters of 0.3 pc. If we assume that *all* of the carbon is ionized, we have $C^+/H = 3 \times 10^{-4}$. At wavelengths of more than a few meters, the carbon lines are in absorption. Assume that the line formation is hydrogen-like. For the C166α line, we estimate that the peak line temperature is 3 K, and the Doppler FWHP is $\Delta V_{1/2} = 3.5 \, \text{km s}^{-1}$(see Kantharia et al. (1998) for a model and references).
(a) Show that for $n > 300$ collisions dominate radiative decay, so that the populations are thermalized, but that the populations are dominated by radiative decay for $n < 150$.
(b) An observer claims that "Since the C166α line is in *emission*, the excitation temperature must be larger than the background temperature, or negative." Do you agree or disagree? Cite equations to justify your decision.
(c) If $hv \ll kT$ and LTE conditions hold, show that a reformulation of (Eq. 14.28) gives the relation between column density and line intensity

$$T_L = \frac{576}{T_e^{3/2}} \frac{\text{EM}}{v_0 \Delta V_{1/2}} \,,$$

where $\Delta V_{1/2}$ is in km s^{-1}, v_0 in GHz, EM in cm^{-6} pc and all temperatures in Kelvin.
(d) Estimate the maximum brightness temperature of Cas A at 1.425 GHz. Assuming that the level populations are *not* inverted, the excitation temperature, T_{ex}, of the C166α transition is >0. Given the background continuum temperature of Cassiopeia A, estimate a lower limit for T_{ex} from the continuum brightness temperature of Cassiopeia A. Next, use the cloud parameters in part **(a)** to determine the emission measure of C^+. Finally, make use of the expression in part **(c)** to determine the

integrated C166α line intensity. From a comparison with the observed result, is it more reasonable to assume that $T_{ex} > 0$ or that population inversion is more likely? In this case, population inversion will give rise to line masering effects.

22. The LTE version of the Saha equation is (Eq. 14.24). Assume that $N_e = N_p$.
(a) Evaluate the constants in this relation.
(b) For hydrogen, for $T = 5000$ K, 10^4 K and 2×10^4 K, determine the ratio of populations in the $n = 1$ and $n = 100$ states with respect to the number of ions (i.e. the left hand side of the relation). From these results can you explain why the Orion nebula is fully ionized given that the temperature of the gas is 8000 K?

Chapter 15
Overview of Molecular Basics

We begin with a summary of the basic facts of molecular line physics. In the next chapter we will present an overview of molecular line astronomy. For more complete accounts of molecular structure relevant to the microwave region, consult the extensive treatments by Herzberg and Herzberg (1960), Bingel (1969), Townes and Schawlow (1975), Gordy and Cook (1970), Flygare (1978) and Kroto (1992). The texts by Gordy and Cook (1970) and Kroto (1992) have the most modern notation, the text by Flygare (1978) treats more topics, the introduction in Kroto has applications to astronomy, but the presentation by Townes and Schawlow (1975) is the standard. In the following presentation, we start with simpler species, extend this treatment to include vibrational states, and then include symmetric and asymmetric top molecules. In the Section on symmetric top molecules, we include nuclear spin. As examples, we treat molecules that are widespread in the Interstellar Medium (ISM). For linear molecules, examples are carbon monoxide, CO, SiO, N_2H^+; for symmetric top molecules, these are ammonia, NH_3, CH_3CN and CH_3CCH and for asymmetric top molecules these are water vapor, H_2O, formaldehyde, H_2CO and H_2D^+. Finally we have a short secton on molecules with non-zero electronic angular momentum in the ground state, using OH as an example, and then present an account of methanol, CH_3OH which has hindered motion. In each section, we relate the molecular properties to a determination of column densities and local densities.

15.1 Basic Concepts

Compared to atoms, molecules have a complicated structure and the Schrödinger equation of the system will be correspondingly complex, involving positions and moments of all constituents, both the nuclei and the electrons. All particles, however, are confined to a volume with the diameter of a typical molecule diameter a, and therefore each particle will possess an average momentum $\hbar/a$ due to the uncertainty principle $\Delta p \Delta q \geq \hbar$. The kinetic energy will then have states with a typical spacing $\Delta E \cong \Delta p^2/2m \cong \hbar^2/2ma^2$. For electrons these energies are about an eV, corresponding to a temperature of 10^4 Kelvins; for rotational energies of the nuclei they are in the milli eV (10 K) range.

In the Schrödinger equation of a molecule, therefore, those parts of the Hamiltonian operator that describe the kinetic energy of the nuclei can be neglected compared to the kinetic energy of the electrons. The nuclei enter into the Coulomb potential only through their position and therefore these positions enter as parameters into the solution. Because the motion of the nuclei is so slow, the electrons make many cycles while the nuclei move to their new positions. This separation of the nuclear and electronic motion in molecular quantum mechanics is called the Born-Oppenheimer approximation.

Transitions in a molecule can therefore be put into three different categories according to different energies, W:

a) electronic transitions with typical energies of a few eV – that is lines in the visual or UV regions of the spectrum;
b) vibrational transitions caused by oscillations of the relative positions of nuclei with respect to their equilibrium positions. Typical energies are $0.1 - 0.01$ eV, corresponding to lines in the infrared region of the spectrum;
c) rotational transitions caused by the rotation of the nuclei with typical energies of $\cong 10^{-3}$ eV corresponding to lines in the cm and mm wavelength range.

$$W^{\text{tot}} = W^{\text{el}} + W^{\text{vib}} + W^{\text{rot}}. \tag{15.1}$$

W^{vib} and W^{rot} are the vibrational and rotational energies of the nuclei of the molecule and W^{el} is the energy of the electrons. Under this assumption, the Hamiltonian is a sum of $W^{\text{el}} + W^{\text{vib}} + W^{\text{rot}}$. From quantum mechanics, the resulting wavefunction will be a product of the electronic, vibrational and rotational wavefunctions.

In general, molecular line radiation arises from a transition between two states described by different electronic, vibrational and rotational quantum numbers. If transitions involve different electronic states, the corresponding spectral lines will lie in the optical range. If we confine ourselves to the centimeter/millimeter/sub-mm wavelength ranges, only transitions between different rotational levels and perhaps different vibrational levels (e.g., rotational transitions of SiO or HC_3N from vibrationally excited states) will be involved. This restriction results in a much simpler description of the molecular energy levels. Occasionally differences between geometrical arrangements of the nuclei result in a doubling of the energy levels. An example of such a case is the inversion doubling found for the Ammonia molecule.

The position of the nuclei of a molecule in equilibrium can usually be approximated by some average distance r between the nuclei and the potential energy is then specified as $P(r)$ (Fig. 15.1). If r_e is the equilibrium value, then

$$D_e = P(\infty) - P(r_e) \tag{15.2}$$

is the dissociation energy of the molecule. This is a unique value if we neglect, for the time being, that in a molecule consisting of more than 2 nuclei, D_e might depend on which of the nuclei would increase its distance r.

Fig. 15.1 The *solid line* shows the potential curve $P(r)$ of a binary molecule. The *dashed line* shows a harmonic approximation. The quantities r_e and D_e are explained in the text
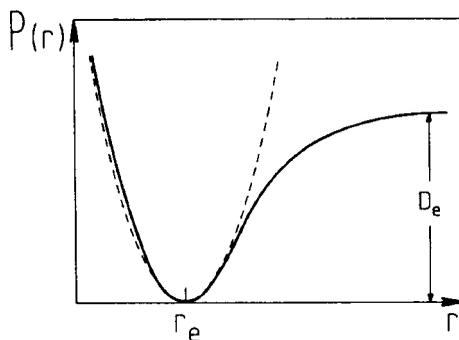
For diatomic molecules the potential curve $P(r)$ is well represented by the Morse potential:

$$P(r) = D_e \left[1 - \exp[-a(r - r_e)]\right]^2 \tag{15.3}$$

where D_e is the value of the potential energy at large distances (in the equilibrium, or zero vibration state), and r_e is the corresponding minimum point of potential energy. Often the even simpler harmonic approximation

$$P(r) = \frac{k}{2}(r - r_e)^2 = a^2 D_e (r - r_e)^2 \tag{15.4}$$

is sufficient.

15.2 Rotational Spectra of Diatomic Molecules

Because the effective radius of even a simple molecule is about 10^5 times the radius of the nucleus of an atom, the moment of inertia Θ_e of such a molecule is at least 10^{10} times that of an atom of the same mass. The kinetic energy of rotation is

$$H_{rot} = \tfrac{1}{2}\Theta_e \omega^2 = J^2/2\Theta_e, \tag{15.5}$$

where J is the angular momentum. J is a quantity that cannot be neglected compared with the other internal energy states of the molecule, especially if the observations are made in the centimeter/millimeter/sub-mm wavelength ranges. (Note that J is *not* the same as the quantum number used in atomic physics.)

For a rigid molecule consisting of two nuclei A and B, the moment of inertia is

$$\Theta_e = m_A r_A^2 + m_B r_B^2 = m r_e^2 \tag{15.6}$$

where

$$r_e = r_A - r_B \tag{15.7}$$

and

$$m = \frac{m_A m_B}{m_A + m_B},$$ (15.8)

and

$$\mathbf{J} = \Theta_e \boldsymbol{\omega}$$ (15.9)

is the angular momentum perpendicular to the line connecting the two nuclei. For molecules consisting of three or more nuclei, similar, more complicated expressions can be obtained. Θ_e will depend on the relative orientation of the nuclei and will in general be a (three-axial) ellipsoid. In (15.9) values of Θ_e appropriate for the direction of $\boldsymbol{\omega}$ will then have to be used.

This solution of the Schrödinger equation then results in the *eigenvalues* for the rotational energy

$$E_{rot} = W(J) = \frac{\hbar^2}{2\Theta_e} J(J+1),$$ (15.10)

where J is the quantum number of angular momentum, which has integer values

$$J = 0, 1, 2, \dots.$$

Equation (15.10) is correct only for a molecule that is completely rigid; for a slightly elastic molecule, r_e will increase with the rotational energy due to centrifugal stretching. (There is also the additional complication that even in the ground vibrational state there is still a zero point vibration; this will be discussed after the concept of centrifugal stretching is presented.) For centrifugal stretching, the rotational energy is modified to first order as:

$$E_{rot} = W(J) = \frac{\hbar^2}{2\Theta_e} J(J+1) - hD[J(J+1)]^2.$$ (15.11)

Introducing the rotational constant

$$B_e = \frac{\hbar}{4\pi\Theta_e}$$ (15.12)

and the constant for centrifugal stretching D, the pure rotation spectrum for electric dipole transitions $\Delta J = +1$ (emission) or $\Delta J = -1$ (absorption) is given by the following expression:

$$v(J) = \frac{1}{h}[W(J+1) - W(J)] = 2B_e(J+1) - 4D(J+1)^3.$$ (15.13)

Since D is positive, the observed line frequencies will be lower than those predicted on the basis of a perfectly rigid rotator. Typically, the size of D is about 10^{-5} of the magnitude of B_e for most molecules. In Fig. 15.2, we show a parameterized

Fig. 15.2 A schematic plot of rotational energy levels for a molecular rotator. The horizontal bars in the upper part represent the rotational energy levels for a rigid rotator (*right part*) and one deformed by centrifugal stretching with $D/B_e = 10^{-3}$ (*left part*). In fact, most molecules have $D/B_e \approx 10^{-5}$. The resulting line frequencies, v, are shown in the lower part. The numbers next to v refer to the J values
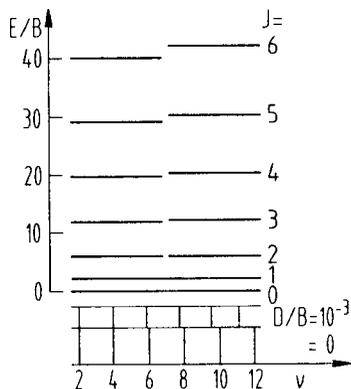
plot of the behavior of energy above ground and line frequency of a rigid rotor with and without the centrifugal distortion term. The function plotted vertically on the left, E_{rot}/B_e, is proportional to the energy above the molecular ground state. This function is given by

$$E_{rot}/B_e = 2\pi\hbar J(J+1) - 2\pi\hbar D/B_e \left[J(J+1)\right]^2 \tag{15.14}$$

while those on the right are given by the first term of (15.14) only. Directly below the energy level plots is a plot of the line frequencies for a number of transitions with quantum number J. The deviation between rigid rotor and actual frequencies becomes rapidly larger with increasing J, and in the sense that the actual frequencies are always lower than the frequencies predicted on the basis of a rigid rotor model. In Fig. 15.3 we show plots of the energies above ground state for a number of diatomic and triatomic linear molecules. As mentioned previously, because of zero-point vibrations, actual measurements give the value B_0, not B_e. In the molecular vibrational ground state, these are related by $B_0 = B_e - \alpha_e$, where the value of α_e is less than 1% of the value of B_e; see the discussions in Townes and Schawlow or Kroto for details.

Allowed dipole radiative transitions will occur between different rotational states only if the molecule possesses a permanent electric dipole moment; that is, the molecule must be polar. Homonuclear diatomic molecules like H_2, N_2 or O_2 do not possess a permanent electric dipole moment. Thus they cannot undergo allowed transitions. This is one reason why it was so difficult to detect these species. In the interstellar medium, the H_3^+ molecule has been identified on the basis of vibrational transitions in the near-infrared and the deuterium isotoomer, H_2D^+, has been detected on the basis of mm/sub-mm transitions.

For molecules with permanent dipole moments, a classical picture of molecular line radiation can be used to determine the angular distribution of the radiation. In the plane of rotation, the dipole moment can be viewed as an antenna, oscillating as the molecule rotates. Classically, the acceleration of positive and negative charges

Fig. 15.3 Rotational energy levels of the vibrational ground states of some linear molecules which are commonly found in the interstellar medium (ISM). To convert the vertical scale from GHz to Kelvins, multiply by 4.8×10^{-2}
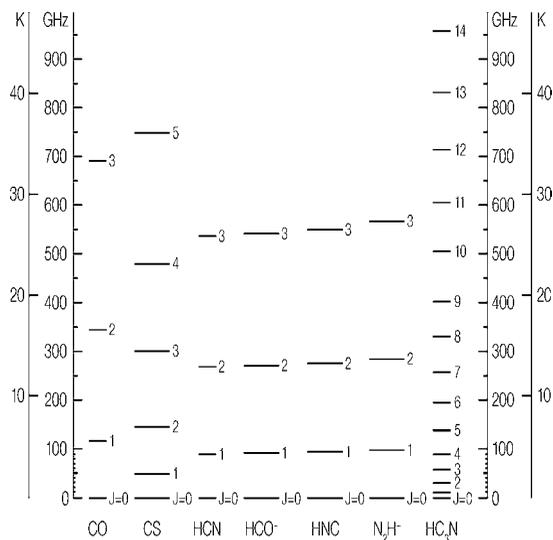
gives rise to radiation whose frequency is that of the rotation frequency. For a dipole transition the most intense radiation occurs in the plane of rotation of the molecule. In the quantum mechanical model, the angular momentum is quantized, so that the radiation is emitted at discrete frequencies. Dipole radiative transitions occur with a change in the angular momentum quantum number of one unit, that is, $\Delta J = \pm 1$. The parity of the initial and final states must be opposite for dipole radiation to occur.

15.2.1 Hyperfine Structure in Linear Molecules

The magnetic dipole or electric quadrupole of nuclei interact with electrons or other nuclei. These give rise to hyperfine structure. For example, the ^{14}N and Deuterium nuclei have spin $I = 1$ and thus a nonzero quadrupole moment. This gives rise to quadrupole splitting in molecules such as HCN, HNC and HC$_3$N. For nuclei with electric quadrupole moments, such as ^{14}N, the hyperfine splitting of energy levels depends on the position of the nucleus in the molecule. The effect is smaller for HNC than for HCN. In general, the effect is of order of a few MHz, and decreases with increasing J. For nuclei with magnetic dipole moments, such as ^{13}C or ^{17}O, the hyperfine splitting is smaller. In the case of hyperfine structure, the total quantum number $\boldsymbol{F} = \boldsymbol{J} + \boldsymbol{I}$ is conserved. Allowed transitions obey the selection rule $\Delta \boldsymbol{F} = \pm 1, 0$ but not $\Delta \boldsymbol{F} = 0 \overrightarrow{\leftarrow} 0$.

Fig. 15.4 Vibrational energy
levels, eigenstates (- - -) and
probability densities (–) for a
harmonic oscillator
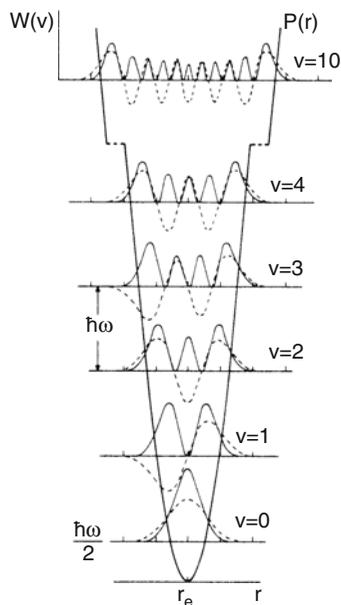
15.3   Vibrational Transitions

If any of the nuclei of a molecule suffers a displacement from its equilibrium distance r_e, it will on release perform an oscillation about r_e. The Schrödinger equation for this is

$$\left(\frac{p^2}{2m} + P(r)\right)\psi^{\mathrm{vib}}(x) = W^{\mathrm{vib}}\,\psi^{\mathrm{vib}}(x),\tag{15.15}$$

where $x = r - r_e$ and $P(r)$ is the potential function given in (15.3). If (15.4) can be used, we have the simple harmonic approximation (Fig. 15.4) with the classical oscillation frequency

$$\omega = 2\pi\nu = \sqrt{\frac{k}{m}} = a\sqrt{\frac{2D_e}{m}},\tag{15.16}$$

and (15.15) has the eigenvalue

$$W^{\mathrm{vib}} = W(v) = \hbar\omega\left(v + \tfrac{1}{2}\right)\tag{15.17}$$

with

$$v = 0, 1, 2, \dots.\tag{15.18}$$

The solutions $\psi^{\mathrm{vib}}(x)$ can be expressed with the help of Hermite polynomials. For the same rotational quantum numbers, lines arising from transitions in different vibrational states, in a harmonic potential, are separated by a constant frequency interval.

For large x, the precision of relation (15.4) is no longer sufficient and the Morse potential (15.3), or even an empirical expression, will have to be introduced into (15.15). The resulting differential equation can no longer be solved analytically, so numerical methods have to be used. Often it is sufficient to introduce *anharmonic* factors x_e, y_e into the solution corresponding to (15.17), viz.

$$W(v) = \hbar\omega\left(v + \tfrac{1}{2}\right) + x_e\hbar\omega\left(v + \tfrac{1}{2}\right)^2 + y_e\hbar\omega\left(v + \tfrac{1}{2}\right)^3 + \dots. \tag{15.19}$$

Usually x_e and y_e are small numbers which can be determined either empirically or by a fit to a numerical solution of (15.15). For example, for the H_2 molecule in the ground state we have $x_e = -2.6 \times 10^{-2}$ and $y_e = 6.6 \times 10^{-5}$. The negative sign of x results in a decrease of the step size of the harmonic energy ladder.

A molecule consisting of only two nuclei can vibrate only in one direction; it has only one vibrational mode. The situation is more complex for molecules with three or more nuclei. In this case, a multitude of various vibrational modes may exist, each of which will result in its own ladder of vibrational states, some of which may be degenerate. For a certain molecular vibrational state, there are many internal rotational states. Vibrational motions along the molecular axis can be and usually are hindered in the sense that these are subject to centrifugal forces, and thus must overcome an additional barrier.

It is possible to have rotational transitions between energy levels in a vibrationally excited state. An example is the $J = 1 - 0$ rotational transition of the SiO molecule from the $v = 0, 1$ and 2 levels. The dipole moment in a vibrational state is usually the same as in the ground state. The dipole moment for a purely vibrational transition in the case of diatomic molecules is usually about 0.1 Debye. A more complex example is the polyatomic linear molecule HC_3N, for which a number of transitions have been measured in the ISM (see Lafferty and Lovas (1978) for details of molecular structure and Lovas (1992) for references to astronomical measurements).

15.4 Line Intensities of Linear Molecules

In this section, we will give the details needed to relate the observed line intensities to column densities of the species emitting the transition. In the Born-Oppenheimer approximation, the total energy can be written as a sum of the electronic, vibrational and rotational energies in (15.1). In the line spectrum of a molecule, transitions between electronic, vibrational and rotational states are possible. We will restrict the discussion to rotational transitions, and in a few cases to vibrational transitions.

Computations of molecular line intensities proceed following the principles outlined in Chap. 12. The radial part of molecular wavefunctions is extremely complex. For any molecular or atomic system, the spontaneous transition probability, in s^{-1}, for a transition between an upper, u, and lower, l, level is given by the Einstein A coefficient A_{ul}. In the CGS system of units, A_{ul} is given by

$$A_{ul} = \frac{64\,\pi^4}{3\,h\,c^3}\,\nu^3\,|\mu_{ul}|^2\,.$$
(15.20)

After inserting numerical constants into the above relation, we have:

$$A_{ul} = 1.165 \times 10^{-11}\,\nu^3|\mu_{ul}|^2\,.$$
(15.21)

The units of the line frequency ν are GHz, and the units of μ are Debyes (i.e., 1 Debye $= 10^{-18}$ e.s.u.) Eq. 15.21 is a completely general relation for *any* transition. The expression $|\mu|^2$ contains a term which depends on the integral over the angular part of the wavefunctions of the final and initial states; the radial part of the wavefunctions is contained in the value of the dipole moment, μ (This is usually determined from laboratory measurements). For dipole transitions between two rotational levels of a linear molecule, $J \overrightarrow{\leftarrow} J + 1$, there can be either absorption or emission. For the case of absorption, for a dipole moment $|\mu_{ul}|^2 = |\mu_J|^2$, we have:

$$|\mu_J|^2 = \mu^2\,\frac{J+1}{2J+1} \qquad \text{for} \quad J \to J+1$$
(15.22)

while for emission, the expression is given by:

$$|\mu_J|^2 = \mu^2\,\frac{J+1}{2J+3} \qquad \text{for} \quad J+1 \to J$$
(15.23)

where ν_r is the spectral line frequency. Here, μ is the permanent electric dipole moment of the molecule. For a number of species μ is tabulated in Townes and Schawlow. In Table 16.2, we give a collection of Einstein A coefficients for some molecules of astrophysical significance, together with some other essential data.

After inserting (15.23) into (15.21) we obtain the expression for dipole emission between two levels of a linear molecule:

$$A_J = 1.165 \times 10^{-11}\,\mu^2\,\nu^3\,\frac{J+1}{2J+3} \qquad \text{for} \quad J+1 \to J$$
(15.24)

where A is in units of s^{-1}, μ_J is in Debyes (i.e. 10^{18} times e.s.u. values), and ν is in GHz. This expression is valid for a dipole transition in a linear molecule, from a level $J+1$ to J.

Inserting the expression for A in (12.17), the general relation between line optical depth, column density in a level l and excitation temperature, T_{ex}, is:

$$N_l = 93.5\,\frac{g_l\,\nu^3}{g_u\,A_{ul}}\,\frac{1}{[1 - \exp(-4.80 \times 10^{-2}\nu/T_{\text{ex}})]}\int \tau\,d\nu$$
(15.25)

where the units for ν are GHz and the linewidths are in km s^{-1}. n is the local density in units of cm^{-3}, and $N = nl$ is the column density, in cm^{-2}.

Although this expression appears simple, this is deceptive, since there is a dependence on T_{ex}. The excitation process may cause T_{ex} to take on a wide range of values. If $T_{ex}/v \gg 4.80 \times 10^{-2}$ K, the expression becomes:

$$N_l = 1.94 \times 10^3 \frac{g_l \, v^2 \, T_{ex}}{g_u A_{ul}} \int \tau \, dv \qquad . \tag{15.26}$$

Values for T_{ex} are difficult to obtain in the general case. Looking ahead a bit, for the $J = 1 \rightarrow 0$ and $J = 2 \rightarrow 1$ transitions, CO molecules are found to be almost always close to LTE, so it is possible to obtain estimates of T_K from (15.30). This result could be used in (15.26) if the transition is close to LTE. Expression (15.26) can be simplified even further if $\tau \ll 1$. Then, if the source fills the main beam (this is the usual assumption) the following relation holds:

$$T_{ex} \, \tau \cong T_{MB} \tag{15.27}$$

where the term T_{MB} represents the main beam brightness temperature. In the general case, we will use T_B, which depends on source size. For a discussion, see Sects. 8.2.4, 8.2.5, and 12.4.1. Inserting this in (15.26), we have:

$$N_l = 1.94 \times 10^3 \frac{g_l \, v^2}{g_u A_{ul}} \int T_B \, dv \qquad . \tag{15.28}$$

In this relation, T_{ex} appears nowhere. Thus, for an optically thin emission line, excitation plays *no role* in determining the column density in the energy levels giving rise to the transition. The units are as before; the column density, N_l, is an average over the telescope beam.

15.4.1 Total Column Densities of CO Under LTE Conditions

We apply the concepts developed in the last section to carbon monoxide, a simple molecule that is abundant in the ISM. Microwave radiation from this molecule is rather easily detectable because CO has a permanent dipole moment of $\mu = 0.112$ Debye. CO is a diatomic molecule with a simple ladder of rotational levels spaced such that the lowest transitions are in the millimeter wavelength region. A first approximation of the abundance of the CO molecules can be obtained by a very standard LTE analysis of the CO line radiation; this is also fairly realistic since the excitation of low rotational transitions is usually close to LTE. Stable isotopes exist for both C and O and several isotopic species of CO have been measured in the interstellar medium; among these are $^{13}C^{16}O$, $^{12}C^{18}O$, $^{12}C^{17}O$, $^{13}C^{16}O$ and $^{13}C^{18}O$. The temperature scales used in mm-wave radio astronomy have been treated in Sects. 8.2

and 13.4.1. For the distribution of CO, we adopt the simplest geometry, that is, an isothermal slab which is much larger than the telescope beam. Then the solution (1.37) may be used. If we recall that a baseline is usually subtracted from the measured line profile, and that the 2.7 K microwave background radiation is present everywhere, the appropriate formula is

$$T_B(v) = T_0 \left(\frac{1}{e^{T_0/T_{ex}} - 1} - \frac{1}{e^{T_0/2.7} - 1} \right) (1 - e^{-\tau_v}), \qquad (15.29)$$

where $T_0 = hv/k$. On the right side of (15.29) there are two unknown quantities: the excitation temperature of the line, T_{ex}, and the optical depth, τ_v. If τ_v is known it is possible to solve for the column density N_{CO} as in the case of the line $\lambda = 21$ cm of H I. But in the case of CO we meet the difficulty that lines of the most abundant isotope $^{12}C^{16}O$ always seem to be optically thick. It is therefore not possible to derive information about the CO column density from this line without a model for the molecular clouds. Here we give an analysis based on the measurement of weaker isotope lines of CO. This procedure can be applied if the following assumptions are valid.

- All molecules along the line of sight possess a uniform excitation temperature in the $J = 1 \rightarrow 0$ transition.
- The different isotopic species have the same excitation temperatures. Usually the excitation temperature is taken to be the kinetic temperature of the gas, T_K.
- The optical depth in the $^{12}C^{16}O$ $J = 1 \rightarrow 0$ line is large compared to unity.
- The optical depth in a rarer isotopomer transition, such as the $^{13}C^{16}O$ $J = 1 \rightarrow 0$ line is small compared to unity.
- The ^{13}CO and CO lines are emitted from the same volume.

Given these assumptions, we have $T_{ex} = T_K = T$, where T_K is the kinetic temperature, which is the only parameter in the Maxwell-Boltzmann relation for the cloud in question. In the remainder of this section and in the following section we will use the expression T, since all temperatures are assumed to be equal. This is certainly *not* true in general. Usually, the molecular energy level populations are often characterized by *at least* one other temperature, T_{ex}.

In general, the lines of $^{12}C^{16}O$ are optically thick. Then, in the absence of background continuum sources, the excitation temperature can be determined from the appropriate T_B^{12} of the optically thick $J = 1 - 0$ line of $^{12}C^{16}O$ at 115.271 GHz:

$$\boxed{T = 5.5 \Big/ \ln \left(1 + \frac{5.5}{T_B^{12} + 0.82} \right)} \qquad . \qquad (15.30)$$

The optical depth of the $^{13}C^{16}O$ line at 110.201 GHz is obtained by solving (15.29) for

$$\tau_0^{13} = -\ln\left[1 - \frac{T_B^{13}}{5.3}\left\{\left[\exp\left(\frac{5.3}{T}\right) - 1\right]^{-1} - 0.16\right\}^{-1}\right] \quad . \quad (15.31)$$

Usually the total column density is the quantity of interest. To obtain this for CO, one must sum over all energy levels of the molecule. This can be carried out for the LTE case in a simple way. For non–LTE conditions, the calculation is considerably more complicated. We discuss some cases in the next Chapter. In this section, we concentrate on the case of CO populations in LTE.

For CO, there is no statistical weight factor due to spin degeneracy. In a level J, the degeneracy is $2J + 1$. Then the fraction of the total population in a particular state, J, is given by:

$$N(J)/N(\text{total}) = \frac{(2J+1)}{Z} \exp\left[-\frac{hB_eJ(J+1)}{kT}\right] . \quad (15.32)$$

Z is the sum over all states, or the Partition function. If vibrationally excited states are not populated, Z can be expressed as:

$$Z = \sum_{J=0}^{\infty} (2J+1) \exp\left[-\frac{hB_e J(J+1)}{kT}\right] . \quad (15.33)$$

The total population, $N(\text{total})$ is given by the measured column density for a specific level, $N(J)$, divided by the calculated fraction of the total population in this level:

$$N(\text{total}) = N(J) \frac{Z}{(2J+1)} \exp\left[\frac{hB_e J(J+1)}{kT}\right] . \quad (15.34)$$

This fraction is based on the assumption that all energy levels are populated under LTE conditions. For a temperature, T, the population will increase as $2J + 1$, until the energy above the ground state becomes large compared to T. Then the negative exponential becomes the significant factor and the population will quickly decrease. If the temperature is large compared to the separation of energy levels, the sum can be approximated by an integral,

$$Z \approx \frac{kT}{hB_e} \qquad \text{for} \quad hB_e \ll kT . \quad (15.35)$$

Here B_e is the rotation constant (15.12), and the molecular population is assumed to be characterized by a single temperature, T, so that the Boltzmann distribution can be applied. Applying (15.34) to the $J = 0$ level, we can obtain the total column density of ^{13}CO from a measurement of the $J = 1 \to 0$ line of CO and ^{13}CO, using the partition function of CO, from (15.35), and (15.25):

$$N(\text{total})^{13}_{\text{CO}} = 3.0 \times 10^{14} \frac{T \int \tau^{13}(v)\,dv}{1 - \exp\{-5.3/T\}} \quad . \tag{15.36}$$

It is often the case that in dense molecular clouds ^{13}CO is optically thick. Then we should make use of an even rarer substitution, C^{18}O. For the $J = 1 \rightarrow 0$ line of this substitution, the expression is exactly the same as (15.36). For the $J = 2 \rightarrow 1$ line, we obtain a similar expression, using (15.34):

$$N(\text{total})^{13}_{\text{CO}} = 1.5 \times 10^{14} \frac{T \exp\{5.3/T\} \int \tau^{13}(v)\,dv}{1 - \exp\{-10.6/T\}} \quad . \tag{15.37}$$

In both (15.36) and (15.37), the beam averaged column density of carbon monoxide is in units of cm^{-2} the line temperatures are in Kelvin, main beam brightness temperature and the velocities, v, are in km s^{-1}. If the value of $T \gg 10.6$ or 5.3 K, the exponentials can be expanded to first order and then these relations are considerably simplified.

Furthermore, in the limit of optically thin lines, integrals involving $\tau(v)$ are equal to the integrated line intensity $\int T_{\text{MB}}(v)\,dv$, as mentioned before. However, there will be a dependence on T_{ex} in these relations because of the Partition function. The relation $T\,\tau(v) = T_{\text{MB}}(v)$ is only approximately true. However, optical depth effects can be eliminated to some extent by using the approximation

$$T \int_{-\infty}^{\infty} \tau(v)\,dv \cong \frac{\tau_0}{1 - e^{-\tau_0}} \int_{-\infty}^{\infty} T_{\text{MB}}(v)\,dv. \tag{15.38}$$

This formula is accurate to 15% for $\tau_0 < 2$, and it always overestimates N when $\tau_0 > 1$. The formulas (15.30), (15.31), (15.36) and (15.37) permit an evaluation of the column density N^{13}_{CO} *only* under the assumption of LTE.

15.4.1.1   Astronomical Considerations

CO is by far the most widespread molecule with easily measured transitions. However, even though the excitation of CO is close to LTE and the chemistry is thought to be well understood, there are limits to the accuracy with which one can measure the CO column densities, set by excitation and interstellar chemistry. Even if all of the concepts presented in this section are valid, there can be uncertainties in the calculation of the column densities of CO. These arise from several sources which can be grouped under the general heading *non-LTE effects*. Perhaps most important is the uncertainty in the excitation temperature. While the ^{12}CO emission might be thermalized even at densities $< 100\,\text{cm}^{-3}$, the less abundant isotopes may be

sub-thermally excited, i.e., populations characterized by $T_{ex} < T_K$. (This will be discussed in conjunction with the large velocity gradient (LVG) approximation in the next Chapter). Alternatively, if the cloud in question has no small scale structure, ^{13}CO emission will arise primarily from the cloud interior, which may be either hotter or cooler than the surface; the optically thick ^{12}CO emission may only reflect conditions in the cloud surface. Another effect is that, although T_{ex} may describe the population of the $J = 0$ and $J = 1$ states well, it may not for $J > 1$. That is, the higher rotational levels might not be thermalized because their larger Einstein A coefficients lead to a faster depopulation. This lack of information about the population of the upper states leads to an uncertainty in the partition function. Measurements of other transitions and use of a large velocity gradient model will allow better accuracy. For most cloud models, LTE gives overestimates of the true ^{13}CO column densities by factors from 1 to 4 depending on the properties of the model and of the position in the cloud. Thus a factor of $\sim$ two uncertainties should be expected when using LTE models.

As is well established, CO is dissociated by line radiation. Since the optical depth of CO is large, this isotopomer will be self-shielded. If there is no fine spatial structure, selective dissociation will cause the extent of ^{12}CO to be greater than that of ^{13}CO, which will be greater than the extent of $C^{18}O$. This will cause some uncertainty since the geometry of molecular clouds is complex. Finally, for H_2 densities $> 10^6\,cm^{-3}$, one might expect a freezing out onto grains for dense regions of small size. For such regions, the thermal emission from dust gives an alternative method to derive H_2 abundances. In spite of all these uncertainties, one most often attempts to relate measurements of the CO column density to that of H_2; estimates made using lines of CO (and isotopomers) are probably the best method to obtain the H_2 column density and mass of molecular clouds. This will be discussed further in Sect. 16.4.

For other linear molecules, such as HC_3N, HC_5N, etc., the expressions for the dipole moments and the partition functions are similar to that for CO and the treatment is similar to that given above. There is one very important difference however. The simplicity in the treatment of the CO molecule arises because of the assumption of LTE. This may not be the case for molecules such as HC_3N or HC_5N since these species have dipole moments of order 3 Debye. Thus populations of high J levels (which have faster spontaneous decay rates) may have populations lower than predicted by LTE calculations. Such populations are said to be *subthermal*, because the excitation temperature characterizing the populations would be $T_{ex} < T_K$.

15.5 Symmetric Top Molecules

15.5.1 Energy Levels

The rotation of a rigid molecule with an arbitrary shape can be considered to be the superposition of three free rotations about the three principal axes of the inertial

ellipsoid. Depending on the symmetry of the molecule these principal axes can all be different: in that case the molecule is an asymmetric top. If two principal axes are equal, the molecule is a symmetric top. If all three principal axes are equal, it is a spherical top. In order to compute the angular parts of the wavefunction, the proper Hamiltonian operator must be solved in the Schrödinger equation and the stationary state eigenvalues determined.

In general, for any rigid rotor asymmetric top molecule in a stable state, the total momentum J will remain constant with respect to both its absolute value and its direction. As is known from atomic physics, this means that both $(J)^2$ and the projection of J into an arbitrary but fixed direction, for example J_z, remain constant. If the molecule is in addition symmetric, the projection of J on the axis of symmetry will be constant also.

Let us first consider the *symmetric top molecule*. Suppose J is inclined with respect to the axis of symmetry z. Then the figure axis z will precess around the direction J forming a constant angle with it, and the molecule will simultaneously rotate around the z axis with the constant angular momentum J_z. From the definition of a symmetric top, $\Theta_x = \Theta_y$. Taking $\Theta_x = \Theta_y = \Theta_\perp$ and $\Theta_z = \Theta_\parallel$, we obtain a Hamiltonian operator:

$$H = \frac{J_x^2 + J_y^2}{2\Theta_\perp} + \frac{J_z^2}{2\Theta_\parallel} = \frac{J^2}{2\Theta_\perp} + J_z^2 \cdot \left(\frac{1}{2\Theta_\parallel} - \frac{1}{2\Theta_\perp} \right) . \tag{15.39}$$

Its eigenvalues are:

$$W(J,K) = J(J+1)\frac{\hbar^2}{2\Theta_\perp} + K^2\hbar^2 \left(\frac{1}{2\Theta_\parallel} - \frac{1}{2\Theta_\perp} \right) \tag{15.40}$$

where K^2 is the eigenvalue from the operator J_z^2 and $J^2 = J_x^2 + J_y^2 + J_z^2$ is the eigenvalue from the operator $J_x^2 + J_y^2 + J_z^2$.

The analysis of linear molecules is a subset of that for symmetric molecules. For linear molecules, $\Theta_\parallel \to 0$ so that $1/(2\Theta_\parallel) \to \infty$. Then finite energies in (15.40) are possible only if $K = 0$. For these cases the energies are given by (15.10). For symmetric top molecules each eigenvalue has a multiplicity of $2J + 1$.

$$J = 0, 1, 2, \ldots \quad K = 0, \pm 1, \pm 2, \cdots \pm J . \tag{15.41}$$

From (15.40), the energy is independent of the sign of K, so levels with the same J and absolute value of K coincide. Then levels with $K > 0$ are doubly degenerate.

It is usual to express $\dfrac{\hbar}{4\pi\Theta_\perp}$ as B, and $\dfrac{\hbar}{4\pi\Theta_\parallel}$ as C. The units of these rotational constants, B and C are usually either MHz or GHz. Then (15.40) becomes

$$W(J,K)/h = BJ(J+1) + K^2 (C - B). \tag{15.42}$$

15.5.2 Spin Statistics

In the case of molecules containing identical nuclei, the exchange of such nuclei, for example by the rotation about an axis, has a spectacular effect on the selection rules. Usually there are no interactions between electron spin and rotational motion. Then the total wavefunction is the product of the spin and rotational wavefunctions. Under an interchange of fermions, the total wave function must be antisymmetric (these identical nuclei could be protons or have an uneven number of nucleons). The symmetry of the spin wavefunction of the molecule will depend on the relative orientation of the spins. If the spin wavefunction is symmetric, this is the *ortho*-modification of the molecule; if antisymmetric it is the *para*-modification. In thermal equilibrium in the ISM, collisions with the exchange of identical particles will change one modification into the other only very slowly, on time scales of $> 10^6$ years. This could occur much more quickly on grain surfaces, or with charged particles. If the exchange is slow, the ortho and para modifications of a particular species behave like different molecules; a comparison of ortho and para populations might give an estimate of temperatures in the distant past, perhaps at the time of molecular formation.

For the H_2 molecule, the symmetry of the rotational wavefunction depends on the total angular momentum J as $(-1)^J$. In the $J = 0$ state the rotational wavefunction is symmetric. However, the total wavefunction must be antisymmetric since protons are fermions. Thus, the $J = 0, 2, 4$, etc., rotational levels are para-H_2, while the $J = 1, 3, 5$, etc., are ortho-H_2. Spectral lines can connect only one modification. In the case of H_2, dipole rotational transitions are not allowed, but quadrupole rotational transitions ($\Delta J = \pm 2$) are. Thus, the $28\,\mu m$ line of H_2 connects the $J = 2$ and $J = 0$ levels of para-H_2. Transitions between the ground and vibrational states are also possible.

Finally, as a more complex example of the relation of identical nuclei, we consider the case of three identical nuclei. This is the case for NH_3, CH_3CN and CH_3C_2H. Exchanging two of the nuclei is equivalent to a rotation by 120°. An exchange as was used for the case of two nuclei would not, in general, lead to a suitable symmetry. Instead combinations of spin states must be used (see Townes and Schawlow). These lead to the result that the ortho to para ratio is two to one if the identical nuclei are protons. That is, NH_3, CH_3CN or CH_3C_2H the ortho form has $S(J, K) = 2$, while the para form has $S(J, K) = 1$. In summary, the division of molecules with identical nuclei into ortho and para species determines selection rules for radiative transitions and also rules the for collisions.

15.5.3 Hyperfine Structure

For symmetric top molecules, the simplest hyperfine spectra is found for the inversion doublet transitions of NH_3. Since both the upper and lower levels have the same quantum numbers (J, K), there will be 5 groups of hyperfine components separated

Table 15.1 Intensities of satellite groups relative to the Main Component [after Mauersberger (1983)]

(J,K)	(1,1)	(2,2)	(3,3)	(4,4)	(5,5)	(6,6)	(2,1)
I_{inner}	0.295	0.0651	0.0300	0.0174	0.0117	0.0081	0.0651
I_{outer}	0.238	0.0628	0.0296	0.0173	0.0114	0.0081	0.0628

by a few MHz. Because of interactions between the spins of H nuclei there will be an additional splitting, within each group, of order a few kHz. In Table 15.1 we give the relative intensities of the NH_3 satellites for the case of low optical depth and LTE. For a molecule such as OH, one of the electrons is unpaired. The interaction of the nuclear magnetic moment with the magnetic moment of an unpaired electron is described as magnetic hyperfine structure. This splits a specific line into a number of components. In the case of the OH molecule, this interaction gives rise to a hyperfine splitting of the energy levels, in addition to the much larger Λ doubling. Together with the Λ doublet splitting, this gives rise to a quartet of energy levels in the OH ground state. Transitions between these energy levels produces the four ground state lines of OH at 18 cm wavelength (see Fig. 15.9).

NH_3 is an example of an oblate symmetric top molecule commonly found in the ISM. A diagram of the lower energy levels of NH_3 are shown in Fig. 15.5. A prolate top molecule has a cigar-like shape. Then A replaces C, and $A > B$. The energy-level diagrams for prolate symmetric top molecules found in the ISM, such as CH_3CCH and CH_3CN, follow this rule. However, since these molecules are much heavier than NH_3, the rotational transitions give rise to lines in the millimeter wavelength range (see Churchwell and Hollis (1983)).

Differences in the orientation of the nuclei can be of importance. If a reflection of all particles about the center of mass leads to a configuration which cannot be obtained by a rotation of the molecule, these reflections represent two different states. For NH_3, we show this situation in the upper part of Fig. 15.5. Then there are two separate, degenerate states which exist for each value of (J,K) for $J \geq 1$. (The $K = 0$ ladder is an exception because of the Pauli principle.) These states are doubly degenerate as long as the potential barrier separating the two configurations is infinitely high. However, in molecules such as NH_3 the two configurations are separated only by a small potential barrier. This gives rise to a measurable splitting of the degenerate energy levels, which is referred to as inversion doubling. For NH_3, transitions between these inversion doublet levels are caused by the quantum mechanical tunneling of the nitrogen nucleus through the plane of the three protons. The wavefunctions of the two inversion doublet states have opposite parities, so that dipole transitions are possible. Thus dipole transitions occur between states with the same (J,K) quantum numbers. The splitting of the (J,K) levels for NH_3 shown in Fig. 15.5 is exaggerated; the inversion transitions give rise to spectral lines in the wavelength range near 1 cm. For CH_3CCH or CH_3CN, the splitting caused by inversion doubling is very small since the barrier is much higher than for NH_3.
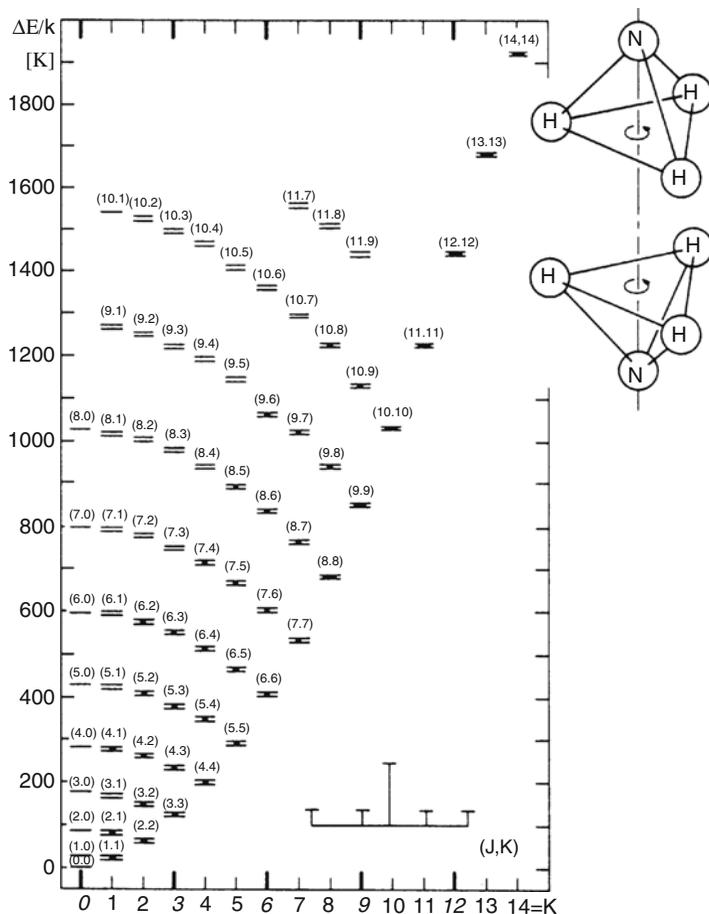
Fig. 15.5 The energy-level diagram of the vibrational ground state of NH_3, a prolate symmetric top molecule. Ortho-NH_3 has $K = 0, 3, 6, 9, \cdots$, while para-NH_3 has all other K values (see text). Rotational transitions with $\Delta J = 1, \Delta K = 0$, give rise to lines in the far IR. This molecule also has transitions with $\Delta J = 0, \Delta K = 0$ between inversion doublet levels. The interaction of the nuclear spin of ^{14}N with the electrons causes quadrupole hyperfine structure. In the $\Delta J = 0, \Delta K = 0$ transitions, the line is split into 5 groups of components. A sketch of the structure of the groups of hyperfine components of the $(J, K) = (1, 1)$ inversion doublet line is indicated in the *lower right*; the separation is of order of MHz. In the *upper right* is a sketch of the molecule before and after an inversion transition, which gives rise to a 1.3 cm photon [adapted from Wilson et al. (2006)]

The direction of the dipole moment of symmetric top molecules is parallel to the K axis. Spectral line radiation can be emitted only by a changing dipole moment. Since radiation will be emitted perpendicular to the direction of the dipole moment, there can be no radiation along the symmetry axis. Thus the K quantum number *cannot* change in dipole radiation, so allowed dipole transitions cannot connect different K ladders. The different K ladders are connected by octopole radiative transitions which require $\Delta K = \pm 3$. These are very slow, however, and collisions are far more likely to cause an exchange of population between different K ladders. We will discuss this in Sect. 15.6.3 in connection with the determination of T_K using the ratio of populations of different (J, K) states in symmetric top molecules.

15.5.4 Line Intensities and Column Densities

The extension of this analysis to symmetric top molecules is only slightly more complex. The dipole moment for an allowed transition between energy level $J+1, K$ and J, K for a symmetric top such as CH_3CN or CH_3C_2H is

$$|\mu_{JK}|^2 = \mu^2 \frac{(J+1)^2 - K^2}{(J+1)(2J+3)} \qquad \text{for} \quad (J+1, K) \to (J, K). \qquad (15.43)$$

For these transitions, $J \geq K$ always.

For NH_3, the most commonly observed spectral lines are the inversion transitions at 1.3 cm between levels (J, K) and (J, K). The dipole moment is

$$|\mu_{JK}|^2 = \mu^2 \frac{K^2}{J(J+1)} \qquad \text{for} \quad \Delta J = 0, \Delta K = 0. \qquad (15.44)$$

When these relations are inserted in (15.21), the population of a specific level can be calculated following (15.25). If we follow the analysis used for CO, we can use the LTE assumption to obtain the entire population

$$N(\text{total}) = N(J, K) \frac{Z}{(2J+1)\, S(J, K)} \exp\left[\frac{W(J, K)}{kT} \right], \qquad (15.45)$$

where W is the energy of the level above the ground state, and the nuclear spin statistics are accounted for through the factor $S(J, K)$ for the energy level corresponding to the transition measured. For symmetric top molecules, we have, using the expression for the energy of the level in question (15.42),

$$N(\text{total}) = \frac{Z N(J, K)}{(2J+1)\, S(J, K)} \exp\left[\frac{BJ(J+1) + (C - B)K^2}{kT} \right]. \qquad (15.46)$$

For prolate tops, A replaces C in (15.46), and in (15.47, 15.48, 15.49, 15.50, 15.51). If we sum over all energy levels, we obtain $N(\text{total})$, the partition function, Z in the following:

$$Z = \sum_{J=0}^{\infty} \sum_{K=0}^{K=J} (2J+1)\, S(J,K)\, \exp\left[-\frac{BJ(J+1)+(C-B)K^2}{kT}\right]. \qquad (15.47)$$

If the temperature is large compared to the spacing between energy levels, one can replace the sums by integrals, so that:

$$Z \approx \sqrt{\frac{\pi(kT)^3}{h^3 B^2 C}}. \qquad (15.48)$$

If we assume that $h\nu \ll kT$, use CGS units for the physical constants, and GHz for the rotational constants A, B and C, the partition function, Z, becomes

$$Z \approx 168.7 \sqrt{\frac{T^3}{B^2 C}}. \qquad (15.49)$$

Substituting into (15.47), we have:

$$N(\text{total}) = N(J,K)\, \frac{168.7 \sqrt{\dfrac{T^3}{B^2 C}}}{(2J+1)\, S(J,K)}\, \exp\left[\frac{W(J,K)}{kT}\right]. \qquad (15.50)$$

Here, $N(J,K)$ can be calculated from (15.25) or (15.26), using the appropriate expressions for the dipole moment, (15.44), in the Einstein A coefficient relation, (15.21) and W is the energy of the level above the ground state. In the ISM, ammonia inversions lines up to (J,K)=(18,18) have been detected (Wilson et al. 2006).

We now consider a situation in which the NH_3 population is *not* thermalized. This is typically the case for dark dust clouds. We must use some concepts presented in the next few sections for this analysis. If $n(H_2) \sim 10^4\,\text{cm}^{-3}$, and the infrared field intensity is small, a symmetric top molecule such as NH_3 can have a number of excitation temperatures. The excitation temperatures of the populations in doublet levels are usually between 2.7 K and T_K. The rotational temperature, T_{rot}, which describes populations for metastable levels ($J = K$) in different K ladders, is usually close to T_K. This is because radiative transitions between states with a different K value are forbidden to first order. The excitation temperature which describes the populations with different J values within a given K ladder will be close to 2.7 K, since radiative decay with $\Delta K=0$, $\Delta J=1$ is allowed. Then the non–metastable energy levels, ($J > K$), are not populated. In this case, Z is simply given by the sum over the populations of metastable levels:

$$Z(J = K)$$
$$= \sum_{J=0}^{\infty} (2J+1)\, S(J,K=J)\, \exp\left[-\frac{BJ(J+1)+(C-B)J^2}{kT}\right]. \qquad (15.51)$$

For the NH_3 molecule in dark dust clouds, where $T_K = 10\,K$ and $n(H_2) = 10^4\,cm^{-3}$, we can safely restrict the sum to the three lowest metastable levels:

$$Z(J = K) \approx N(0,0) + N(1,1) + N(2,2) + N(3,3). \qquad (15.52)$$

Substituting the values for NH_3 metastable levels:

$$Z(J = K)$$
$$\approx N(1,1) \left[\frac{1}{3} \exp\left(\frac{23.1}{kT} \right) + 1 + \frac{5}{3} \exp\left(-\frac{41.2}{kT} \right) + \frac{14}{3} \exp\left(-\frac{99.4}{kT} \right) \right]. \quad (15.53)$$

For NH_3 we have given two extreme situations: in the first case, described by (15.53), is a low-density cloud for which only the few lowest metastable levels are populated. The second case is the LTE relation, given in (15.50). This represents a cloud in which the populations of the molecule in question are thermalized. More complex are those situations for which the populations of some of the levels are thermalized, and others not. Methods needed to describe these situations will be discussed in Sect. 16.2.3.

15.6 Asymmetric Top Molecules

15.6.1 Energy Levels

For an *asymmetric top molecule* there are no internal molecular axes with a time-invariable component of angular momentum. So only the total angular momentum is conserved and we have only J as a good quantum number. The moments of inertia about each axis are different; the rotational constants are referred to as A, B and C, with $A > B > C$. The prolate symmetric top ($B = C$) or oblate symmetric top ($B = A$) molecules can be considered as the limiting cases. But neither the eigenstates nor the eigenvalues are easily expressed in explicit form. Each of the levels must be characterized by three quantum numbers. One choice is $J_{K_a K_c}$, where J is the total angular momentum, K_a is the component of J along the A axis and K_c is the component along the C axis. If the molecule were a prolate symmetric top, J and K_a would be good quantum numbers; if the molecule were an oblate symmetric top, J and K_c would be good quantum numbers. Intermediate states are characterized by a superposition of the prolate and oblate descriptions. In Fig. 15.7, we show the energy level diagram for the lower levels of H_2CO. H_2CO is almost a prolate symmetric top molecule with the dipole moment along the A axis. Since radiation must be emitted perpendicular to the direction of the dipole moment, for H_2CO there can be no radiation emitted along the A axis, so the quantum number K_a will not change in radiative transitions.

15.6.2 Spin Statistics and Selection Rules

The case of a planar molecule with two equivalent nuclei, such as H_2CO, shows is a striking illustration of these effects (see Fig. 15.7). The dipole moment lies along the A axis. A rotation by $180°$ about this axis will change nothing in the molecule, but will exchange the two protons. Since the protons are fermions, this exchange must lead to an antisymmetric wavefunction. Then the symmetry of the spin wave function and the wave function describing the rotation about the A axis must be antisymmetric. The rotational symmetry is $(-1)^{K_a}$. If the proton spins are parallel, that is ortho-H_2CO, then the wave function for K_a must be anti-symmetric, or K_a must take on an odd value. If the proton spins are anti-parallel, for para-H_2CO, K_a must have an even value (Fig. 15.7). For ortho-H_2CO, the parallel spin case, there are three possible spin orientations. For para-H_2CO, there is only one possible orientation, so the ratio of ortho-to-para states is three. Such an effect is taken into account in partition functions (15.33) by spin degeneracy factors, which are denoted by the symbol $S(J,K)$. For ortho-H_2CO, $S(J,K) = 3$, for para-H_2CO, $S(J,K) = 1$. This concept will be applied in Sect. 15.6.3.

Allowed transitions can occur only between energy levels of either the ortho or the para species. For example, the $6\,cm$ H_2CO line is emitted from the ortho-modification only (see Fig. 15.7). Another example is the interstellar H_2O maser line at $\lambda = 1.35\,cm$ which arises from ortho-H_2O (see Fig. 15.6). The H_2O molecule is a more complex case since the dipole moment is along the B axis. Then in a radiative transition, both K_a and K_c must change between the initial and final state.

15.6.3 Line Intensities and Column Densitiess

For asymmetric molecules the moments of inertia for the three axes are all different; there is no symmetry, so three quantum numbers are needed to define an energy level. The relation between energy above the ground state and quantum numbers is given in Appendix IV of Townes and Schawlow or in databases for specific molecules (see figure captions for references). The relation for the dipole moment of a specific transition is more complex; generalizing from (15.21), we have, for a spontaneous transition from a higher state, denoted by u to a lower state, denoted by l:

$$A = 1.165 \times 10^{-11}\, v_x^3\, \mu_x^2\, \frac{\mathscr{S}(u;l)}{2J'+1}\,. \tag{15.54}$$

This involves a dipole moment in a direction x. As before, the units of v are GHz, and the units of μ are Debyes (i.e., $1\,\mathrm{Debye} = 10^{-18}$ times the e.s.u. value). The value of the quantum number J' refers to the lower state. The expression for $\mathscr{S}(u;l)$, the line strength is an indication of the complexity of the physics of asymmetric top molecules. The expression $\mathscr{S}(u;l)$ is the angular part of the dipole moment between the initial and final state. The dipole moment can have a direction which is *not* along

Fig. 15.6 Energy level diagrams for ortho- and para-H_2O. This is an asymmetric top molecule, with the dipole moment along the B axis, that is, the axis with an intermediate moment of inertia. Because of the two identical nuclei, the energy level diagram is split into ortho and para, that show almost no interaction under interstellar conditions. The transitions marked by *arrows* are masers [adapted from Menten (1994)]. In the *upper part* of the diagram is a sketch of the structure of H_2O. See DeLucia et al. (1972) for details, including line strengths and selection rules
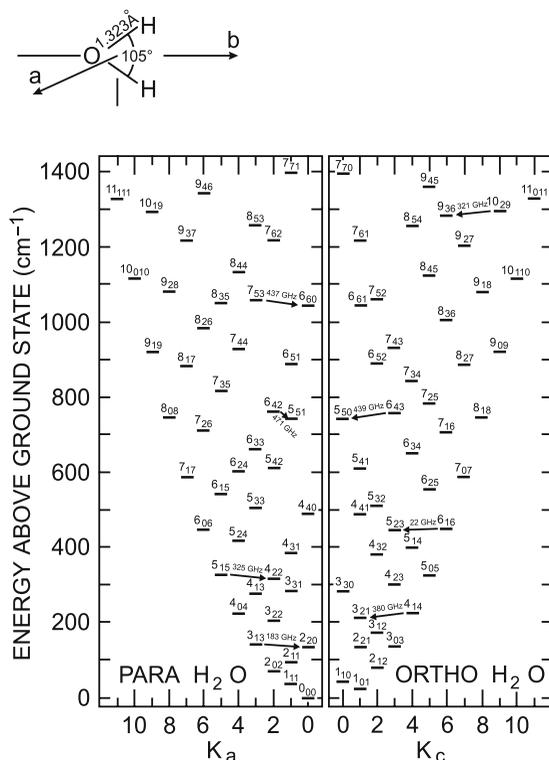
a single axis. In this case there are different values of the dipole moment along different molecular axes. In contrast, for symmetric top or linear molecules, there is *a* dipole moment for rotational transitions. Methods to evaluate transition probabilities for asymmetric molecules are discussed at length in Townes and Schawlow; a table of $\mathscr{S}(u;l)$ in their Appendix V. We give references for $\mathscr{S}(u;l)$ in our figure captions. From the expression for the Einstein A coefficient, the column density in a given energy level can be related to the line intensity by (15.25). Following the procedures used for symmetric top molecules, we can use a relation similar to (15.47) to sum over all levels, using the appropriate energy, W, of the level $J_{K_aK_c}$ above the ground state and the factor $S(J_{K_aK_c})$ for spin statistics:

$$N(\text{total}) = N(J_{K_aK_c}) \frac{Z}{(2J+1)\,S(J_{K_aK_c})} \exp\left(\frac{W}{kT}\right). \tag{15.55}$$

If the populations are in LTE, one can follow a process similar to that used to obtain (15.49). Then we obtain the appropriate expression for the partition function:

$$Z = 168.7 \sqrt{\frac{T^3}{ABC}}. \tag{15.56}$$

When combined with the Boltzmann expression for a molecule in a specific energy level, this gives a simple expression for the fraction of the population in a specific rotational state if LTE conditions apply:

$$N(\text{total}) \approx N(J_{K_a K_c}) \frac{168.7 \sqrt{\dfrac{T^3}{ABC}}}{(2J+1) S(J_{K_a K_c})} \exp\left(\frac{W}{kT}\right). \qquad (15.57)$$

In this expression, $S(J_{K_a K_c})$ accounts for spin statistics for energy level $J_{K_a K_c}$, and A, B and C are the molecular rotational constants in GHz. W, the energy of the level above the ground state, and T, the temperature, are given in Kelvin. Given the total molecular column density and the value of T, the feasibility of detecting a specific line can be obtained when the appropriate A coefficient value is inserted into (15.25) or (15.26).

As pointed out in connection with NH_3, T need not be T_K. In reality, a number of different values of T may be needed to describe the populations. We will investigate the influence of excitation conditions on molecular populations and observed line intensities next.

Two important interstellar molecules are H_2CO and H_2O. Here we summarize the dipole selection rules. Rotating the molecule about the axis along the direction of the dipole moment, we effectively exchange two identical particles. If these are fermions, under this exchange the total wavefunction must be antisymmetric. For H_2CO, in Sect. 15.5.2 we reviewed the spin statistics. Since the dipole moment is along the A axis, a dipole transition must involve a change in the quantum numbers along the B or C axes. From Fig. 15.7, the $K_a = 0$ ladder is para-H_2CO, so to have a total wavefunction which is antisymmetric, one must have a space wavefunction which is symmetric. For a dipole transition, the parities of the initial and final states must have different parities. This is possible if the C quantum number changes. For H_2O, the dipole moment is along the B axis, from Fig. 15.6. In a dipole transition, the quantum number for the B direction will not change. For ortho-H_2O, the spin wavefunction is symmetric, so the symmetry of the space wavefunction must be antisymmetric. In general, this symmetry is determined by the product of K_a and K_c. For ortho-H_2O, this must be $K_a K_c = (\text{odd})(\text{even})$, i.e. oe, or eo. For allowed transitions, one can have $oe - eo$ or $eo - oe$. For para-H_2O, the rule is $oo - ee$ or $ee - oo$. Clearly H_2S follows the selection rules for H_2O. These rules will be different for SO_2 since the exchanged particles are bosons. More exotic are D_2CO, ND_3 and D_2O (Butner et al. 2007).

The species H_3^+ has the shape of a planar triangle. It is a key to ion-molecule chemistry (see next Chapter), but has no rotational transitions because of its symmetry (see Oka et al. 2005). The deuterium isotopomer, H_2D^+ is an asymmetric top molecule with a permanent dipole moment. The spectral line from the 1_{10}-1_{11} levels ortho species was found at 372.421 GHz. A far infrared absorption line from the 2_{12}-1_{11} levels was also detected (Cernicharo et al. 2007; see references therein). We show an energy level diagram in Fig. 15.8. The doubly deuterated species, D_2H^+,has
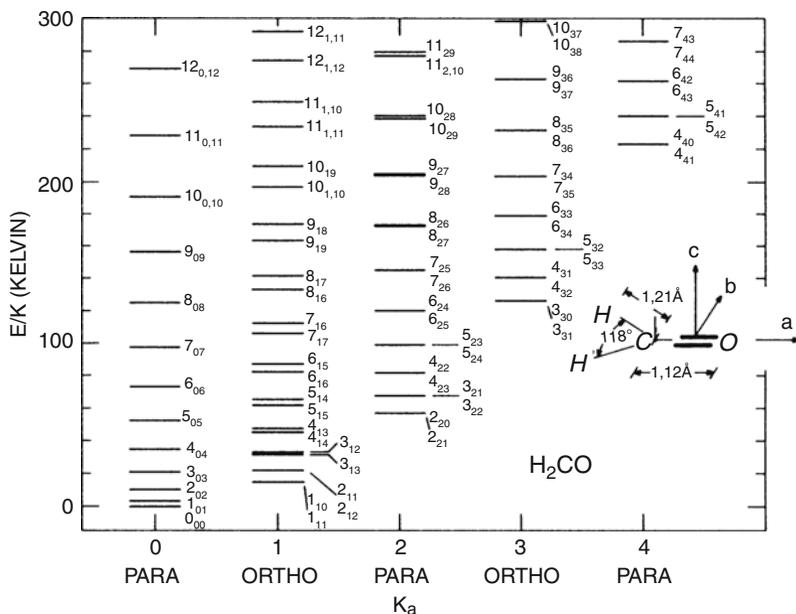
Fig. 15.7 Energy-level diagram of formaldehyde, H_2CO. This is a planar asymmetric top, but the asymmetry is very small. The energy-level structure is typical of an almost prolate symmetric top molecule. In the *lower right* is a sketch of the structure of the molecule [adapted from Mangum and Wootten (1993)]. See Johnson et al. (1972) for details about line strengths and selection rules
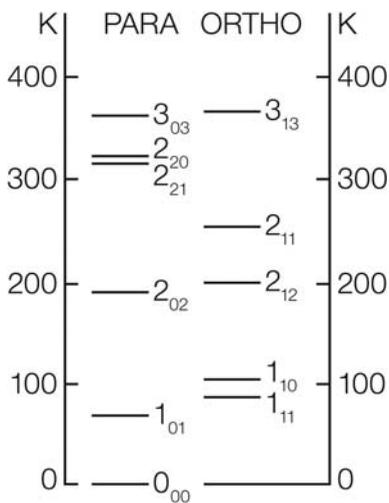
Fig. 15.8 Energy-level diagram of formaldehyde, H_2D^+. This is a planar, triangular-shaped asymmetric top. The diagram is adapted from Gerlich et al. (2006)

been detected in the 1_{10}-1_{01} line at 691.660 GHz from the para species (see, e.g. Vastel et al. 2006).

15.6.4 Electronic Angular Momentum

In many respects the description of electronic angular momentum is similar to that of atomic fine structure as described by Russell-Saunders (LS) coupling. Each electronic state is designated by the symbol $^{2S+1}\Lambda_\Omega$, where $2S+1$ is the multiplicity of the state with S the electron spin and Λ is the projection of the electronic orbital angular momentum on the molecular axis in units of $\hbar$. The molecular state is described as Σ, Π, Δ etc., according to whether $\Lambda = 0, 1, 2, \ldots$.

Σ is the projection of the electron spin angular momentum on the molecular axis in units of $\hbar$ (not to be confused with the symbol Σ, for $\Lambda = 0$). Finally, Ω is the total electronic angular momentum. For the Hund coupling case A, $\Omega = |\Lambda + \Sigma|$ [see e.g., Hellwege (1974) for other cases].

Since the frequencies emitted or absorbed by a molecule in the optical range are due to electronic state changes, many of the complications found in optical spectra are not encountered when considering transitions in the cm and mm range. However the electronic state does affect the vibrational and rotational levels even in the radio range. For most molecules, the ground state has zero electronic angular momentum, that is, a singlet sigma, $^1\Sigma$ state. For a small number of molecules such as OH, CH, C_2H, or C_3H, this is not the case; these have ground state electronic angular momentum. Because of this fact, the rotational energy levels experience an additional energy-level splitting, which is Λ doubling. This is a result of the interaction of the rotation and the angular momentum of the electronic state. This splitting causes the degenerate energy levels to separate. This splitting can be quite important for Π states; for Δ and higher states it is usually negligible. The OH molecule is a prominent example for this effect. Semi-classically, the Λ doubling of OH can be viewed as the difference in rotational energy of the (assumed rigid) diatomic molecule when the electronic wave function is oriented with orbitals in a lower or higher moment of inertia state. We show a sketch of this in the upper part of Fig. 15.9. Since the energy is directly proportional to the total angular momentum quantum number and inversely proportional to the moment of inertia, the molecule shown on the left has higher energy than the one shown on the right.

There are also a few molecules for which the orbital angular momentum is zero, but the electron spins are parallel, so that the total spin is unity. These molecules have triplet sigma $^3\Sigma$ ground states. The most important astrophysical example is the SO molecule; another species with a triplet Σ ground state is O_2. These energy levels are characterized by the quantum number, N, and the orbital angular momentum quantum number J. The most probable transitions are those within a ladder, with $\Delta J = \Delta N = \pm 1$, but there can be transitions across N ladders. As with the OH molecule, some states of SO are very sensitive to magnetic fields. One could then use the Zeeman effect to determine the magnetic field strength. This may be difficult

Fig. 15.9 The lower energy
levels of OH showing
Λ-doubling. F is the total
angular momentum, including
electron spin, while J is the
rotational angular momentum
due to the nuclear motion.
The quantum number F
includes hyperfine splitting of
the energy levels. The parities
of the states are also shown
under the symbol F. The Λ
doubling causes a splitting of
the J states. In the sketch of
the OH molecule, the shaded
regions represent the electron
orbits in the Λ state. The two
unshaded spheres represent
the O and H nuclei. The
configuration shown on the
left has the higher energy OH
diagram [taken from Barrett
(1964)]. The sketch of the
energy levels is adapted from
Wilson et al. (1990)
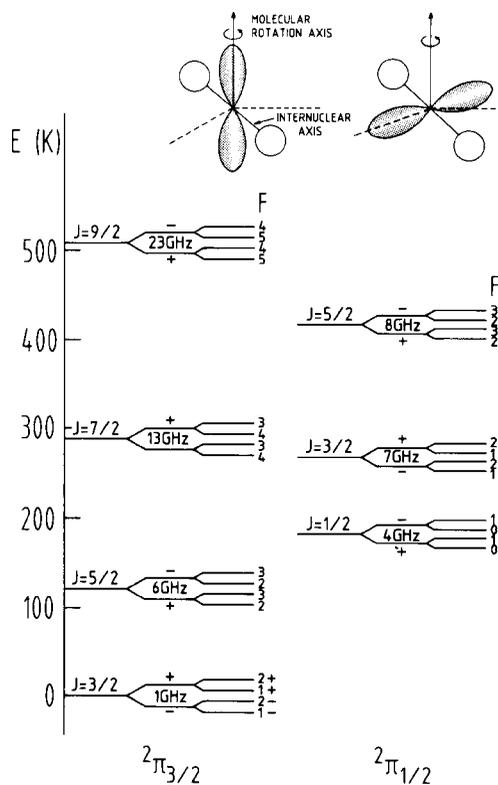
since a 1 μGauss field will cause a line splitting of only about 1 Hz in linear polar-
ization. Even so, measurements of the polarization of the $J_N = 1_0 - 0_1$ line of SO
(see Tiemann (1974) for structural details) may allow additional determinations of
interstellar magnetic fields.

15.6.5   Molecules with Hindered Motions

The most important hindered motion involve quantum mechanical tunneling; such
motions cannot occur in classical mechanics because of energy considerations. A
prime example of this phenomenon is the motion of the hydrogen atom attached to
oxygen in CH_3OH, methanol. This H atom can move between 3 positions between
the three H atoms in the CH_3 group. Another example is motion of the CH_3 group
in CH_3COOH, methyl formate (see Plummer et al. 1987). These are dependent on
the energy. At low energy these motions do not occur, while at larger energies are
more important. For both methanol and methyl formate these motions allow a large
number of transitions in the millimeter and sub-millimeter range.

The description of energy levels of methyl formate follows the standard nomenclature. For methanol, however, this is not the case, due to historical developments. Lees (1973) has presented a description of the energy levels of E type methanol. The energy levels are labelled as J_k, where K can take on both positive and negative values. Figure 15.10 is taken from that paper. There is a similar scheme for naming energy levels of A type methanol, as $A_k^{\pm}$. A and E type methanol are analogous to ortho and para species, in that these states are not normally coupled by collisions.

Torsionally excited states of methanol have been found in the interstellar medium. These are analogous to transitions from vibrationally excited states to the ground state of a molecule. Another complexity is that because of its structure, there are two dipole moments, along either the c or the a axis (see the plot in the upper right of Fig. 15.10.
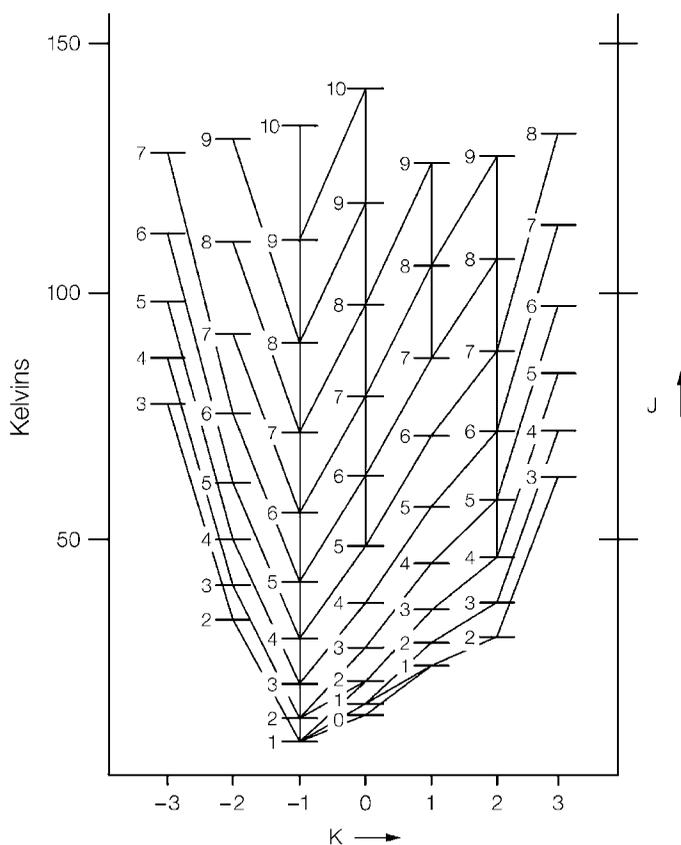
Fig. 15.10 Energy-level diagram of E type methanol, CH_3OH. This is an asymmetric top. The energy-level structure is typical of an almost prolate symmetric top molecule. The lines connecting the levels show the spontaneous transitions with the largest A coefficients from each level; where the two largest A coefficients are within a factor of 2, both transitions are shown. [adapted from Lees (1973)]. See Anderson et al. (1992) for details and further references

Problems

1. (a) For $T = 273$ K and pressure 1 atmosphere, that is 10^6 dyne cm^{-2} (760 mm of Hg), find the density, n, of an ideal gas in cm^{-3}. Repeat for conditions in a molecular cloud, that is $T = 10$ K, pressure 10^{-12} mm of Hg.
(b) For both sets of conditions, find the mean free path, λ, which is defined as $1/(\sigma n)$, and the mean time between collisions, τ, which is $1/(\sigma n v)$, where v is the average velocity. In both cases, take $\sigma = 10^{-16}$ cm^{-3}. For the laboratory, take the average velocity to be 300 m s^{-1}; for the molecular cloud, take the average velocity of H_2 as 0.2 km s^{-1}.
(c) Suppose that the population of the upper level of a molecule decays in 10^5 s. How many collisions in both cases occur before a decay?
(d) For extinction we define the penetration depth, λ_v, in analogy with the mean free path. When $\lambda_v = 1$ the light from a background star is reduced by a factor 0.3678. For a density of atoms n, λ_v in cm is $2 \times 10^{21}/n$. Calculate the value of λ_v for a molecular cloud and for standard laboratory conditions. The parameters for both are given in part **(a)** of this problem.

2. (a) The result of Problem 2, Chap. 13 is $T_k = 21.2 \, (m/m_H) \, (\Delta V_t)^2$ where ΔV_t is the FWHP thermal width, i.e. there is no turbulence and the gas has a Maxwell–Boltzmann distribution. Apply this formula to the CO molecule (mass 28 m_H) for a gas of temperature T. What is ΔV_t for $T = 10$ K, $T = 100$ K, $T = 200$ K?
(b) The observed linewidth is 3 km s^{-1} in a dark cloud for which $T = 10$ K. What is the turbulent velocity width in such a cloud if the relation between the observed FWHP linewidth, $\Delta V_{1/2}$, the thermal linewidth, ΔV_t and the turbulent linewidth ΔV_{turb} is

$$\Delta V_{1/2}^2 = \Delta V_t^2 + \Delta V_{turb}^2 \, ?$$

3. The following expression is apropriate for a linear molecule: $A_{ul} = 1.165 \times 10^{-11} \mu_0^2 v^3 \, (J+1)/(2J+3)$ where v is in GHz, μ_0 is in Debyes and J is the lower level in the transition from $J+1 \rightarrow J$. Use this to estimate the Einstein A coefficient for a system with a dipole moment of 0.1 Debye for a transition from the $J = 1$ level to the $J = 0$ level at 115.271 GHz.

4. To determine whether a given level is populated, one frequently makes use of the concept of the "critical density", n^*

$$A_{ul} = n^* \, \langle \sigma v \rangle \, .$$

Here u is the quantum number of the upper level, and l is that for the lower level. If we take $\langle \sigma v \rangle$ to be 10^{-10} cm^3 s^{-1}, determine n^* from the following A_{ul} coefficients

$$CS : A_{10} = 1.8 \times 10^{-6} \, s^{-1}$$
$$CS : A_{21} = 2.2 \times 10^{-5} \, s^{-1}$$
$$CO : A_{10} = 7.4 \times 10^{-8} \, s^{-1}.$$

5. (a) Suppose the effective radius $r_e = 1.1 \times 10^{-8}$ cm and the reduced mass, m_r, of a perfectly rigid molecule is 10 atomic mass units, AMU (an AMU is 1/16 of the mass of a 16-oxygen atom; 1 AMU $= 1.660 \times 10^{-24}$ g), where $\Theta = m_r r_e^2$.
(a) Calculate the lowest four rotational frequencies and energies of the levels above the ground state.
(b) Repeat if the reduced mass is (2/3) AMU with a separation of 0.75×10^{-8} cm; this is appropriate for the HD molecule. The HD molecule has a dipole moment $\mu_0 = 10^{-4}$ Debye, caused by the fact that the center of mass is not coincident with the center of charge. Take the expression for $A(ij)$ from Problem 3 and apply to the $J = 1 - 0$ and $J = 2 - 1$ transitions.
(c) Find the "critical density" $n^* \approx 10^{10} A(ij)$.

6. The $^{12}C^{16}O$ molecule has $B_e = 57.6360$ GHz and $D_e = 0.185$ MHz. Calculate the energies for the $J = 1, 2, 3, 4, 5$ levels and line frequencies for the $J = 1 - 0$, $2 - 1, 3 - 2, 4 - 3$ and $5 - 4$ transitions. Use the expression energy $E(J)/h \approx B_e J$ $(J + 1) - D_e J^2 (J + 1)^2$ for the energy calculation. Check the results against Table 14.2.

7. Apply for $J = 0, 1$ the analysis in Problem 10 to the linear molecule $HC_{11}N$, which has $B_e = 169.06295$ MHz and $D_e = 0.24$ Hz. Estimate J for a transition near 20 GHz. What is the error if one neglects the distortion term?

8. In the following, we neglect the distortion term D_e and assume that the population is in LTE. The population in a given J level for a linear molecule is given by (Eq. 15.32)
$$n(J)/n(\text{total}) = (2J + 1)e^{B_0 J(J+1)/kT}/Z$$
where Z, the partition function, does not depend on J. Differentiate $n(J)$ with respect to J to find the state which has the largest population for a fixed value of temperature, T. Calculate this for CO if $T = 10$ K and $T = 100$ K. Repeat for CS ($B_0 = 24.584$ GHz) and $HC_{11}N$, for $T = 10$ K.

9. Extend (Eq. 15.32) to include the optical depth relation (Eq. 15.26) to obtain an estimate of which J level has the largest optical depth, τ, in the case of emission for a linear molecule.
(a) Show that when the expression for the A coefficient for a linear molecule is inserted into (Eq. 15.26), we have $N_l = \frac{1.67 \times 10^{14}}{\mu_0^2 v[\text{GHz}]} \frac{2J+1}{J+1} T_{ex} \tau \Delta v$, where μ is in Debyes and v is in km s^{-1}.
(b) Use the above expression to estimate whether the J for the maximum $T_{MB} = T_{ex} \tau$ is larger or smaller than the J obtained in Problem 12.

10. Find the ratio of the intensities of the $J = 2 - 1$ to $J = 1 - 0$ transitions for a linear molecule if the excitation temperature of the system, T, is very large compared to

the energy of the $J = 2$ level above the ground state, and both lines are optically thin. What is the ratio if both are optically thick?

11. The ammonia molecule, NH_3, is an oblate symmetric top. For ammonia, $A = 298\,GHz$, $C = 189\,GHz$. If $T \gg A, C$, the value of Z, the partition function, with C and B in GHz, is $Z = 168.7\sqrt{(T^3)/(B^2 A)}$.
(a) Evaluate Z for NH_3 for $T = 50\,K$, $100\,K$, $200\,K$, $300\,K$. For this approximation to be valid, what is a lower limit to the value of T?
(b) The (3,3) levels are $120\,K$ above ground. Use the partition function and

$$n(J)/n(\text{total}) = (2J+1)e^{120/T}/Z$$

to calculate the ratio of the total population to that in the (3,3) levels.
(c) If only metastable ($J = K$) levels are populated, use the definition of Z as a sum over all populated states, and

$$n(J)/n(\text{total}) = (2J+1)e^{(BJ(J+1)+K^2(C-B))/kT}/Z$$

and the A and C values for NH_3 to obtain the ratio between the population of the (3,3) levels and all metastable levels.

12. The selection rules for dipole transitions of the doubly deuterated isotopomer D_2CO differ from that of H_2CO since D_2CO has two Bosons, so the symmetry of the total wavefunction must be symmetric. Determine these rules following the procedure in Sect. 15.6.2.

Chapter 16
Molecules in Interstellar Space

We begin with a history of the field of molecular line astronomy starting from the simpler situations typical of the ISM, including radiative processes. In Chap. 15, we analyzed the excitation of carbon monoxide, CO, under local thermodynamic equilibrium (LTE). This is a simple situation but has limited validity; more common are deviations from LTE. In this Chapter, we give examples of two level and three level systems, the latter being used for an analysis of one-dimensional maser amplification. We then derive the large velocity gradient (LVG) approximation, which is one method of analyzing the transport of moderately optically thick lines, and an account of Photon Dominated Regions (PDRs). A brief exposition of the use of molecules as probes of the interstellar medium follows.000 These tools are those used to investigate molecular line sources in our galaxy as well as other galaxies. We close the chapter with an introduction to interstellar chemistry.

Since the literature in this field is so immense, we quote only the most recent references. In these, one can find citations to earlier publications. One general recent publication is "Protostars and Planets V" (2007) edited by Reipurth, Jewitt and Keil, hereafter "PPV"; previous editions of Protostars and Planets by Mannings et al. (2000) and Levy and Lunine (1993) are also given in the references. General texts related to the results in this Chapter are by Lequeux (2005), Stahler and Palla (2005) and Sparke and Gallagher (2000). An interactive data base for molecular lines is described in Remijan, A. J. et al. (2007).

16.1 Introduction

As we have seen in Chaps. 13 and 14, only a few atomic species have been detected at radio wavelengths, the study of molecular line radiation provides a vastly richer field of study. It is well established that stars form in molecular clouds and therefore a determination of the physical conditions in these clouds will help us to understand the star formation process. Also, newly formed stars greatly influences their birthplaces. Molecular line studies are essential to understanding how stars such as our sun and our solar system formed. From molecular line studies we can determine

conditions on the surface of planets and in the interiors of comets. From these results we may be able to estimate conditions in the primitive solar nebula.

Molecular clouds consist mostly of H_2. The production of H_2 must occur on dust grains whose surfaces act as catalysts for the conversion of hydrogen into H_2. Efficient production requires a minimum density, which is generally thought to be $\geq 50\,cm^{-3}$. Since the H_2 molecule is dissociated by spectral line photons with energies $>11\,eV$, H_2 must also be shielded against the interstellar radiation field (ISRF). Some shielding is supplied by dust grains, but H_2 in the outer layers of the clouds provides more protection from the ISRF since H_2 is destroyed by spectral line radiation. These constituents prevent dissociating radiation from penetrating deeply into the cloud. The minimum H_2 column density is $\approx 10^{20}\,cm^{-2}$, equivalent to a visual extinction, $A_v \sim 0.1^m$. By mass, clouds consist of $\sim 63\%\ H_2$, $\sim 36\%$ He and $\sim 1\%$ dust, other molecules and atoms.

The study of molecules is far more complex than the study of neutral hydrogen (Chap. 13) for three reasons:

(1) deviations of the populations of energy levels from LTE. Then the populations are characterized by *excitation temperatures*, T_{ex}. Sometimes the population of a given species must be specified by a number of different values of T_{ex};
(2) non-equilibrium interstellar chemistry, including depletion from the gas phase onto dust grain surfaces and the formation of species in the gas phase and on grain surfaces;
(3) the presence of small scale structure, or *clumping* that leads to large inhomogeneities; these lead to differences between beam-averaged and source-averaged abundances.

16.1.1 History

Molecular lines in the optical range were detected in late type stars with low surface temperatures, planetary atmospheres, and comets in the 1930s. About this time, interstellar absorption lines of CN were found in a dust cloud toward the star ζ Oph; later, lines of CH^+ and CH were identified. These results showed that diatomic molecules can exist in the interstellar medium, given the proper physical conditions.

Molecular line radio astronomy began in 1963 when two OH lines were detected by the absorption (see Fig. 15.9) of continuum radiation from the supernova remnant Cassiopeia A. The clouds containing OH are not associated with the intense radio source Cas A, but are line-of-sight objects. There are four ground state OH transitions; the two with the largest line strengths at 1.665 and 1.667 GHz were detected in absorption. It soon became clear that the excitation of OH deviates from LTE, since toward Cassiopeia A, the highest frequency line from the OH ground state at 1.720 GHz, appeared in emission, while the other three OH ground state lines appeared in absorption. This *cannot* occur under LTE conditions. Such line emission is an indication that the upper level is overpopulated and amplifies the continuum background, so this is a natural maser. For clouds near H II regions these deviations

are much more spectacular. The peak OH line intensities as observed with single dishes are even stronger than H I lines and show narrow spectral features. These results were first attributed to an as yet unknown molecule "Mysterium". Further measurements showed that the emission is polarized and time variable. Data taken with interferometers showed that the OH emission arises from very compact sources, with sizes of milli arc seconds. Thus the true brightness temperatures are $> 10^{12}$ K. At this temperature the OH would be dissociated, so these brightness temperatures cannot represent the kinetic temperature, T_K. Thus it was concluded that the peak temperatures are not related to kinetic temperatures, but caused by non-LTE processes. Ultimately, it was shown that stimulated emission or Maser line emission is the cause of this effect. The excitation of OH masers is now thought to be well understood (see Sect. 16.2.2).

Until 1968 all interstellar molecules detected consisted of only two atoms. This was believed to be a natural limit caused by low densities in the ISM. Then, however, the line radiation of ammonia, NH_3 (see Fig. 15.5), was found at 1.3 cm. Later a centimeter wavelength line of water vapor, H_2O (see Fig. 15.6), was found in the same frequency band by the same group. Toward some of the sources, the $\lambda = 1.35$ cm water vapor line showed intense radiation consisting of features with narrow linewidths. It was soon found that this emission is time variable. Later, a series of radio interferometer measurements showed that the true brightness temperatures are $> 10^{15}$ K, so it was concluded that the centimeter wavelength emission of H_2O is caused by strong maser action. At present, the excitation of H_2O masers is not completely understood; determinations of the abundance of H_2O using non-maser lines is a subject of current research.

In 1969, the 6 cm K-doublet line of H_2CO, formaldehyde, was discovered (see Fig. 15.7). In some regions, H_2CO is seen in absorption against the 2.7 K microwave background. The 2.7 K background pervades all of space. Then in LTE all kinetic temperatures must be equal to or larger than 2.7 K, i.e. $2.7 K < T_{ex} < T_K$ must hold, so the absorption of the 2.7 K background must be caused by non-LTE effects. In terms of (11.13), the population of the lower level is increased, so that the absorption is enhanced. This enhanced absorption is the opposite of maser action.

After this initial discovery period, perhaps the most important molecule found was carbon monoxide (see Fig. 15.3), in mid-1970. Early empirical studies indicated that the CO to H_2 ratio appeared to be $\approx 10^{-4}$ in dense molecular clouds. In recent years there is an indication that in small (< 0.1 pc) cold ($T_K \approx 10 K$) regions, CO may freeze out onto grains. At about this time, it was recognized that such complex molecules must be in dense clouds of molecular hydrogen, H_2, since the excitation is in part through collisions and because these species are easily dissociated so these must be protected from the ISRF. At the kinetic temperatures of such clouds, H_2 lines are not excited, so the properties of molecular clouds must be traced by molecules with permanent dipole moments. In 1976, vibrational lines of H_2 were found in shocked regions and later, rotational lines of H_2 were found in warm molecular clouds (see the account in Townes 1994).

The molecule which showed the importance of ion-molecule chemistry was termed "X-ogen" in 1970. This was later shown to be HCO^+ (see Herbst 1999,

2001). The number of detected molecular species increased greatly during this time. By 1970, six molecules had been found, by 1980 there were 51, by 1990 there were 85, by 2004 there were 130, and today there are 151. In the last few years, the rate of detection of new molecules is a few per year. Most of these species were found in the radio range, although in the last years a few species without permanent dipole moments have been found from infrared (FIR) vibrational transitions. In the mid-1970s, there were a number of studies of isotopic ratios of carbon, nitrogen and oxygen. These continued with varying amounts of effort into the 1990s.

While large scale surveys of atomic hydrogen date back to the 1950s (see Chap. 13) imaging of the galaxy in the $J = 1 - 0$ line of CO began in the early 1970s. The first measurement of CO in another galaxy was in 1975. This work continues to the present. From these data, estimates of cloud masses, correlations of cloud parameters, the distribution of H_2 (compared to that of HI) and histograms of clump masses have been made. Our galactic center has been imaged repeatedly in CO; the distribution is asymmetric with most of the emission north of the center. This is associated with the source Sgr B2.

The study of specific sources began with the discovery of an intense source of molecular lines in Orion and this has been expanded to include a number of sources. Toward the end of the 1970s, it was found that very young stars pass through a period where material is ejected in a bipolar outflow. Apparently this is the case for both high and low mass stars (see PPV). The relation between molecular clouds and star formation is a close one, since molecular clouds are cold and dense, so collapse is likely. For a kinetic temperature of 10 K, the time in years for the collision between a gas particle and a grain is

$$t_{\text{gas}-\text{grain}} = \frac{1.2 \times 10^{10}}{n_{H_2}} \tag{16.1}$$

However, this does not seem to be the case for all molecules, since NH_3, H_2D^+ and N_2H^+ are found in dense quiescent cores. This empirical result has generated interest in these species. With high resolution measurements of these species one can image the dynamics of dense cores that may be collapsing.

It is well established that when low mass stars end their Main Sequence lives, moving to the Asymptotic Giant Branch (AGB), these eject processed material, and enrich the interstellar medium. Molecular line studies of this phase have concentrated on the source IRC+10216, a nearby carbon star but recently has been extended to other AGB stars and Planetary Nebulae. From molecular line studies, we know of a number of such post-main sequence sources in different evolutionary phases.

Beginning in the mid-1970s, CO was found in nearby galaxies. This has opened the vast field of extragalactic molecular line astronomy. This study has been extended to a wide range of molecules, as well as fine structure lines (see Table 13.1) of neutral carbon (C I) and C^+ ([C II]), oxygen [OI] and nitrogen [N II]. The correlation between the intensities of [C II] and CO have led to the concept of a

Photon Dominated Regions, or PDR, in which the radiation field is $\sim 10^2$ to 10^5 times that near the Sun, but in which simple molecules such as CO can survive and produce intense emission. Such regions are prominent in galaxies undergoing bursts of star formation (see Ivison et al. (2002)). Searches for CO and dust emission was extended to high redshift objects, reaching a record of $z = 6.42$ (see e.g. Maiolino et al. 2007; Solomon and vanden Bout 2005). Objects with the most intense emission may be gravitational lensed, but are at least abnormally intense. Most are dust enshrouded star forming regions which are best studied in the radio or infrared.

For solar system studies, radio astronomical molecular line measurements of the planets and comets started in the 1970s. The most remarkable results for comets were obtained when Hale-Bopp and Hyakutake passed close to the Sun in the mid-1990s since these had a large amount of gas-phase molecules. Studies of disks around low mass stars showed the presence of dust and gas. The H_2 densities are very large by ISM standards, and geometries are complex, so that high angular resolutions are needed to accurately determine abundances as a function of position from the star.

In the following sections, we consider solutions to the following problem: given a line intensity, one wishes to separate excitation, radiative transfer and abundance effects. In the last chapter, we presented some particularly simple situations for CO, and NH_3. LTE formulae were also given. However, LTE is the exception in the ISM. Thus, more specific models must be applied to interpretet the data. We present an overview of excitation in the next section. Following this, we present a more complex model of radiative transport, and then present a discussion of molecules as probes of the ISM. Following this a brief overview of results.

16.2   Molecular Excitation

16.2.1   *Excitation of a Two-Level System*

So far, we have only considered molecules populated under LTE conditions. We will now start a much more general analysis. In this section, we will consider a two-level approximation to molecular energy level populations and we will apply this formalism to two- and three-level models to investigate masers. Then we will generalize the photon transport and apply this to quasi-thermally excited molecules, and use these results to probe physical conditions in molecular clouds.

The methods and expressions used in this study have practically all been covered in Chap. 12. Although molecules are vastly more complex, a first discussion of molecular excitation will be made in analogy with the analysis used for H I. The following discussion will show that the observation of emission from a single line is, contrary to common belief, not sufficient to establish a minimum gas density. For such effects, at least three energy levels must be involved.

The emissivity using the two-level approximation has been derived in Chap. 12 resulting in the expression (12.15):

$$\varepsilon_v = \frac{h\,v_0}{4\pi}\,n_u A_{ul}\,\varphi(v)\,, \tag{16.2}$$

where n_u is the population of the upper level and $\varphi(v)$ the line shape [cf. (12.1)]. This emissivity is proportional to A_{ul}, and this seems to be the reason for the above-mentioned belief. A line with an exceedingly small A_{ul} would be very weak, unless n_u is very large.

But ε_v is proportional to the population n_u of the excited level, and in the limit of small A we will have to consider collisional excitation even in a low density situation. The rate equation (12.30) then results in a stationary population as given by (12.35); that is

$$n_l\,(C_{lu} + B_{lu}\bar{U}) = n_u\,(A_{ul} + B_{ul}\bar{U} + C_{ul})\,. \tag{16.3}$$

Note that $\bar{U}$ is the average radiation field intensity in (12.32). To solve for the emissivity, we substitute (16.3) into (16.2), obtaining:

$$\varepsilon_v = \frac{h\,v_0}{4\pi}\,\frac{n_l(C_{lu} + B_{lu}\bar{U})}{A_{ul} + B_{ul}\bar{U} + C_{ul}}\,\varphi(v)\,. \tag{16.4}$$

If collisions dominate, (16.3) shows C_{lu} and C_{ul}, the collision rates, are connected by the principle of detailed balance (12.37):

$$\frac{C_{lu}}{C_{21}} = \frac{g_u}{g_l}\,\exp^{-h\,v_0/k\,T_k} \tag{16.5}$$

and the A and B coefficients are connected by the Einstein relations given in Chap. 11, (12.8) and (12.9). Using these, and substituting for brevity

$$\tilde{S}_v = \frac{2\,h\,v_0^3}{c^2}\,, \tag{16.6}$$

(16.4) then becomes (Fig. 16.1)

$$\varepsilon_v = \frac{h\,v_0}{4\pi}\,\frac{g_u\,n_l}{g_l}\,\frac{\bar{U}/\tilde{S}_v + \dfrac{C_{ul}}{A_{ul}}\exp(-h\,v_0/k\,T_K)}{1 + \bar{U}/\tilde{S}_v + C_{ul}/A_{ul}}\,\varphi(v)\,. \tag{16.7}$$

As expected, in the limit of $C_{ul}/A_{ul} \gg 1$, the population of the two-level system is determined by the kinetic temperature, T_K. However, even in the limit of small C_{ul}/A_{ul}, that is, a very sub thermally excited state, there will be *some* emission. For a detectable line however, the abundance must be sufficiently large. From (16.7)

Fig. 16.1 Emissivity and
excitation temperature as a
function of C/A in the case of
weak lines
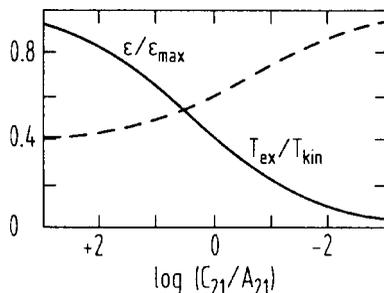
the emissivity also depends on column density. Because CO has a small A coeffi-
cient, the levels are populated at densities of a few $100\ \mathrm{cm}^{-3}$. Emission from the
$J = 1 \rightarrow 0$ or $J = 2 \rightarrow 1$ lines of CO is widespread because of the large column
density.

If a level is populated by collisions with H_2, the main collision partner, a first
approximation to the H_2 density can be had if one sets the collision rate, $n(H_2)\langle \sigma v\rangle$
equal to the A coefficient. The brackets indicate an average over velocities of H_2,
which are assumed to be Boltzmann distributed. The H_2 density for a given transi-
tion which will bring T_{ex} midway between the radiation temperature and T_K is re-
ferred to as the *critical density*, and is denoted by n^*. That is $n^*\langle \sigma v\rangle = A$. As noted
at the beginning of this section, this can be at best only an approximate estimate.

In some cases, the collision cross section, σ, is known from calculations or from
experiment. As a first approximation one can assume that $\sigma = 10^{-16}\ \mathrm{cm}^{-2}$. If the
velocity, v, is taken to be $1\ \mathrm{km\,s}^{-1}$, the critical density is $n_{H_2} = A \times 10^{10}\ \mathrm{cm}^{-2}$.
Taking the Einstein A coeffients from Table 14.2, we find that to produce significant
emission from the $J = 1 \rightarrow 0$ line of CO, we need densities of $\sim 740\ \mathrm{cm}^{-3}$. To
produce emission from the $J = 1 \rightarrow 0$ line of CS, we require densities of $1.8 \times
10^4\ \mathrm{cm}^{-3}$. Thus, CO emission can arise from both lower and higher density regions,
while CS emission can arise only from higher density regions.

For an optically thin line, every collision results in a photon being emitted. If
the collision rate is less than the A coefficient, the line is said to be sub thermally
excited, that is $2.7\,\mathrm{K} < T_{\mathrm{ex}} < T_K$. Then, the line emitted will be weaker, but may
be detectable. This two-level model is a first approximation for the case of the
$J = 1 \rightarrow 0$ line of molecule HCO^+. For this line, $n^* = 3 \times 10^5\ \mathrm{cm}^{-3}$. Then, in clouds
where densities are of order $10^3\ \mathrm{cm}^{-3}$, this line is sub thermally excited. The line
can be detected only because the optical depth is large; estimates of HCO^+ opti-
cal depths are obtained from comparisons with isotopic lines, following methods
similar to those used for CO and ^{13}CO.

From Fig. 16.1 there is a smooth variation in the emissivity between the limits
where the radiation field dominates and where collisions dominate. When calculat-
ing the emissivity, one must account for the radiation field itself.

Before considering photon transport in optically thick lines, we will briefly con-
sider three-level systems. As is clear from Fig. 16.1, a two-level system *cannot* give
rise to maser emission, or anomalous absorption $T_{\mathrm{ex}} < 2.7\,K$.

16.2.2   Maser Emission Processes in One Dimension

In radio astronomy, the concept of maser emission can be used to explain many of the observed molecular lines. The maser phenomenon is very natural in interstellar space, since there are usually deviations from LTE. In fact, it is very unusual to find molecular excitation which is close to LTE. Our treatment is a simplified version of that given in Chap. 4 of Elitzur (1992).

Masers arise when the population of the upper energy level, n_u is more than the factor $(g_u/g_l)n_l$. This is formally represented by a negative T_{ex} in the expression

$$\frac{n_u}{n_l} = \frac{g_u}{g_l} \exp^{-h\nu_0/kT_{ex}} . \tag{16.8}$$

For this situation, the transition connecting n_u with n_l can amplify a background source of radiation, if one is present. As we will show later in this section, if there is no background continuum and if the line optical depth is small, no masering line will be produced.

Shortly after Weinreb et al. (1963) detected the OH line in absorption, OH lines were found with intensity ratios that could not be explained at all by line radiation emitted under LTE conditions. This conclusion was strengthened when OH line emission spectra in some sources were found to be linearly and circularly polarized, and that the sources of the emission lines had exceedingly small angular diameters leading to very high brightness temperatures for the sources. Another intense maser line is the 6_{16}–5_{23} rotational transition of H_2O. After this discovery it was generally accepted that deviations from LTE were a general feature of molecular excitation, and that OH and H_2O emission is merely an extreme case. Since then many more masering transitions have been detected for many different molecules. Deviations from LTE could also enhance absorption. For example, in a large number of galactic sources, the K doublet transitions of H_2CO even absorb the 2.7 K background, that is, absorption is found even at positions where no discrete background source is present. Such phenomena have been found for some transitions of CH_3OH. These transitions act as "cosmic refrigerators". However, an overpopulation of the lower level for a given transition can only lead to a limited degree of absorption.

Masers are classed as either *strong* or *weak*. Strong masers produce intense spectral line radiation, such as the 18 cm lines of OH, or the 1.3 cm line of H_2O. The large brightness temperatures produced by maser emission allow the use of high-resolution interferometry, which allows us to investigate phenomena on a very small scale. The excitation processes which lead to strong masers are very difficult to understand, since the processes are very nonlinear.

In the following, we will show a highly simplified version of the one-dimensional galactic maser process. We emphasize that this is an example to show the basic physical principles. At the end of this section, we will describe specific astrophysical maser models, explaining some details of the maser excitation. For a more detailed discussion of maser models, see Watson (1994).

Let us consider a cloud of molecules with three energy levels l, u and 3. For the sake of simplicity the statistical weights of the levels are taken to be the same. The radiation transfer equation (12.13) corresponding to the transition $l\overrightarrow{\leftarrow}u$ with the frequency v_0, is then

$$\frac{dI_v}{ds} = \frac{h v_0}{4\pi} \left[(n_u - n_l) B \frac{4\pi}{c} I_v + n_u A \right] \varphi(v),\tag{16.9}$$

where we write A for A_{ul}, and B for B_{lu} and B_{ul}, the last two being equal now because we assumed $g_l = g_u$. The populations n_l and n_u of the levels can be changed by spontaneous emission, described by the transition rate A; by stimulated emission M given by

$$M = \frac{4\pi}{c} B \bar{I} = \frac{4\pi}{c} B I \frac{\Omega_m}{4\pi}\tag{16.10}$$

where Ω_m is the beam solid angle of the radiation; or by collisions described by the rate C; and finally by the pumping rates from levels l and u to a third level, P_{lu} and P_{ul}. Such a transfer to a third energy level is essential to produce population inversion. If the system is stationary then, as in (12.35),

$$n_l (C_{lu} + M_{lu} + P_{lu}) = n_u (C_{ul} + M_{ul} + P_{ul} + A_{ul}).\tag{16.11}$$

In the study of microwave masers we can assume that A is negligible compared to C and M, and that $C_{lu} \approx C_{ul} \approx C$ and $M_{lu} \approx M_{ul} \approx M$. Then we can write:

$$\frac{n_u}{n_l} = \frac{P_{lu} + M + C}{P_{ul} + M + C}.\tag{16.12}$$

The population inversion which the pump would establish if $M = C = 0$, is then

$$\Delta n_0 = (n_u - n_l)\big|_{M=C=0} = n \frac{P_{lu} - P_{ul}}{P_{lu} + P_{ul}}\tag{16.13}$$

with

$$n = n_l + n_u.\tag{16.14}$$

For $C \neq 0$ and $C \neq M$ we have from (16.12) with $P = P_{lu} + P_{ul}$:

$$\Delta n = \frac{\Delta n_0}{1 + \dfrac{2(C+M)}{P}}.\tag{16.15}$$

Substituting this into (16.9) results in

$$\frac{dI_v}{ds} = \frac{\alpha I_v}{1 + I_v/I_s} + \varepsilon\tag{16.16}$$

where

$$\alpha = \frac{h \, v_0}{c} B \frac{\Delta n_0}{1 + \dfrac{2C}{P}} \varphi(v), \tag{16.17}$$

$$I_s = \frac{cP}{2B\Omega_m} \left(1 + \frac{2C}{P}\right), \tag{16.18}$$

and

$$\varepsilon = \frac{h \, v_0}{4\pi} n_u A \, \varphi(v). \tag{16.19}$$

In most astrophysical applications the term ε is taken to be a constant and often this can be dropped altogether.

For $I_v \ll I_s$ we have the solution for the *unsaturated maser* for $v = v_0$

$$I_{v0} = I_0 \, e^{\alpha_0 l} + \frac{\varepsilon}{\alpha_0} \left(e^{\alpha_0 l} - 1\right) \tag{16.20}$$

where l is the length along the line of sight within the maser region. Converting this relation to temperature, we have

$$T_b = T_c \, e^{\alpha_0 l} + |T_{ex}| \left(e^{\alpha_0 l} - 1\right). \tag{16.21}$$

This is equal to the isothermal solution (1.37) with $\tau = -\alpha_0 l$.

For $I_v \gg I_s$ the right-hand side of (16.16) is a constant, and so the solution for the *saturated maser* becomes

$$I_{v0} = I_0 + (\alpha_0 I_s + \varepsilon) l. \tag{16.22}$$

In this case the intensity increases linearly with l compared to the exponential growth with l in the unsaturated maser. If a maser is unsaturated the width of the line will steadily decrease with increasing maser gain (Fig. 16.2). A wide band background source with a brightness temperature T_c will, according to (16.21), produce the output signal

$$T_b(v) = T_c \, e^{\alpha(v) l}. \tag{16.23}$$

Assuming a gaussian line shape φ in (16.17), $\alpha(v)$ is given by

$$\alpha(v) = \alpha_0 \exp\left(-(v - v_0)^2 / 2 \sigma_0^2\right).$$

As long as we concentrate our attention on the center of the line this can be approximated by

$$\alpha(v) = \alpha_0 \left[1 - \frac{1}{2} \left(\frac{v - v_0}{\sigma_0}\right)^2\right],$$

so that (16.23) then becomes
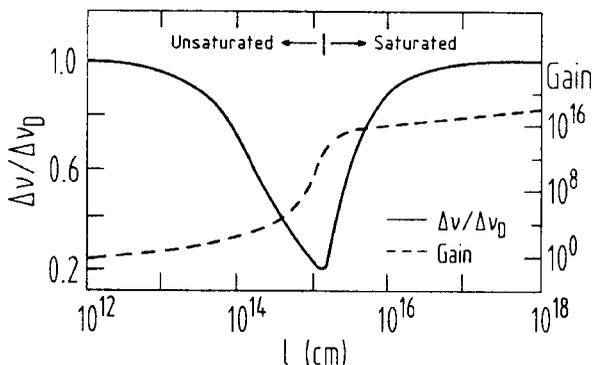
Fig. 16.2 A plot of the gain and linewidth as a function of distance for a linear maser

$$T_b(\nu) = T_c \, e^{\displaystyle \alpha_0 \exp\left(-\frac{\alpha_0 \, l}{2\sigma_0^2}(\nu - \nu_0)^2\right)}.$$

But this is a gaussian with the dispersion

$$\sigma(\alpha_0 \, l) = \frac{\sigma_0}{\sqrt{\alpha_0 \, l}} \tag{16.24}$$

where $\sigma_0 = 3/8\ln 2\,\Delta\nu_{1/2}$. From (16.23), the width will decrease until the center of the line begins to saturate. The linewidth then broadens as the wings of the line continue to experience exponential growth until the line again has the original line width σ_0. In a saturated maser the linewidth remains constant. The saturated and unsaturated masers are extreme cases. Intense masers are usually a mixture of these two categories: unsaturated when the signal intensity is small, but then saturated when the intensity becomes large.

What we have given here is only a much simplified phenomenological description of the maser effect in one dimension. If the astrophysics of any given maser source is to be fully understood this is only the very first step. A detailed specification of the pump mechanism, the detailed solution of the rate equations and an account of the energy sources are needed just as a geometrical model. In the following we give a few examples of interstellar masers, emphasizing the qualitative aspects.

All masers can be understood in terms of thermodynamics, in the sense that these are like Carnot engines. In the case of interstellar masers this consists of producing useful work, or more generally, organization, in the form of directional radiation, which is sometimes highly polarized, at the expense of increasing entropy, or disorganization, in the place where the radiation is produced. In terms of Carnot engines, there must be a heat source and a heat sink. That is, there must always be two influences operating on the molecular species. Usually these are radiation and collisions, because in the ISM, in general, the radiation color temperature is very different from the kinetic temperature. However, two radiation sources, for example

in different wavelength ranges, or two collision sources, for example electrons and neutrals *with different temperatures* are also possible.

16.2.3 Non-LTE Excitation of Molecules

In this section, we present a few observational examples to which the concepts in the last two sections can be applied. The discussion will begin with a consideration of LTE excitation, and then move onto larger and larger deviations from LTE, especially masers (A more detailed discussion of interstellar masers is given in Elitzur (1992), Chaps. 9, 10, 11, 12, 13 and 14).

First, as noted, the CO molecule is exceptional in that the excitation *always* seems to be close to LTE. However, even for CO, there is a level J for which the value of the Einstein A coefficient is large enough that the population is subthermal. For molecules such as CS, with larger dipole moments, subthermal excitation occurs even in lower J levels. For symmetric top molecules such as NH_3, the inversion doublet transitions require densities of order $10^4 \, cm^{-3}$, but transitions across K ladders can occur only as octopole transitions, that is, these are first and second order forbidden. Thus, the value of T_{ex} within a ladder can be small, but the temperature which characterizes populations in different K ladders will be much higher. In the case of NH_3, this temperature, T_{rot}, is close to T_K. There are many examples of different excitation temperatures, even within the same K ladder. This is found in the case of H_2CO, where a transition between the K doublet levels, for example the 1_{10} and 1_{11} levels at 6 cm, can have a T_{ex} of less than 2.7 K, whereas the rotational transitions, for example between the 2_{12} and 1_{11} levels, have T_{ex} larger than 2.7 K. The fact that $T_{ex} < 2.7$ K for the K doublet levels depends on the geometry of the H_2CO molecule and the parity of the levels.

Weak masers are very often found in the ISM. These can be more easily understood than intense or *strong* masers which involve very nonlinear processes. One excellent example is the $J = 1 \rightarrow 0$ emission line of HC_3N. The energy-level scheme is shown in Fig. 15.3. As one example, the galactic continuum source Sgr B2 is behind an extended molecular cloud with $n\,(H_2) \approx 10^3 \, cm^{-3}$. Toward Sgr B2, the $J = 1 \rightarrow 0$ line is found in emission, the $J = 3 \rightarrow 2$ and $J = 4 \rightarrow 3$ lines are found in absorption. The simplest way to explain these observations is the following: The populations of HC_3N are determined by collisions. The collisions follow "hard sphere" rules, meaning that collisional selection rules, ΔJ, can have any value. Then populations can be transferred from $J = 0$ to higher J values. These high J populations decay radiatively following the dipole rule $\Delta J = +1$. For higher J, the decay rate is faster, since the frequency is higher, from the dependence of the Einstein A coefficient. However, for the transition $J = 1 \rightarrow 0$, this rate is slower than the rate for the $J = 2 \rightarrow 1$ transition. Then population builds up in the $J = 1$ level, and there is an overpopulation in the $J = 1$ level relative to $J = 0$ level. These lines are optically thin, but in the centimeter wavelength range there are a number of intense continuum sources. Thus the overpopulation of the $J = 1$ level leads to stimulated

emission, which amplifies the background continuum. In this example, the Carnot heat source is provided by the hot molecular collision partners, and the Carnot heat sink by the cold, centimeter wavelength radiation field.

Strong maser lines are characterized by line profiles which have narrow features, which may vary on short time scales, and are sometimes highly polarized. The explanations offered for *strong* masers must be more vague than for weak masers, since nonlinear effects are central, and small influences may play a major role. In the following, we summarize the currently accepted maser models. For the 18 cm wavelength masering lines from the OH molecule, there are 1.612 GHz maser lines from circumstellar envelopes and a large number of lines from interstellar masers. The circumstellar masers are explained by the absorption of 35 μm far infrared photons, which are copiously produced by the cool, dust enshrouded star. The OH populations are transferred via the 35 μm radiation field from the $^2\Pi_{3/2}, J = 3/2$ states to the $^2\Pi_{1/2}, J = 5/2$ state (see Fig. 15.9). The radiative decay down the $^2\Pi_{1/2}$ ladder occurs at much longer wavelengths, where there are many fewer photons. These decays then occur without much of an interaction with the radiation field, through optically thin transitions. The bottom of the $^2\Pi_{1/2}$ ladder connects radiatively by allowed transitions to the $^2\Pi_{3/2}$ ground state. The net result is a population transfer from the $F = 2$ to the $F = 1$ levels in the OH ground state.

There has been some recent progress in our understanding of the excitation of interstellar OH masers from rotationally excited levels. The ground state populations are transferred to higher rotational states by the absorption of far infrared radiation. Since these masers are located in dusty regions near young stars, this radiation field is very large. This excitation is followed by a radiative decay in the far infrared, back to the ground state. Collisions play an important role in determining populations during this decay, but the most important process leading to these OH masers is the overlap between different far infrared line components.

The H_2O masers are thought to arise in regions where the H_2 densities are $\approx 10^9 \, cm^{-3}$, and T_K values are $\approx 500 \, K$ or more. In the ISM, these conditions are thought to be due to *hydromagnetic* shock waves caused by very young nearby stars. The excitation of the H_2O maser at 1.35 cm can be qualitatively understood (see Fig. 15.6) in that the 6_{16} level is populated by collisions, and that this population has only a small number of allowed radiative decay routes. In contrast, the population of the 5_{23} level can decay via a number of different routes. Thus, a population inversion is almost inevitable. The difficulty is to explain the line intensities, which in some cases reach $10^{15} \, K$. One possibility is that these masers are very elongated cylinders pointing in our direction; another is that the high-brightness masers are amplified in two regions which are aligned in our direction. Boboltz et al. (1998) find a result for a flare in the source W49N that they interpret in terms of the amplification by a foreground cloud of H_2O emission from a background cloud. H_2O masers in the circumstellar envelopes of late type stars are thought to be caused by collisonal excitation to high rotational levels, followed by radiative decays. A number of millimeter and submillimeter wavelength transitions of H_2O masers show maser emission; some are indicated in Fig. 15.6.

The *strong* SiO masers arise from rotational transitions $J+1 \rightarrow J$ in vibrationally excited states. Most of these masers are found in the circumstellar envelopes of late type stars, such as Mira variables, supergiants and M stars. The maser lines are emitted from vibrationally excited states (see Table 16.2). These lines are highly time variable and rather intense. It is thought that the masers are excited by collisions, but this is not completely conclusive. The SiO masers are located closer to the star than are circumstellar H_2O masers, which are, in turn, closer than circumstellar OH masers. At present, it appears that the SiO masers are a part of the extended photosphere of the star, and their behavior reflects the motions in turbulent cells. The measured sizes of these maser regions are between 0.1 and 5 Astronomical Units.

The most recently discovered intense, widespread masers in the ISM are those of methanol CH_3OH. These can be placed in two distinct categories, depending on whether they are found close to compact H II regions or not. Methanol has an asymmetric top structure, with the $'O\text{-}H'$ arm tunneling through the maxima at the locations of the hydrogen atoms in the CH_3 group. As expected, there is a very large number of transitions. At first glance, the energy-level diagram for E type methanol (Fig. 15.10) has an overall structure which is somewhat like that of NH_3. However, there are no doublet levels, and transitions between K ladders are allowed. There is an additional difficulty with methanol in that the labelling of the energy levels is somewhat different from that of other asymmetric top molecules. For CH_3OH, there are two modifications, $A^{\pm}$ and E. These are based on the behavior of the $'O\text{-}H'$ arm of CH_3OH. The energy levels within each modification are labelled as J_k, where K can take on both positive and negative values. The masers observed are just those due to transitions between adjacent ladders. In contrast, the populations within a given ladder appear to be close to LTE. This difference of populations between ladders may be related to population transfers between the ground and vibrationally excited states, due to far infrared radiation, or by collisional selection rules which favor a change in ladders.

16.3   Models of Radiative Transfer

There are a number of radiative transfer approximations. Two sophisticated examples of radiative transfer models are Monte Carlo and Accelerated Lambda Iteration methods [see van Zadelhoff et al. (2002) for a discussion]. In the following we present a simple commonly used method.

16.3.1   The Large Velocity Gradient Model

In this section we will describe the treatment of photon transport in order to characterize the interaction between molecular populations and the radiation field due to the radiation from the molecule. Our previous treatment in regard to the CO

molecule was carried out under the assumption that the molecular populations were in LTE. The treatment we present next does not use this assumption. Furthermore as shown in the preceding section, a model of molecular excitation involving only two levels cannot account for complex excitation processes. The sketch of maser excitation in Sect. 16.2.2 illustrates this point dramatically. To obtain a more complete picture of molecular excitation, several spectral lines of the same molecule have to be observed. Commonly, some of these lines have $\tau > 1$, that is, they are optically thick. In addition, there may be deviations from LTE, so an analysis should include effects which lead to deviations from LTE.

In the following, we will present the most widely used method of radiative transfer which satisfactorily accounts for photon transport when lines are optically thick. This is the large velocity gradient (or LVG) approximation. First, we consider the LVG method qualitatively. It is assumed that the spherically symmetric cloud possesses large scale systematic motions so that the velocity is a function of distance from the center of the cloud, that is, $V = V_0(r/r_0)$. Furthermore, the systematic velocity is much larger than the thermal line width. Then the photons emitted at one position in the cloud can only interact with molecules which are nearby, and the global problem of photon transport is reduced to a local problem.

For excitation close to LTE, the qualitative effect of taking line optical depths into account can be illustrated by the following example: If a cloud is filled with 2 level molecular systems, we find that for a certain H_2 density, a portion of the systems must be in the upper level, if the photons can freely escape. If these photons are reabsorbed within the cloud the portion in the upper level will be larger than if the photons can freely escape, if all other factors remain the same. This is because the photons slowly diffuse out of the cloud.

In general, the excitation involves collision and radiation. The radiative excitation can take several forms: The most obvious such interaction is caused by radiation emitted by the same species. A photon emitted by one molecule in the cloud is absorbed by another nearby molecule. This is referred to as *photon trapping*. In addition to this molecular line radiation, there can be interactions with broadband radiation. These radiation sources include the 2.7 K background radiation; in some cases emission from dust may excite some transitions.

We present the LVG treatment for a uniform density cloud. This discussion, from (16.25) to (16.38) is based on a general treatment of radiative transfer. From (16.39) onward, the treatment is restricted to two-energy-level molecular systems. (This is *not* an essential feature of the LVG approximation, but is used here to illustrate basic principles.) We assume that all parameters depend only on the coordinate r. Rather than using u and l, in the following we will use i and j to label the levels, in order to preserve generality. A potentially important effect *not* included is the effect of local dust (for a discussion, see Deguchi 1981). This effect may be important for sub-millimeter line transitions in clouds with high densities.

The populations, n_i, of a level i are given by

$$n_i(r) \sum_j P_{ij} = \sum_j n_j P_{ji} \qquad (16.25)$$

where P_{ij} can be expressed in terms of the Einstein A and B coefficients, and the collision rates, C_{ij}. Then P_{ij} has the form,

$$P_{ij} = A_{ij} + B_{ij}\langle U_{ij}\rangle + C_{ij} \qquad \text{for} \quad i > j \tag{16.26}$$

and

$$P_{ij} = B_{ij}\langle U_{ij}\rangle + C_{ij} \qquad \text{for} \quad i < j. \tag{16.27}$$

Here $\langle U_{ij}\rangle$ is the average radiation field at the frequency of the transition from level i to j. $\langle U_{ij}\rangle$ is related to the line source function, S_{ij}, by a mathematical kernel $\langle K_{ij}\rangle$ containing the physics of the radiative transfer, cloud physics and structure,

$$\langle U_{ij}(r)\rangle = \int K_{ij}(|\mathbf{r} - \mathbf{r}'|)\, S_{ij}(r')dr'. \tag{16.28}$$

The boundary condition is that outside the cloud $\langle U_{ij}(r)\rangle$ equals the Planck function for a 2.7 K background, $\boldsymbol{B}_{ij}$, at the frequency of interest. Equation (16.28) can be simplified if there are large monotonic velocity gradients in the cloud. Then photons emitted at some point can be absorbed only within a distance $l \approx v_t R/V \ll R$, where V is the large scale velocity and v_t is the thermal velocity of the molecules. We assume that there is a complete redistribution of frequencies; that is, when a photon is absorbed and afterwards emitted, there is no "memory" of the previous frequency, direction and polarization. Then the source function can be taken outside the integral, and (16.28) becomes:

$$\langle U_{ij}(r)\rangle = (1 - \beta_{ij}(r))\, S_{ij} + \beta\, \boldsymbol{B}_{ij}(\nu_{ij}, T_{\text{BB}}). \tag{16.29}$$

Here $\beta_{ij}(r)$ is the probability that a photon emitted in the transition from level i to j at a radius r will escape from the cloud. The second term represents the interaction with the 2.7 K radiation field. If all the photons escape, $\beta_{ij}(r) = 1$ and $\langle U_{ij}(r)\rangle$ is the blackbody radiation field intensity. If no photons escape from the cloud, $\langle U_{ij}(r)\rangle$ is the source function. This source function must depend on the molecular level populations by the expression given below:

$$S_{ij} = \frac{2h\nu_{ij}^3}{c^2}\frac{1}{g_j n_j/g_i n_i - 1}. \tag{16.30}$$

Equation (16.30) couples the radiation field and the molecular populations. The expression for the optical depth is needed to obtain the form of $\beta_{ij}(r)$. For a given direction, given by the direction cosine, μ, and frequency, ν, the optical depth is:

$$\tau_{ij}(\nu, r, \mu) = \int k_0(r, s, \mu)\, \varphi\left(\nu - \nu_0 + \frac{\nu_0 s}{c}\frac{dv_s}{ds}\right)ds \tag{16.31}$$

where k_0 is the standard absorption coefficient, given in (12.17), and φ is the normalized line shape function (taken to be a Gaussian in most cases). If the velocity gradient is large, φ is nonzero only at $s = 0$ and can be taken outside the integral.

By assumption the value of the gradient $dv_s/ds = \alpha$ is also a constant, defined as α_0 and can be taken outside the integral. This is *the* crucial step in the analysis

$$\tau(v,r,\mu) = k_0(r)\,\frac{c}{\alpha_0\,v_0}\int \varphi(x')\,dx' \tag{16.32}$$

where $x' = v - v_0 + \dfrac{v_0}{c}\,\alpha_0\,\dfrac{s}{c}$. The escape probability in the line, averaged over the line and all angles, is

$$\beta(r) = \frac{1}{4\pi}\int\!\!\int \varphi(x')\,\exp(-\tau_{ij}(x,r,\mu))\,d\Omega\,dx'. \tag{16.33}$$

Changing the variable to

$$y = \int \varphi(x')\,dx' \tag{16.34}$$

and using the normalization

$$\int \varphi(x')\,dx' = 1, \tag{16.35}$$

we obtain

$$\beta(r) = \frac{1}{2}\int \frac{1 - \exp\big(-\tau_{ij}(r,\mu)\big)}{\tau_{ij}(r,\mu)}\,d\mu \tag{16.36}$$

where $\tau_{ij}(r,\mu) = (k_0(r)\,c)/(v_0\,\alpha_0(r,\mu))$. In this plane parallel geometry, we have set $(dv_s/ds) = \mu^2(dv/dz)$, where z is the depth and v is the velocity normal to the plane. Applying the Eddington approximation, which is equivalent to replacing μ^2 by $1/3$, the escape probability becomes

$$\beta(r) = \frac{1 - \exp(-3\,\tau_{ij})}{3\,\tau_{ij}}. \tag{16.37}$$

The plane parallel result differs from the spherical cloud result for large values of τ_{ij}. Substituting this result into (16.29), we have

$$\langle U_{ij}(r)\rangle = 1 - \frac{1 - \exp(-3\,\tau_{ij})}{3\,\tau_{ij}}\,S_{ij} + \frac{1 - \exp(-3\,\tau_{ij})}{3\,\tau_{ij}}\,B_{ij}(v_{ij},T_{BB}). \tag{16.38}$$

This is the radiation field in the cloud. In order to qualitatively show the effect of this field on molecules, we will first analyze a two-level system, and then we will show results for numerical models of multilevel systems of CO and CS.

To illustrate the effects of photon trapping in the LVG approximation, we will use (16.38) in (16.3):

$$n_i\,(C_{ij} + B_{ij}\bar{U}) = n_j\,(A_{ji} + B_{ji}\bar{U} + C_{ji}). \tag{16.39}$$

We neglect the 2.7 K background in (16.38) and substitute this expression into (16.39). After grouping terms and using relations $T_0 = h\,v/k$ and (16.5) and (16.8), we have

$$\frac{T}{T_0} = \frac{T_k/T_0}{1 + T_k/T_0 \ln\left[1 + \dfrac{A_{ji}}{3C_{ji}\,\tau_{ij}}\left(1 - \exp\left(-3\,\tau_{ij}\right)\right)\right]} \ . \tag{16.40}$$

The term $(1 - \exp(-3\,\tau_{ij}))/\tau_{ij}$ is caused by "photon trapping" in the cloud. If $\tau_{ij} \gg 1$, the case of interest, then A_{ij} is replaced by A_{ij}/τ_{ij}. In a qualitative way this additional factor explains why the spontaneous decay out of the $J = 1$ level of CO is 100 times lower (for $\tau \approx 100$), and thus the $J = 1 \to 0$ line of CO can arise in the lower density part of the cloud. In contrast, decay out of the $J = 1$ level for ^{13}CO with $\tau \approx 2$, is only two times lower, and thus the $J = 1 \to 0$ line arises from denser regions. For C^{18}O emission, with $\tau < 1$ and the $J = 1 \to 0$ line must arise from still denser regions.

In Fig. 16.3, we show calculations which can be used to determine X, the ratio of total abundance of CO relative to H$_2$ (per velocity gradient) and local H$_2$ density from the measurement of two transitions of CO. The plots show results for two different values of T_K. A comparison of the two sets of curves shows that for $T_K = 10$ K,
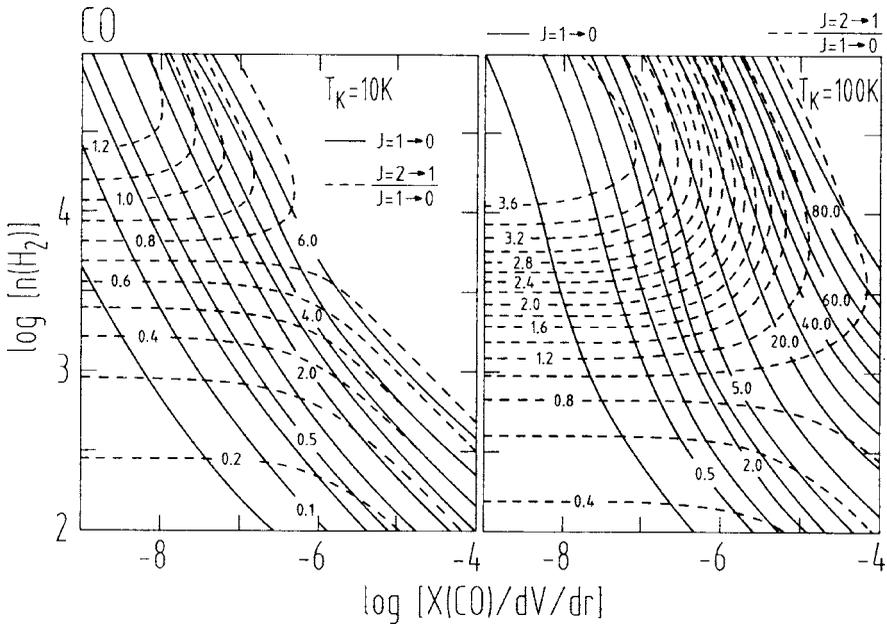
Fig. 16.3 Plots of the line intensities calculated using the LVG approximation for the two lowest rotation transitions of the CO molecule. There are four input quantities. The first two are the kinetic temperature of the cloud, and X, where X is the abundance of CO relative to H$_2$, per velocity gradient, dV/dR. The other two inputs are measurements of the two lines in question. It is usual to estimate this gradient from the ratio of the cloud size to typical linewidth, expressed in units of km s^{-1} per pc. Typical values are in the range 1–100 km s^{-1} pc^{-1}. An upper limit for the value of X(CO) is 10^{-4}. The emission region is expected to be larger than the telescope beam, so the peak temperatures are not beam diluted. [Adapted from Goldsmith et al. (1983)]

the kinetic temperature is more important for setting a limit to the ratio of the line temperatures. That is, if the CO in the cloud is smoothly distributed and if the line ratio of the two lowest transitions of CO give a low value, the cloud has either a low H_2 density or a low T_K. Using additional information, it might be possible to decide between these alternatives. For 100 K, in the limit of low X and high H_2 density, the ratio approaches 4. This is the limit expected for optically thin, thermalized CO lines from a hot medium [cf. (15.23) and (15.25)]. In examining the plot for $T_K = 100$ K, we see that the analysis of these CO lines cannot be used to determine densities larger than 10^4 cm^{-3}: the line ratio of unity usually indicates large optical depths and thermalization in the transitions. Provided that the measurements are not strongly affected by noise and that the lines of CO arise from the same volume, then the measurements of CO correspond to a point on these curves, so that the H_2 density and column density of CO *can* be obtained. For a consistent pair of measurements, the LVG approximation gives a value of $n(H_2)$ and $X/(dV/dR)$. Here dV is usually taken to be the measured linewidth, and dR the cloud size. If one multiplies $X/(dV/dR)$ by $n(H_2)$, one obtains the total CO abundance per velocity gradient. Multiplying the CO abundance per velocity gradient by the observed linewidth, one obtains the CO column density. To obtain the H_2 column density from this result, one must then *assume* a CO to H_2 ratio. This is usually close to the "cosmic" value of $(C/H) = 3 \times 10^{-4}$. A commonly used value is $(CO/H_2) = 10^{-4}$; this is based on the *assumption* that 15% of C is in the form of CO.

An application of the LVG analysis to carbon monosulfide, CS, allows the estimation of larger densities since the spontaneous decay rates for CS transitions are faster than for the corresponding transitions of CO (see our Table 16.1). Also, the abundance of CS is much lower than that of CO, so line optical depths should be smaller (Fig. 16.4). It is *not* at all clear what the relative abundance of CS to H_2 should be. There have been a number of estimates; these are in the range $\sim 10^{-9}$, but a factor of 10 larger or smaller could be possible. This is because CS is much more chemically reactive than CO. So CS might be chemically converted into other species at higher densities. In contrast, CO is chemically stable and quite abundant, so if CO were chemically converted into other species, the chemistry of the cloud would be greatly affected, and this would give rise to observable differences.

In summary, the LVG approximation requires at least two measurements of a species in addition to the kinetic temperature. From the data and the model one can determine $n(H_2)$. One could also use the LVG approach to determine another parameter, such as either the abundance of the molecule in question relative to H_2, per velocity gradient, or the column density of the molecule per linewidth. These two results are equivalent, since the H_2 density is determined in this analysis and the gradient is taken to be $\Delta V/\Delta R$, where ΔV is the linewidth and ΔR is the cloud size. It is important to note that in the case of linear molecules estimates of H_2 density and kinetic temperature are coupled closely. For example, if one detects the $J = 1 - 0$ line of CO but has a limit of < 0.4 for the ratio of the $J = 2 - 1$ to $J = 1 - 0$ line, and no estimate of kinetic temperature,

Table 16.1 Parameters of the more commonly observed molecular lines

Chemical[a] formula	Molecule name	Transition	ν/GHz[b]	E_u/K[c]	A_{ij}/s^{-1}[d]
OH	hydroxyl radical	$^2\Pi_{3/2} F = 1 - 2$	1.612231	0.1	1.3×10^{-11}
OH	hydroxyl radical	$^2\Pi_{3/2} F = 1 - 1$	1.665400	0.1	7.1×10^{-11}
OH	hydroxyl radical	$^2\Pi_{3/2} F = 2 - 2$	1.667358	0.1	7.7×10^{-11}
OH	hydroxyl radical	$^2\Pi_{3/2} F = 2 - 1$	1.720529	0.1	0.9×10^{-11}
H_2CO	ortho-formaldehyde	$J_{K_a K_c} = 1_{10} - 1_{11}$	4.829660	14	3.6×10^{-9}
CH_3OH	methanol*	$J_K = 5_1 - 6_0 A^+$	6.668518	49	6.5×10^{-10}
HC_3N	cyanoacetylene	$J = 1 - 0, F = 2 - 1$	9.009833	0.4	3.8×10^{-8}
CH_3OH	methanol**	$J_K = 2_0 - 3_{-1} E$	12.178593	12	8.2×10^{-9}
H_2CO	ortho-formaldehyde	$J_{K_a K_c} = 2_{11} - 2_{12}$	14.488490	22	3.2×10^{-8}
C_3H_2	ortho-cyclopropenylidene	$J_{K_a K_c} = 1_{10} - 1_{01}$	18.343137	0.9	3.9×10^{-7}
H_2O	ortho-water*	$J_{K_a K_c} = 6_{16} - 5_{23}$	22.235253	640	1.9×10^{-9}
NH_3	para-ammonia	$(J, K) = (1, 1) - (1, 1)$	23.694506	23	1.7×10^{-7}
NH_3	para-ammonia	$(J, K) = (2, 2) - (2, 2)$	23.722634	64	2.2×10^{-7}
NH_3	ortho-ammonia	$(J, K) = (3, 3) - (3, 3)$	23.870130	122	2.5×10^{-7}
SiO	silicon monoxide*	$J = 1 - 0, v = 2$	42.820587	3512	3.0×10^{-6}
SiO	silicon monoxide*	$J = 1 - 0, v = 1$	43.122080	1770	3.0×10^{-6}
SiO	silicon monoxide	$J = 1 - 0, v = 0$	43.423858	2.1	3.0×10^{-6}
CS	carbon monosulfide	$J = 1 - 0$	48.990964	2.4	1.8×10^{-6}
DCO$^+$	deuterated formylium	$J = 1 - 0$	72.039331	3.5	2.2×10^{-5}
SiO	silicon monoxide*	$J = 2 - 1, v = 2$	85.640456	3516	2.0×10^{-5}
SiO	silicon monoxide*	$J = 2 - 1, v = 1$	86.243442	1774	2.0×10^{-5}
H^{13}CO$^+$	formylium	$J = 1 - 0$	86.754294	4.2	3.9×10^{-5}
SiO	silicon monoxide	$J = 2 - 1, v = 0$	86.846998	6.2	2.0×10^{-5}
HCN	hydrogen cyanide	$J = 1 - 0, F = 2 - 1$	88.631847	4.3	2.4×10^{-5}
HCO$^+$	formylium	$J = 1 - 0$	89.188518	4.3	4.2×10^{-5}
HNC	hydrogen isocyanide	$J = 1 - 0, F = 2 - 1$	90.663574	4.3	2.7×10^{-5}
N_2H^+	diazenylium	$J = 1 - 0, F_1 = 2 - 1,$			
		$F = 3 - 2$	93.173809	4.3	3.8×10^{-5}
CS	carbon monosulfide	$J = 2 - 1$	97.980968	7.1	2.2×10^{-5}
C^{18}O	carbon monoxide	$J = 1 - 0$	109.782182	5.3	6.5×10^{-8}
^{13}CO	carbon monoxide	$J = 1 - 0$	110.201370	5.3	6.5×10^{-8}
CO	carbon monoxide	$J = 1 - 0$	115.271203	5.5	7.4×10^{-8}
H$_2$^{13}CO	ortho-formaldehyde	$J_{K_a K_c} = 2_{12} - 1_{11}$	137.449959	22	5.3×10^{-5}
H_2CO	ortho-formaldehyde	$J_{K_a K_c} = 2_{12} - 1_{11}$	140.839518	22	5.3×10^{-5}
CS	carbon monosulfide	$J = 3 - 2$	146.969049	14.2	6.1×10^{-5}
C^{18}O	carbon monoxide	$J = 2 - 1$	219.560319	15.9	6.2×10^{-7}
^{13}CO	carbon monoxide	$J = 2 - 1$	220.398714	15.9	6.2×10^{-7}
CO	carbon monoxide	$J = 2 - 1$	230.538001	16.6	7.1×10^{-7}
CS	carbon monosulfide	$J = 5 - 4$	244.935606	33.9	3.0×10^{-4}
HCN	hydrogen cyanide	$J = 3 - 2$	265.886432	25.5	8.5×10^{-4}
HCO$^+$	formylium	$J = 3 - 2$	267.557625	25.7	1.4×10^{-3}
HNC	hydrogen isocyanide	$J = 3 - 2$	271.981067	26.1	9.2×10^{-4}

[a] If isotope not explicitly given, this is the most abundant variety, i.e., ^{12}C is C, ^{16}O is O, ^{14}N is N, ^{28}Si is Si, ^{32}S is S.

[b] The line rest frequency from Lovas (1992).

[c] Energy of upper level above ground, in Kelvin.

[d] Spontaneous transition rate, i.e., the Einstein A coefficient.

* Always found to be a maser transition.

** Often found to be a maser transition.
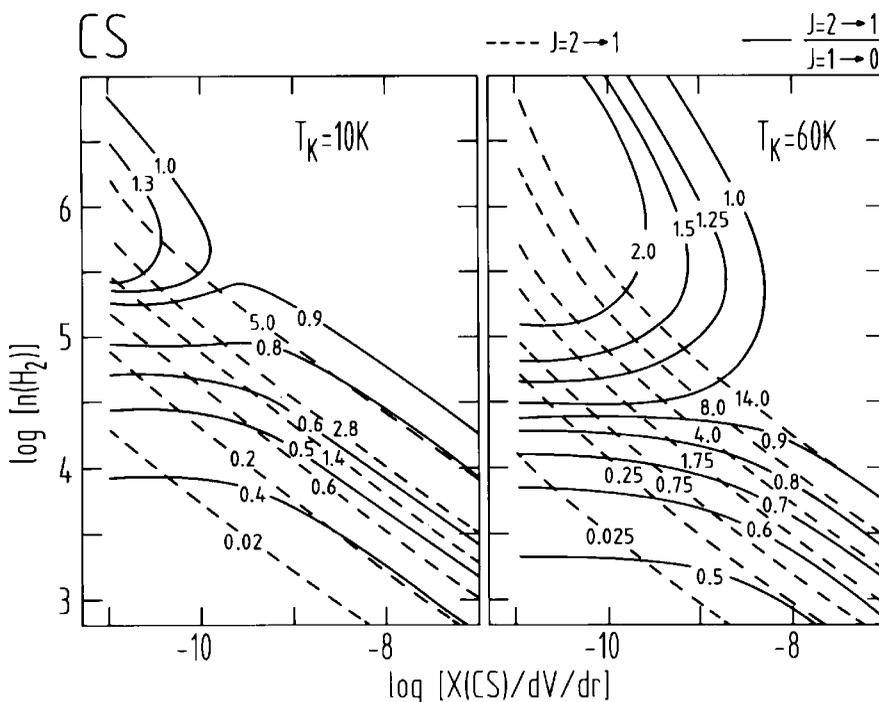
Fig. 16.4 Plots of the line intensities calculated using the LVG approximation for the lower rotational transitions of the CS molecule. The input quantities are the kinetic temperature of the cloud, dV/dR, and measurements of the two lines in question. An internally consistent solution gives a value of $n(H_2)$ and the abundance of CS relative to H_2, X, per velocity gradient. In this analysis, the emission region is expected to be larger than the telescope beam, so the calculated peak temperatures are the observed temperatures; otherwise the line emission must be corrected for beam dilution. [Adapted from Linke and Goldsmith (1978)]

either the density is 150 cm^{-3} and $T_k > 100$ K, or $T_k = 10$ K or less and density could be $> 10^3$ cm^{-3}.

16.4 Spectral Lines as Diagnostic Tools

We will now indicate applications of the methods described in the previous sections in order to determine cloud parameters. Inherent in these measurements is the limitation that for a given direction, spectra provide weighted, line of sight averaged estimates of FWHP linewidth, $\Delta V_{1/2}$, radial velocity, V_{lsr}, beam averaged line intensity in Kelvin, polarization percentage, and in some cases, line optical depths.

16.4.1 Kinetic Temperatures

A crucial input parameter for the LVG calculations is T_K. Linewidths of thermally excited species provide a definite value for T_K if the turbulent velocity can be neglected. The relation is

$$T_k = 21.2 \, (m/m_H) \, (\Delta V_t)^2$$

where m is the molecule, m_H is the mass of hydrogen, and ΔV_t is the FWHP thermal width. Such a relation has been applied to NH_3 lines in quiescent dust clouds.

Historically, T_K was obtained at first from the peak intensities of rotational transitions of CO, which has a small spontaneous decay rate. From the ratio of CO to ^{13}CO line intensities, $\tau(CO) \gg 1$. After correcting for cloud size, from (15.30) T_{MB} and T_{ex} are directly related. In addition, the large optical depths reduce the critical density by a factor τ, the line optical depth, so that for the $J = 1 \rightarrow 0$ line, the value of the critical density $n^* \approx 50 \, cm^{-3}$. Then, the level populations are determined by collisions, so the excitation temperature for the $J = 1 \rightarrow 0$ transition is T_K. Usually, it is assumed that the beam filling factor is unity; for distant clouds or external galaxies, the filling factor is clearly less.

An alternative to CO measurements makes use of the fact that radiative transitions between different K ladders in symmetric top molecules are forbidden. Then the populations of the different K ladders are determined by collisions. Thus a method of determining T_K is to use the ratio of populations in different K ladders of molecules such as NH_3, CH_3CCH or CH_3CN. This is also approximately true for different K_a ladders of H_2CO. Since ratios are involved, beam filling factors play no role. Even for extended clouds, T_k values from CO and NH_3 may not agree. This is because NH_3 is more easily dissociated so must arise from the cloud interior. Thus for a cloud heated externally, the T_k from CO data will be larger than that from NH_3.

For NH_3, the rotational transitions, $(J+1,K) \rightarrow (J,K)$, occur in the far infrared and have Einstein A coefficients of order $1 \, s^{-1}$; for inversion transitions, A values are $\sim 10^{-7} \, s^{-1}$ (see, e.g., Table 14.2). These non metastable states require extremely high H_2 densities or intense far infrared fields to be populated. Thus NH_3 non metastable states ($J > K$) are not suitable for T_K determinations. Rather, one measures the inversion transitions in different metastable ($J = K$) levels. Populations cannot be transferred from one metastable state to another via allowed radiative transitions. This occurs via collisions, so the relative populations of metastable levels are directly related to T_K. The column densities are obtained from the inversion transitions from different metastable states, and converts these to column densities using (15.25, 15.26, 15.27 and 15.28). If the NH_3 lines are optically thick, one can use the ratios of satellite components to main quadrupole hyperfine components, in most cases, to determine optical depths. A large number of T_K determinations have been made using NH_3 in dark dust clouds. Usually these involve the $(J,K) = (1,1)$ and (2,2) inversion transitions.

From the ratio of column densities in the Boltzmann relation, one obtains a rotational temperature, T_{rot}. This temperature describes the relative populations in
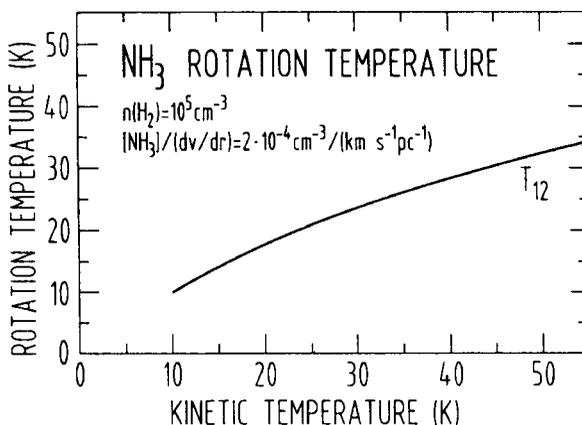
Fig. 16.5 A plot of the kinetic temperature, T_K, as a function of T_{rot} obtained from the column densities of the (1,1) and (2,2) inversion lines of NH_3. The H_2 density and NH_3 density per gradient are as shown [Adapted from Schilke (1989)]

different K ladders of the NH_3 molecule. Although the (2,2) and (1,1) populations are not linked by radiative transitions, collisions can cause a transfer of population from the (2,2) levels to the (2,1) levels and by radiative decay to the (1,1) levels. The very fast decay from the (2,1) to the (1,1) levels normally causes T_{rot} to be an underestimate of T_K. Only at very high densities is this population exchange equal in both directions. From model calculations, such as the one shown in Fig. 16.5, one can correct for this bias, to obtain a reliable estimate of T_K.

16.4.2 Linewidths, Radial Motions and Intensity Distributions

From the spectra themselves, the linewidths, $\Delta V_{1/2}$, and radial velocities, V_{lsr}, give an estimate of motions in the clouds. The $\Delta V_{1/2}$ values are a combination of thermal and turbulent motions. Observations show that the widths are supersonic in most cases. In cold dense cores, motions barely exceed Doppler thermal values. Detailed measurements of lines with moderate to large optical depths show that the shapes are nearly Gaussian. However, simple models in which unsaturated line shapes are Gaussians would give flat topped shapes at high optical depths. This is not found. More realistic models of clouds are those in which shapes are determined by the relative motion of a large number of small condensations, or clumps, which emit optically thick line radiation. If the motions of such small clumps are balanced by gravity, one can apply the virial theorem. Images of isolated sources can be used for comparing with models. One example is the attempt to characterize the kinetic temperature and the H_2 density distributions from spectral line or thermal dust emission data. Clearly there are a number of parameters in any comparison. An example

of fits to a pre-protostellar region is given in Galli et al. (2002). For lower density, more extended clouds, one can interpret sharp velocity changes in a cloud as evidence for the presence of shock waves (see e.g. Megeath and Wilson (1997)), but a confirmation requires high angular resolution imaging.

16.4.3 Determination of H_2 Densities

To obtain a reliable determination of H_2 densities $n(H_2)$ one must measure at least two spectral lines of a given species. To interpret these data, one needs an estimate of the kinetic temperature, collision rates, and a radiative transport model. One can assume that the lines are optically thin, and use a statistical equilibrium model, but the present approach is to apply the LVG model to these data. We show examples of this approach in Figs. 16.3 and 16.4. Clearly, the more lines measured, the more reliable the result. For linear molecules such as CO or CS, it is not possible to separate kinetic temperature and density effects. For example, CO with $T_K = 10\,K$ will have a $J = 2$-1 line much weaker than the $J = 1$-0 line no matter how high the density. However, if the plot of CO line intensity versus frequency shows a *turn over*, that is a decrease in intensity, it is possible to find a unique combination of T_K and $n(H_2)$.

16.4.4 Estimates of H_2 Column Densities

The CO molecule has four properties which allow a good estimate of the total column density of H_2, namely; a high line intensity, low critical density, an excitation close to LTE, and a large abundance relative to H_2. From all of the radio observations made, it appears that the abundance of CO relative to H_2 seems to be close to 10^{-4} in most cases. There have been long discussions about which isotopomer of CO is best suited for this purpose. For regions with densities $< 10^2\,cm^{-3}$, $^{12}C^{16}O$ is the best choice, for somewhat denser regions, ^{13}CO, and for $n(H_2) > 500\,cm^{-3}$, $C^{18}O$ is the best choice, since the transitions are optically thin. As can be seen from the LTE treatment of the $J = 1 \rightarrow 0$ line (15.36), the total column density of CO varies directly with the value of $T_{ex} = T_K$. For an LTE treatment of the $J = 2 \rightarrow 1$ line, the effect of T_K is somewhat less, since the $J = 1$ energy level is 5.5 K above the ground state (15.37).

An LVG treatment of the dependence of the total column density on the line intensity of the $J = 2 \rightarrow 1$ line shows that a simple relation is valid for T_K from 15 K to 80 K, and $n(H_2)$ from $\sim 10^3$ to $\sim 10^6\,cm^{-3}$. An assumption used in obtaining this relation is that the ratio of $C^{18}O$ to H_2 is 1.7×10^{-7}, which corresponds to $(C/H_2) = 10^4$, and $(^{16}O/^{18}O) = 500$. The latter ratio is obtained from isotopic studies for molecular clouds near the Sun. Then we have

$$N_{H_2} = 2.65 \times 10^{21} \int T_{MB}(C^{18}O, J = 2 \rightarrow 1)\, dv. \tag{16.41}$$

The units of v are $km\,s^{-1}$, of $T_{MB}(C^{18}O, J = 2 \rightarrow 1)$ are Kelvin, and of N_{H_2} are cm^{-2}. This result can be used to determine cloud masses, if the distance to the cloud is known, by a summation over the cloud, position by position, to obtain the total number of H_2. There is an additional 36% of mass in other constituents, mostly in helium. The mass obtained from the method given above, or similar methods, is sometimes referred to as the "CO mass"; this terminology can be misleading, but is frequently found in the literature.

16.4.5 Masses of Molecular Clouds from Measurements of $^{12}C^{16}O$

In large scale surveys of the CO $J = 1 \rightarrow 0$ line in our galaxy and external galaxies, it has been found, on the basis of a comparison of CO with ^{13}CO maps, that the CO integrated line intensities measure mass, even though this line is optically thick. The line shapes and intensity ratios along different lines of sight are remarkably similar for both ^{12}CO and ^{13}CO line radiation. This can be explained if the total emission depends primarily on the number of clouds. If so, ^{12}CO line measurements can be used to obtain estimates of N_{CO}^{12}. Observationally, in the disk of our galaxy, the ratio of these two quantities varies remarkably little for different regions of the sky.

This empirical approach has been followed up by a theoretical analysis. The basic assumption is that the clouds are virial objects, with self-gravity balancing the motions. If these clouds are thought to consist of a large number of clumps, each with the same temperature, but sub thermally excited (i.e., $T_{ex} < T_K$), then from an LVG analysis of the CO excitation, the peak intensity of the CO line will increase with $\sqrt{n(H_2)}$, and the linewidth will also increase by the same factor, as can be seen from (16.45). The exact relation between the integrated intensity of the CO $J = 1 \rightarrow 0$ line and the column density of H_2 must be determined empirically. Such a relation has also been applied to other galaxies and the center of our galaxy. However, the environment, such as the ISRF, may be very different and this may have a large effect on the cloud properties. For the disk of our galaxy, a frequently used conversion factor is:

$$N_{H_2} = X \int T_{MB}(CO, J = 1 \rightarrow 0)\, dv$$
$$= 2.3 \times 10^{20} \int T_{MB}(CO, J = 1 \rightarrow 0)\, dv. \tag{16.42}$$

where $X = 2.3 \times 10^{20}$ and N_{H_2} is in units of cm^{-2}. By summing the intensities over the cloud, the mass in $M_\odot$ is obtained. Strictly speaking, this relation is only valid for whole clouds. The exact value of the conversion factor between CO integrated line intensity and mass, X, is a matter of some dispute. Most large-scale surveys are restricted to the $J = 1 - 0$ line of CO. In Fig. 16.6, we show the distribution
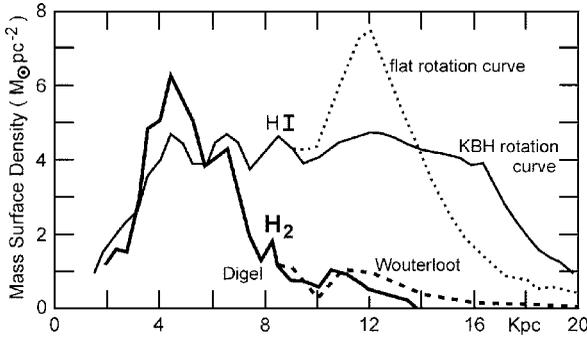
Fig. 16.6 A plot of the surface density of CO and HI as a function of distance from the galactic center. Dame estimates that the total mass in HI in the disk is $5.2 \pm 1.2 \ 10^9 \ M_\odot$, the mass in H2 is $1.5 \pm 0.2 \ 10^9$. Fich and Tremaine (1991) estimate a total mass of about $4 \ 10^{11} \ M_\odot$. [Figure adapted from Dame (1993)]

of HI and H_2 deduced from 21 cm and CO surveys. The H_2 mass estimates were obtained from an application of (16.42). The differing amounts of HI in the outer galaxy are caused by different choices of rotation curves. The general opinion is that the coefficient in (16.42) should be 0.3–0.5 for the galactic center region, and perhaps ~ 5 for the outer galaxy. From the CO surveys, large clouds, so-called Giant Molecular Clouds (GMC's) contain 90% of the total mass of H_2, and follow the relationships

$$\sigma \sim R^{0.5}$$

where σ is the linewidth and R is the cloud size, another relation is,

$$M \sim R^2 \ .$$

If (16.42) holds, one can obtain a relation similar to that of Tully-Fisher, namely $L_{CO} \sim \sigma^4$.

16.4.6 The Correlation of CO and H_2 Column Densities

One widely used method involves estimates of the column densities of carbon monoxide. As far as we know, CO is the most abundant polar molecule in the ISM; thus we believe that there should be other noticeable effects if interstellar chemistry should affect the abundance of CO. The abundance of CO can be determined with some confidence since the rotational transitions are emitted in LTE and there are a large number of isotopomer lines. Clearly it is of interest to investigate whether there is a relation between the column density of CO, N_{CO}, and $N(H_2)$. If so, CO could be used as a mass tracer. Even if we assume that interstellar chemistry and excitation

do not affect the CO molecule, the destruction of CO by UV photons must be taken into account. Outside dust clouds, the interstellar radiation field contains enough UV quanta able to photo dissociate most molecules so that the molecular fractional abundance will be small. But with increasing depth in the clouds, the radiation will be attenuated. This results in a corresponding increase of the fractional abundance of molecular species. The CO molecule is dissociated by UV spectral lines, so different isotopomers will have different destruction thresholds. From this qualitative sketch, it should be expected that CO will be observed only if A_V is above a certain threshold value. Frerking et al. (1982) found a value $A_V > 1^m$ for ^{13}CO and $A_V > 1.9^m$ for $C^{18}O$.

CO can be used as a tracer of molecular hydrogen in interstellar clouds where $A_V > 1^m$. The mass of the molecular cloud should be determined from observations of an isotope of CO that is readily observable, optically thin and not affected by enhancements of the isotope by fractionation. Comparing CO abundances with H_2 column densities obtained in the infrared and the UV, Frerking et al. (1982) give relations for dense cores of clouds. If we use the standard relation between visual extinction in magnitudes and column density of H_2, we have

$$
\begin{aligned}
N(H_2) &= \left[\frac{N(C^{18}O)}{2.29 \times 10^{14}} \right] \times 10^{21}\,cm^{-2} \quad \text{for} N(H_2) < 1.5 \times 10^{22}\,cm^{-2} \\
&= \left[\frac{N(^{13}CO)}{2.18 \times 10^{15}} \right] \times 10^{21}\,cm^{-2} \quad \text{for} N(H_2) < 5 \times 10^{21}\,cm^{-2}.
\end{aligned}
\tag{16.43}
$$

As noted by the authors of various such studies, a scatter of factor two to five in the correlations is always present; Lada et al. (1994) noted that the scatter increases with increasing extinction. If this is random, this is a fundamental limit, perhaps caused by small scale structure in the clouds which allows UV to penetrate deep into the interior. Kramer et al. (1998) noted a decrease in the $C^{18}O$ to H_2 ratio for the darkest parts of IC5146, indicating a freezing out of $C^{18}O$.

16.4.6.1   Column Densities from Dust Measurements

In the ISM, one expects a more-or-less constant dust-to-gas ratio, since dust is made up of *metals* such as carbon and silicon, and the mass fraction in metals is 1%. Thus, from the column density of dust one can estimate the total column density of gas. In molecular clouds, the gas will be in the form of H_2 which cannot be measured under most circumstances. However, there are relations between the column density of dust and visual extinction, A_V. In nearby regions of lower extinction, we can measure A_V from star counts or measurements of the infrared extinction. In darker regions or regions far from the Sun, extinction must be measured using indirect methods, for example by measuring the column density of H_2. Another method to determine column densities is the use of dust continuum emission. This is important in very dense, cold regions, since from a simplified (perhaps oversimplified) theory,

there is a gas-grain collision given by (16.1). For H_2 densities $> 10^6 \, \mathrm{cm}^{-3}$, t_{mg} is 3000 years, short compared to most other time scales. Then CO and most other molecules might be condensed out of the gas phase, so that spectroscopic measurements cannot be used as a probe of very dense regions. An alternative (for column densities but not dynamics) are measurements of the thermal emission from dust. The most critical step is making a quantitative connection between τ_{dust} and $N(H_2)$. All astronomical determinations are based on the analysis of Hildebrand (1983). For $\lambda > 100 \, \mu m$, Mezger et al. (1990) used the Rayleigh-Jeans approximation to obtain the following relation (presented previously in 10.7):

$$N(\mathrm{H}) = 1.93 \times 10^{24} \frac{S_\nu}{\theta^2} \frac{\lambda^4}{Z/Z_\odot \, b \, T_{dust}} , \tag{16.44}$$

where the dust flux densities, S_ν, are in mJy, the source FWHP sizes, θ, in arc seconds, and wavelengths, λ in mm, the column density of hydrogen in all forms, $N(\mathrm{H})$, in cm^{-2}. The value of Z for the Sun is $Z_\odot$, and b is an adjustable factor used to take into account changes in grain sizes. At present it is believed that $b = 1.9$ is appropriate for moderate density gas and $b = 3.4$ for very dense gas, although Ossenkopf and Henning (1994) report that a range of ± 2.5 is likely. Since dust emission measurements using ground based telescopes usually require beam switching, the results emphasize compact structures.

16.4.7 Mass Estimates and Cloud Stability

We repeat here the result obtained in Chap. 12. If we assume that only gravity is to be balanced by the motions in a cloud, then, for a uniform density cloud of radius R, in terms of the line of sight FWHP velocity, virial equilibrium requires:

$$\boxed{\frac{M}{M_\odot} = 250 \left(\frac{\Delta v_{1/2}}{\mathrm{km\,s}^{-1}} \right)^2 \left(\frac{R}{\mathrm{pc}} \right)} \quad . \tag{16.45}$$

Once again very optically thick lines should not be used in determining masses using (16.45).

It is interesting to investigate what occurs if isolated clouds are *not* stable. This is a very complex question in the general case where clumping and large temperature differences are present. If we consider only a uniform density isothermal cloud, there are two possibilities: (1) the gravitational attraction is too weak to bind the cloud, in which case the cloud disperses in a time $(D/\Delta V_{1/2})$, or the cloud must be confined by external pressure or (2), the motion of gas in the cloud is too small to balance gravity, in which case the cloud collapses. The collapse of a uniform, pressure free, non rotating cloud will occur in a *free fall* time. From the integration of the equation of motion of a spherical region, we have

$$t_{ff} = \sqrt{\frac{3\pi}{32\,G\rho(0)}}\,, \tag{16.46}$$

where G is the gravitational constant, and $\rho(0) = m(H_2)n(H_2)$ is the initial density of the cloud. In CGS units, this is

$$t_{ff} = 5 \times 10^7/\sqrt{n(H_2)}\,, \tag{16.47}$$

where the free fall time t_{ff} is given in years if the H_2 density $n(H_2)$ is in cm^{-3}.

The estimate given above leads to an interesting conclusion if applied to clouds in our galaxy. If one takes the average H_2 density in giant molecular clouds as 10^2 cm^{-3}, the collapse time is 5×10^6 years. There are >500 such clouds each with mass $10^6 M_\odot$, so the average star formation rate must be $\sim 5 \times 10^2\,M_\odot$ year^{-1}. This is much too large to be compatible with models of the galaxy, from direct observations of selected clouds, so there must be a support mechanism. This could be either turbulence or magnetic fields. Supersonic turbulence has a rapid decay time, so if present, there must be a source to resupply the turbulent energy.

The support may be provided by magnetic fields. If so, it is likely that magnetohydrodynamic waves are present. An estimate of the contribution of such waves to the observed linewidth can be obtained by considering the Alfvén velocity, v_a, in km s^{-1}:

$$V_a = 1.1 \left(\frac{B}{\mu G}\right)\left(\frac{n}{cm^{-2}}\right)^{-1/2}. \tag{16.48}$$

In regard to linewidths, this situation might be described by the phrase *linewidths appear to be supersonic, but may be sub-Alfvénic*. For a typical cloud of density 10^3 cm^{-3}, with a $10\,\mu$ Gauss field, the Alfvén wave velocity would be 0.3 km s^{-1}.

The strengths of line-of-sight magnetic fields can be obtained from circular polarization measurements of the Zeeman effect (see, e.g., Chap. 12). However linear polarization has been detected in CO lines. This is caused by a combination of magnetic field, excitation and optical depth effects. This is the *Goldreich-Kylafis* effect (Goldreich and Kylafis 1982). Crutcher (2008) has reviewed both line and continuum polarization measurements. Even if the magnetic field strength is known, the coupling of this field to the neutral gas also depends on the relative electron abundance. Estimates of the relative electron abundances can be estimated on the basis of HCO$^+$ measurements, if the electron recombination rates are known. See Sect. 16.61 for a discussion of the use of HCO$^+$ measurements to determine electron abundances. According to standard theory [see, e.g., Spitzer (1978)], if the relative electron abundance is [e] $\geq 10^{-7}$, the magnetic field will be coupled to the gas and this will help to prevent collapse. If the relative electron density is less than 10^{-7}, the gas and magnetic field are decoupled. Then if the molecular gas is gravitationally unstable, magnetic effects will not hinder collapse of the molecular cloud. A detailed discussion of cloud collapse is given in Shu et al. (1987).

If magnetic fields have a large effect on line widths, the virial mass is an upper limit. The magnetic energy density is given by $B^2/8\pi$. If one assumes that there is equipartition between magnetic, kinetic and gravitational energies, one can obtain,

$$\frac{B}{\mu G} = 15.2 \left(\frac{\Delta v_{1/2}}{\mathrm{km\,s^{-1}}}\right)^2 \left(\frac{R}{\mathrm{pc}}\right).\tag{16.49}$$

16.4.8 Signatures of Cloud Collapse

We know that stars form in molecular clouds, because of the close spatial and physical associations. The details of the process are not well determined, so an active field of study to the determination of collapse criteria, and the search for objects that are collapsing. Models for collapse begin with sizes of 10^{16} cm, continuing to pre-planetary disk sizes 10^{14} cm around the resulting star. At present, mm and sub-mm instrumentation limits us to nearby objects which could be imaged on scales $<10^{16}$ cm. After many searches, a number of candidate objects have been found. These are mostly in dark dust clouds where low mass stars are expected to form. The collapse signature is based on a model where the cloud center has higher temperature and density, and where the collapse motions exceed the thermal motions. The measurements of optically thick lines, such as carbon monosulfide, $^{12}C^{32}S$ will show a narrow absorption feature (from gas in the outer part of region). This locates the center of the emission profile. If the cloud is collapsing, the blue shifted portion of this profile has a larger intensity than the red-shifted portion, since the radiation from hotter blue- shifted gas passes through the red-shifted gas without being absorbed. The red-shifted radiation from equally hot gas is absorbed by cooler gas at nearly the same velocity. Thus the blue wing arises closer to the warmer gas. If the cloud is expanding, the red-shifted portion is more intense. A measurement of the rarer isotope $^{13}C^{32}S$ must show no narrow absorption feature or asymmetry. After a large number of attempts, a few well established cases of collapsing clouds have now been found (Evans 1999). The next steps are to establish the chemistry of pre-collapse objects and the details of chemistry and dynamics in disks around young stars.

16.5 A Selected Sample of Results

In Table 16.2, we give a classification of molecular regions together with some of their properties. The diffuse clouds are studied mostly by their absorption lines toward bright background stars. UV photons from the ISRF can penetrate diffuse clouds and rapidly destroy most molecules, so the fraction of atomic species is high and complex species are rare. The translucent clouds form a bridge to the dark molecular clouds, with higher values of density and lower values of T_k. Photoprocesses play a large role in the outer parts of these clouds, but much less in the centers. These clouds are found as IRAS $100\,\mu$m cirrus emission and also in CO lines. Little or no star formation has occurred in these, perhaps due to the high fractional ionization. The cold, dark molecular clouds show complex morphologies. These are

Table 16.2 Physical Characteristics of Molecular Regions in the ISM [after van Dishoeck et al. (1993)]

	Density (cm^{-3})	T (K)	Mass M$_\odot$	A$_v$ mag	Size pc	ΔV km s^{-1}	Example
Diffuse clouds	100–800	30–80	1–100	$\approx$1	1–5	0.5–3	ζ Oph
Transluc. clouds	500–5000	15–50	3–100	1–5	0.5–5		Hi-lat-cl
Cold dark clouds							
Complex	10^2–10^3	>10	10^3–10^4	1–2	6–20	1–3	Taurus
clouds	10^2–10^4	>10	10–10^3	2–5	0.2–4	0.5–1.5	NGC 1333
cores	10^4–10^5	$\approx$10	0.3–10	5–25	0.05–0.4	0.2–0.4	TMC-1
GMC complex	10^2–10^3	15–20	(1–30)10^5	1–2	20–80	6–15	M17, Orion
clouds	10^2–10^4	>20	10^3–10^5	>2	3–20	3–12	Ori. OMC–1
warm clumps	10^4–10^7	25–70	1–10^3	5–1000	0.05–3	1–3	M17 clumps
hot cores	10^7–10^9	100–200	10–10^3	50–1000	0.05–1	1–10	Ori.Hot Core

the birthplaces of low mass stars (i.e. < 2 M$_\odot$) and extend over tens of parsec; in Taurus this is several square degrees of sky. In some parts of these clouds, complex molecules have been detected (see the discussion of TMC 1). The Giant Molecular Clouds (GMC's) have a similarly complex morphology but are warmer and more massive than cold dark clouds. GMC's have the same average density as cold dark clouds, but GMC clumps can have densities n(H$_2$) $\approx 10^6$ cm^{-3}. GMC cores are also the sites of massive star formation, in addition to low mass star formation. It is believed that dark clouds and GMC's are close to virial equilibrium, so self gravity is important in determining the structure and evolution. Diffuse and translucent clouds are not in virial equilibrium, but probably in pressure balance. There is no consensus on ages, with values from 10^7 to 10^8 years in the literature. The cores may have much shorter lifetimes. If star formation has occurred, perhaps 10^4 to 10^5 years; cores without stars may have lifetimes of 10 times longer. From the star formation rate, cores must be supported against collapse for many free fall times.

16.6 Chemistry

In the following, we treat two examples of molecular clouds to illustrate the wide variety of chemistries.

In Table 16.3, we give a list of the known interstellar molecules ordered by the number of atoms. Of the interstellar molecular species identified so far, 90 contain carbon, and are considered organic. From this table, one notes first, that many of the molecules are organic. The remaining 40, mostly diatomic species, belong to what is commonly called inorganic chemistry. Aside from H$_2$ itself, the most widespread inorganic species are OH, NH$_3$ and SiO and H$_2$O. Nearly all of the complex molecular species contain carbon. Among these are formaldehyde, H$_2$CO, cyanoacetylene, HCCCN, formamide, NH$_2$CHO and ethyl alcohol, CH$_3$CH$_2$OH.

Table 16.3 Gas-phase interstellar and circumstellar molecules [a]

(2)	(2)	(3)–(4)	(5)	(6)	(8)
H_2	PO	N_2O	CH_4	C_4H_2	H_2C_6
CH	SO	SO_2	SiH_4	H_2C_4	C_6H_2
CH^+	SO^+	SiCN	H_2COH^+	HC_2CHO	C_7H
NH	FeO	SiNC	CH_2NH	$c\text{-}C_3H_2O$	(9)–(10)
OH	(3)	AlNC	H_2C_3	HC_3NH^+	CH_3CHCH_2
SH	H_3^+	MgCN	$c\text{-}C_3H_2$	C_5N	CH_3OCH_3
C_2	CH_2	MgNC	CH_2CN	HC_4N	CH_3CONH_2
CN	NH_2	NaCN	NH_2CN		CH_3C_4H
CO	H_2O	CH_3	CH_2CO	(7)	C_8H
CO^+	H_2S	NH_3	HCOOH	CH_3CHO	C_8H^-
CF^+	CCH	H_3O^+	C_4H	CH_3NH_2	HC_7N
CP	HCN	H_2CO	C_4H^-	CH_3CCH	C_3H_6
CS	HNC	HCCH	HC_3N	C_2H_3OH	C_2H_5CHO
HF	HCO^+	H_2CN	HC_2NC	$c\text{-}CH_2OCH_2$	CH_3COCH_3
NO	HOC^+	$HCNH^+$	HNCCC	C_2H_3CN	$HOCH_2CH_2OH$
PN	HCO	H_2CS	C_5	HC_5N	CH_3C_5N
NS	HN_2^+	C_3H	C_4Si	C_6H	(11)
AlF	HCP	$c\text{-}C_3H$	CNCHO	C_6H^-	CH_3C_6H
AlCl	HNO	HCCN	CH_2CHO	(8)	HC_9N
NaCl	HCS^+	HNCO	C_2H_4	C_2H_6	(12)
KCl	C_3	$HOCO^+$	CH_3OH	$HCOOCH_3$	C_6H_6
SiC	C_2O	HNCS	CH_3SH	CH_3COOH	(13)
SiN	C_2S	C_3N	CH_3CN	$HOCH_2CHO$	$HC_{11}N$
SiO	$c\text{-}C_2Si$	C_3O	CH_3NC	C_2H_3CHO	
SiS	CO_2	C_3S	CH_2CNH	CH_3C_3N	
N_2?	OCS	$c\text{-}SiC_3$	NH_2CHO	CH_2CCHCN	
O_2	CCP	C_3N^-	C_5H	NH_2CH_2CN	

[a] "c" stands for cyclic species; "?" stands for ambiguous detections; isotopomers excluded. Table provided by E. Herbst and T. Millar

Only a few ring molecules, marked with a prefix "c", have been detected so far, although extensive searches have been conducted. It is not clear yet whether this indicates a true deficiency of ring molecules, since the molecular structures of ring molecules are complex with a large number of energy levels. In addition, some may have no permanent dipole moments. Thus even at low interstellar temperatures, the population in any one level is low since a large number of energy levels are populated. In radio astronomy, usually only one transition is measured at a time, so this discriminates against the detection of complex species which lack internal symmetry. Alternatively, the excitation mechanisms for the lines searched for so far may be unfavorable. At present, the abundance of molecules without permanent dipole moments, such as CH_4 or C_2H_2, can be determined to only a limited extent. For polar molecules, one can use the results of detailed abundance analyses (presented earlier in this chapter) to estimate the relative abundance of molecules in different classes of sources. In Fig. 16.8, we show a plot of relative abundances for the extended molecular region in Orion (a warm dense core, with $T_K \geq 60\,K$), TMC-1S (a cold moderately dense core, with $T_K = 10\,K$) and the envelope of a carbon star IRC+10216.

16.6.1   Clouds for which the UV Field can be Neglected

Even in very dense, hot neutral regions such as the Orion Hot Core, three body collisions hardly ever occur. In the more typical regions where the H_2 density is $\simeq 10^2 \, cm^{-3}$, and $T_K \simeq 20 \, K$, collisions between two neutral partners rarely lead to chemical reactions since there is an activation energy. Rather, as was found in the 1970s, gas phase production of more complex molecules commonly found in the ISM requires a nonstandard chemistry involving ions [for a short history, see Herbst (1999, 2001)].

16.6.2   Models of Photon Dominated Regions

One finds high intensity, extended emission in various CO rotational lines and atomic fine structure lines (Table 12.1) toward a variety of sources such as star burst galaxies, molecular clouds near giant HII regions, Planetary Nebulae and even disks around young stars (see Hollenbach and Tielens 1999). The Orion region has good examples of PDRs. These arise from the interface between the HII region and molecular cloud in a Photon Dominated Region (PDR). One of the most well studied PDRs is the 'Orion Bar', to the SE of the Orion H II region. Here, the radiation field is $\sim 10^5$ times larger than found near the Sun, and the proton density is $\sim 10^5$ cm^{-3}. In PDRs, the molecules are close to a source of far UV (energies between 6 eV and 13.6 eV) radiation from the O stars which give rise to the HII region. This UV field alters the chemistry and heat balance. The most obvious change is that CII and CI become much more abundant. The layer containing CI may extend to a depth which is a significant fraction of the width of the PDR.

   PDR models are complex since in the PDR scheme one must account both for the heating and chemistry of the region. The present view is that the heating in the PDR is caused by the absorption of stellar UV by dust grains, with the photoejection of electrons. To reach the high temperatures observed, small grains must play a large role; the details of this heating process are a topic of current research. The models are characterized by G_0, the ratio of the radiation field to that found near the Sun, and n, the proton density. In a plane parallel geometry, from the HII region toward the molecular cloud interior, one finds thin layers of H_2, then CII, CI and then CO (see Störzer et al. 1996 for spherical geometry). The CO line emission from a PDR arises from a layer corresponding to A_v=1–2^m, or 2–4 $10^{21} \, cm^{-2}$ protons. For a time-stationary PDR, this layer must be rather dense ($n(H_2) \approx 10^5 \, cm^{-3}$), so that H_2 formation times are fast compared to dynamical times. There is a stratification in a PDR, so this is not a uniform region. Closer to the UV source, OI and CII provide most of the cooling. Deeper into the molecular region, CO rotational lines contribute more to the cooling. In Fig. 16.7 we show a set of ratios for far IR fine structure and CO rotational lines. For these plots, it is assumed that there is a single PDR, more extended than the telescope beam. As can be seen, in CO emission from high density PDRs, higher rotational lines have larger peak
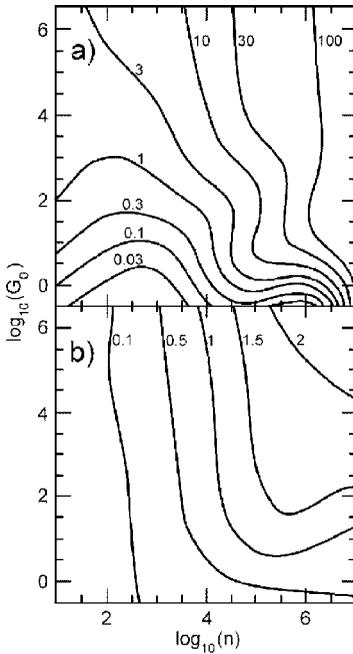
Fig. 16.7 Plots of the intensity ratios of the fine structure lines of OI and CII, and the ratio of the rotational lines of CO as a function of the UV radiation field (in terms of G_0 the field near the Sun) and the proton density, n. In (**a**) we plot the ratio of OI to CII, in (**b**) the ratio of the $J = 3 - 2$ to $J = 1 - 0$ lines of CO are shown. [Adapted from Kaufman et al. (1999)]

temperatures; from the ratios of CO lines one can also characterize conditions in the PDR.

From more detailed studies has it become necessary to introduce high density condensations in the PDRs. This inclusion of such *clumps* complicates the analysis, but clumping is needed to explain sub-mm CO line intensities.

At the surface of clouds, ions can be produced by radiation. However, the radiation would also destroy the molecules. Deeper inside the clouds, the molecules can survive, and there the ions are produced by cosmic rays which can penetrate into the interior of a molecular cloud.

16.6.3 Results

We have plotted selected observational abundances in Fig. 16.8. These are corrected for source and beam size effects. The typical error bars are shown for IRC+10216. These apply to all sources plotted. We restrict ourselves to the abundances of the more commonly found species. For CO, the abundance is about 10^{-4}, which 30% of the carbon to H ratio. For IRC+10216, the ratio may be higher, but this is a star in which carbon is overabundant compared to the usual ISM ratio . It is believed that at
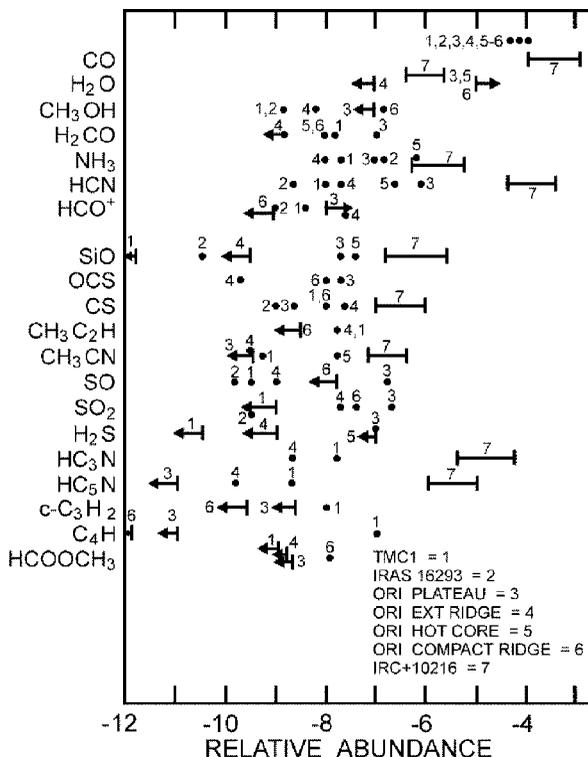
Fig. 16.8 A plot of the relative abundances of molecular species for the sources TMC-1, representing dust clouds, IRC+10216, representing carbon star envelopes, and Orion KL, representing hot dense molecular cores and IRAS 16293-2422, representing low mass star forming regions [adapted from van Dishoeck and Blake (1998), van Dishoeck et al. (1993) with IRC+10216 data from Irvine et al. (1985) and some additional data from Langer et al. (2002)]

the very highest densities all molecules (except H_2, H_3^+ and H_2H^+) are condensed onto grains. Uncertainties are thought to be <2; for other species, the error in the total abundances is likely to be ± 3, and the abundance relative to H_2 must be even larger.

In most cases, the abundances are averaged over the source, not the telescope beam. That is, corrections for source size have been applied. In most cases, this requires high angular resolution imaging. Except for TMC-1S (the HC_xN, $x = 2n + 1$, with $n = 1, 2 \ldots$ abound in Taurus) such images exist for some transitions of the molecules. From top to bottom in the table, one finds a small scatter in the CO abundance, but a large scatter for water vapor. One has to bear in mind that determinations of the H_2O abundance are made difficult since the earth's atmosphere is not transparent at the line frequencies, and that many transitions of H_2O show maser emission. From SWAS (sub-mm water astronomy satellite) data, the relative abundance of H_2O is $\approx 10^{-5}$ in warm sources, but orders of magnitude smaller in cold sources. H_2O is produced in the gas phase at high T_K, but may also be evaporated from grains.

Methanol, CH_3OH and methyl formate, $HCOOCH_3$, are 10–100 times more abundant in the Orion Hot Core than elsewhere. This is an indication that these species are favored in "Hot Core chemistry". One surmises that these are formed on and evaporated from grains when a protostar evolves into a Young Stellar Object (YSO). Both species are asymmetric top molecules, so have many energy levels. Thus even if the lines are weak there is no assurance that the total abundance is low.

Formaldehyde, H_2CO, ammonia, NH_3, and hydrogen cyanide, HCN, also show a large abundance spread. In the case of H_2CO and NH_3, there are indications that in warmer source these are produced on grains and evaporated. There must be additional mechanisms to produce these, and HCN, in regions where T_{kin}=10 K. The formyl ion, HCO^+, and protonated molecular nitrogen, N_2H^+ are produced in the gas phase. At present, we believe that these species exist in the gas phase even at the highest densities. The spread in the HCO^+ and N_2H^+ abundances gives an indication of the effect of local conditions on the abundance. The silicon monoxide molecule, SiO, is found only where strong shock waves are present. The generally accepted scenario is that grains, which consist of Si and carbon, are broken up by the shocks, and the Si is set free. SiO can form only at high temperatures, but the shocks also raise the gas temperature, so this is also provided by shock waves. Carbon monosulfide, CS, is found in cold clouds, so this is very likely produced in the gas phase. In the quiescent TMC-1S region and in the Orion extended ridge, methyl cyanide, CH_3CN, has a much lower abundance than methyl acetylene, CH_3C_2H, which reflects a very different chemistry. This may indicate that CH_3C_2H is a gas phase product. It is thought that sulfur monoxide, SO, sulfur dioxide, SO_2, and hydrogen sulfide, H_2S, are produced only a higher temperatures, so probably in shocks. The long chain polyenes at the bottom of the table are produced in gas phase chemistry.

Some of the molecules are commonly found on earth; examples are NH_3, HCN and NaCl. However, others, such as OH, CN, CO^+, CH^+, HCO^+, and N_2H^+ are chemically unstable even under laboratory conditions and will quickly combine to form other, chemically stable species. An immediate question is how such species are formed. As will be discussed in the next section, the observed abundance of even diatomic molecules requires a chemistry which is different from that normally encountered on earth.

16.6.4 Ion-Molecule Chemistry

In cold clouds prior to star formation the chemistry is dominated by low temperature gas phase ion molecule and neutral-neutral reactions. These lead to small radicals and unsaturated molecules. Ion-molecule chemistry is very successful in explaining many aspects of this chemistry. At typical densities, neutral-neutral chemical reactions between atoms are simply too slow to form even triatomic molecules in a few 10^6 years. The solution to this problem involves ion-molecule chemistry. The ions, mostly hydrogen but about 10% helium by number, are produced by cosmic rays. The hydrogen ions either form H_3^+, or charge exchange with atoms such as

carbon or oxygen. The usual reaction rates for ion molecule reactions are thousands of times faster than neutral reactions, and this satisfactorily explains the abundance of diatomic molecules such as CO or OH in dark clouds. Ion-molecule reactions also play a large part in the formation of molecules in the envelopes of low-mass stars in the red giant phase.

As far as we know, with one exception, simpler species are produced in the gas phase by reactions of ions and molecules. The exception is the most abundant molecule, H_2. Because H_2 is a homopolar molecule, that is, without a permanent dipole moment, the formation via collisions of two H atoms will occur very rarely. This is because of the need to conserve both energy and momentum; the radiation process is simply too slow to allow the excess energy to escape. For this reason, the formation of H_2 must occur on dust grain surfaces. It has been proposed that at densities $\geq 10^{10}$ cm^{-3}, three body reactions in the gas phase could occur, but such high densities are thought to occur only rarely, perhaps in early universe before grains were formed.

Gas phase chemistry consists of 3 types: carbon insertion ($C^+ + CH_4 \rightarrow C_2 + H_2$ H_2^+), condensation reactions ($CH_3^+ + CH_4 \rightarrow C_2H_5^+ + H_2$), and radiative association ($C^+ + C_n \rightarrow C_{n+1}^+ + h\nu$). The build up of polyatomic molecules is limited by photodissociation at the edges of clouds and near young stars. At larger A_V, reactions with O and C^+ tie up atoms in CO, which is very stable. At higher kinetic temperatures, T_k, these reactions lead to H_2O. Element abundances play a role, especially C versus O. In the ISM, one finds that C, N and O are $\approx 25\%$ below the solar system values. Dust must contain nearly 100% of the Si and Fe, and $\approx 25\%$ of the O, and 60% of the C. Dust contains a few % of N. Gas phase O_2 abundance is small, with 40% of the O in the form of OI. In cold clouds, the H_2O abundance is $\approx 10^{-7}$ to 10^{-8}. In warmer clouds the water abundance is much larger. Most of the O is in dust grains or condensed as ices. In the gas phase, oxygen is in OI and CO, and perhaps H_2O.

In warmer, denser clouds, the chemistry is more complex, since it is possible that certain complex molecules are formed on the surfaces of dust grains, and are then liberated. For grain surface chemistry, the reaction sites and reaction mechanisms may not be well determined. Even for the rather well-understood ion-molecule gas phase chemistry, there are many possible reactions leading to an observed product. This is even the case for an abundant species such as CO. Also, reaction rates for low temperatures are not measured in many cases. Usually, reactions involving ions are not very temperature sensitive, but in any production path, there may be reactions involving neutrals and these *are* very temperature dependent. (See the web site www.rate99.co.uk for a database.)

We now consider a simple example of gas phase production of molecules. This is meant to serve as an example for other schemes which have a larger number of production pathways. An excellent example of ion-molecule chemistry is the case of HCO^+. In the following, we show one of the production and destruction paths for this molecule in well shielded regions where the interstellar radiation field, (ISRF) plays only a minor role. This example illustrates many of the general principles of the preceding discussion and shows how laboratory and astronomical data can be

combined to yield results which are otherwise unreachable. It should be noted that CO, OH and other species are also produced in a similar way, although the processes are somewhat more complex. We assume that the ISRF plays no role. Then, the first reaction is cosmic ray (CR) radiation reacting with H_2:

$$H_2 + CR \rightarrow H^+ + H + e^-, \tag{16.50}$$
$$H^+ + H_2 \rightarrow H_3^+. \tag{16.51}$$
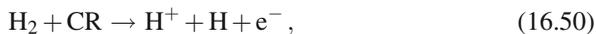
These reactions and similar ones involving ionized helium or H_2^+ lead to H_3^+ rather than to $H^+ + H$. Once H_3^+ has formed, it can react with many partners. For example, one reaction is:

$$H_3^+ + CO \rightarrow HCO^+ + H_2. \tag{16.52}$$

These reactions lead to the production of HCO^+. The destruction of HCO^+ or H_3^+ can occur via recombination with electrons:

$$H_3^+ + e^- \rightarrow H_2 + H, \tag{16.53}$$
$$HCO^+ + e^- \rightarrow H + CO. \tag{16.54}$$

The total cosmic ray ionization rate per H_2 is given by ζ, in units of s^{-1}; the rate per H atom is $\zeta/2$. The destruction of H_3^+ is via reactions with electrons, either from those resulting from the formation of H_3^+, with a rate k_e, or those resulting from other ions, with a rate k_i. When formation and destruction are in equilibrium, we arrive at a steady state relation for the abundance of HCO^+ and H_3^+ in a cloud with a local H_2 density $n(H_2)$. For the reaction rate in (16.53) we use k_e for electrons from H_3^+ and k_i for electrons from metals. Similarly for (16.54) we use k_{HCO^+} for the reaction of HCO^+ with electrons from the ion itself, and k_i' for the reaction charge exchange with metals. Balancing the creation and destruction rates of H_3^+ we have

$$k_e\, n_{H_3^+}\, n_e + k_i\, n_{H_3^+}\, n_{X_i} = \zeta\, n_{H_2}. \tag{16.55}$$

For HCO^+ the reaction balance is

$$k_{HCO^+}\, n_{(HCO^+)}\, n_e + k_i'\, n_{(HCO^+)}\, n_{X_i} = k\, n_{(H_3^+)}\, n_{CO}. \tag{16.56}$$

Dividing by the local density of H_2 in these two relations, we obtain the concentrations of the species involved. We use square brackets to indicate concentrations, then:

$$[H_3^+]\,(k_e[e] + k_i[X_i]) = \zeta/n_{H_2}. \tag{16.57}$$

For the abundance of HCO^+, the steady state result is:

$$[HCO^+]\,(k_{HCO^+}[e] + k_i'[X_i]) = k\,[H_3^+]\,[CO]. \tag{16.58}$$

Combining these two relations, we can solve for the abundance of HCO^+ in terms of the cosmic ray rate, ζ, the abundance of CO and the ionization fraction in

a cloud:

$$[HCO^+] = \frac{k\zeta[CO]/n(H_2)}{(k_{HCO^+}[e] + k_i'[X_i])(k_e[e] + k_i[X_i])} \tag{16.59}$$

where k is the ion-molecule rate which H_3^+ reacts with CO to form HCO^+; this is a measured value. If we take $\sum X_i \ll 10^{-4}$, and define $R = n(H_2)n(HCO^+)/n(CO)$, then (16.59) reduces to:

$$R \le \frac{k\zeta}{(k_{HCO}^+[e])(k_e[e])} \tag{16.60}$$

or, solving for $[e]$,

$$[e] \le \sqrt{\frac{\zeta k}{R k_{HCO^+} k_e}}. \tag{16.61}$$

From the abundances of HCO^+, CO and the cosmic ray rates, one can obtain the electron abundances. We take $k = 10^{-9}$, $k_e = 4 \times 10^{-5} T^{-0.5}$ cm^{-3} s^{-1}, and $k_{HCO^+} = 6 \times 10^{-6} T^{-0.5}$ cm^{-3} s^{-1}. For molecular clouds, the typical values are $T_K = 10$ K, $R \approx 1$, $\zeta = 4 \times 10^{-17}$ s^{-1}. From these values, an *upper* limit to the electron fraction is 10^{-8}. See our discussion of magnetic support in Sect. 14.10.4.

As one might expect, the real situation is not so simple, because of: (1) the inhomogeneous structure of clouds, (2) possible non-LTE excitation and (3) uncertain reaction rates for H_3^+ or other reactions.

The inhomogeneous structure of molecular clouds may have an effect because their envelopes will have a larger abundance of electrons due to ionization of metals with potentials below the ionization energy of H (13.6 eV) by the ISRF. In the cloud interior these are less abundant because of the extensive depletion expected at high densities, and shielding by dust. Thus HCO^+ may be present mainly in the outer parts of clouds, and one cannot simply combine the CO abundance for the whole cloud with HCO^+ results. The excitation will play a role since HCO^+ requires a high H_2 density for thermalization (see Sect. 14.6). Finally, the calculated electron abundances depend on how fast H_3^+ exchanges charge.

The HCO^+ scheme makes it clear that the gas phase abundance of any species in a molecular cloud is a delicate balance between the production and destruction processes. The production process depends on the type of chemistry; either ion-molecule, neutral-neutral or grain surface reaction. These in turn depend on T_K, as well as the abundance of relevant ions and constituents and their form, either molecular or atomic. The destruction processes can be either chemical, that is, processing into more complex species, or physical, that is, the freezing out onto grains, or destruction by the ISRF or cosmic rays. All of these are dynamic processes, which can change on time scales of 10^3 years, much shorter than the lifetime of molecular clouds.

At present, it seems clear that for the vast majority of clouds in most circumstances, the most important gas phase chemical production mechanisms have been identified, but the quantitative reaction rates at low T_K have not been settled in many cases. The question of which chemical paths are most important must be settled

on an individual basis, from laboratory measurements. One sweeping philosophical conclusion is that interstellar chemistry has shown us that standard processes known from laboratory chemistry are only a special case of many possible chemistries.

16.6.5   Grain Chemistry

If the cloud begins to collapse, the densities rise and for a while the kinetic temperature falls. One result will be that a significant fraction of gas phase molecules condense on icy mantles. High abundances of some species and detection of ices demand gas-grain, or surface chemistry. These contain important information on the temperature and irradiation history of the region. There are two regimes for the surface chemistry: In the first, the mobile species moves over the surface faster than the accretion time of the reactant. In the second, the accretion time is faster. Most models of surface chemistry use the second approach, although the first approach can be analyzed via a Monte-Carlo method. After the reaction, the species must return to the gas phase. The dust temperature must reach 20 K for thermal evaporation of CO. Other mechanisms are cosmic ray spot heating, or exothermic heating via chemical reactions, grain-grain collisions. The efficiencies are dependent on the binding energies of the molecules on the grain surfaces.

Grain chemistry can be modified by surface reactions and through processing by UV, X rays from nearby stars and cosmic rays. After the new star has formed its radiation heats up the surrounding gas and dust and the molecules begin to evaporate back into the gas phase with the most volatile returned first. In addition, outflows from the young stars penetrate the surrounding envelope creating high temperature shocks and lower temperature turbulent regions. In which the grain mantles are returned to the gas. Finally the envelope is dispersed by winds and in the case of massive stars, UV photons leading the appearance of photon dominated regions (PDR).

In connection with present-day gas-grain models, one must bear in mind that these have dependences on time. With enough data, one could determine the age of the source. However, given the incomplete state of our knowledge of sources, time-dependence is likely to be used as a free parameter to determine source age.

16.6.6   Searches for New Molecules

Initially, molecular identifications were usually based on one transition. Today, this is done only in exceptional cases for species expected to be abundant in certain regions. Recent examples are D_2O and H_2D^+. For less abundant asymmetric top molecules without characteristic hyperfine signatures, there is a good chance of line confusion. For example, in the frequency range between 100 and 110 GHz there are 362 lines, of which 113 are unidentified ("U") lines

(Lovas (1992)). The weakest lines detected have peak intensities at the 20 milli K level. For a survey in the 325-60 GHz interval, Schilke et al. (1997) can identify

94% of the lines found. There are 29 lines per GHz. Many of the identified lines are due to transitions from the ground or vibrationally excited states of abundant species such as methanol, methyl formate or dimethyl ether. It would appear that the chance of finding new species is better in the band around 100 GHz, where there are fewer intense lines from known species, and where one can obtain sensitive spectra. However, even at 100 GHz, there are so many spectral lines that the chance of line overlap is large. The recent discovery of new complex species such as vinyl alcohol, (CH_2CHOH), glycolaldehyde (CH_2OHCHO, the first sugar) and ethylene glycol ($HOCH_2OH$; see Hollis et al. (2002)) makes it more likely that glycine (NH_2CH_2COOH), the simplest amino acid, may be present at some level. (Amino acids have been found in meteorites; in some proposals life may have been brought to the early earth from outer space). The discoveries of complex organic molecules show that we have not yet reached the ultimate limit. However, it is becoming more likely that an increase in sensitivity might not lead to the identification of new species in the classical sources such as Orion or Sgr B2. The most direct method (see, e.g., Chap. 9), is to increase the angular resolution using interferometry. This is effective since different molecular species are usually found in slightly different spatial locations. In addition, one could conduct searches in sources with narrower linewidths.

We conclude that in order to identify a particular molecule, there must be: (1) good agreement between the astronomical frequencies of at least four, and preferably more transitions known to an accuracy of better than 100 kHz, and (2) a common exitation mechanism.

For such accuracies, there must be laboratory frequencies. For stable species, such results are available, as are molecular structures (see e.g. Harmony et al. (1978)). These are usually based on a series of measurements in the millimeter wavelength region. When combined with a model of the molecular structure one can calculate the frequencies of all allowed transitions. The measured frequencies are usually accurate to a few kHz, the model dependent frequencies to better than a MHz. For unstable molecules, such as ions, laboratory frequencies are more of a problem. For these species, one must measure the frequencies before destruction by reactions in the gas phase or contact with walls. To carry out such measurements, one usually produces the desired species in a volume much larger and a density lower than is normally used in microwave spectroscopy, and then proceeds to measure the frequencies quickly. P. Thaddeus and coworkers (see Gordon et al. (2002)) have produced a variety of long carbon chains. Searches for such species can be made in IRC+10216 or TMC 1, where the lines are narrower and density of lines is smaller.

In order to have a certain detection of a new species, it may be necessary to escape from the problem of confusion of those lines from the molecule sought with lines from other molecules. In addition, the intensities of the lines for a given species should yield a common rotational temperature. That is, one expects that the excitation follows a simple model. Deviations from this rule may indicate that some of the lines assigned to this species are the result of misidentifications.

Problems

1. For CH_3CN, CH_3C_2H and NH_3 there can be no radiative transitions between different K ladders. The populations can however be exchanged via collisions. For ammonia, there must be $J > K$. There is a rapid decay of populations with quantum numbers from $(J+N+1,K)$ to $(J+N,K)$, where N is ≥ 1. Use the relation in Eq. (15.43) to show that rotational transitions of NH_3 fall in the frequency range $\geq 500\,\mathrm{GHz}$. Estimate the Einstein A coefficients for the $J = 1 - 0$ and $J = 2 - 1$ transitions using $\mu = 1.34$ Debye. Compare these values to those for the inversion transitions listed in Table 16.2, which are $\sim 10^{-7}\,\mathrm{s}^{-1}$.

2.(a) Calculate the excitation temperature, T_{ex}, between two energy levels which have the same statistical weights, that is $g_u = g_l$, so that the Boltzmann equation is $n_u/n_l = e^{-h\nu/kT_{ex}}$ with $h\nu/k = 1.14$ K. The values of n_u/n_l are 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5.
(b) Use the relation between optical depth and column density from Problem 13 to calculate the optical depth, τ, for the $J = 1 - 0$ line which has a FWHP of $10\,\mathrm{km\,s}^{-1}$, $T_{ex} = -100$ K, $\mu_0 = 3.6$ Debye, and $\nu = 9.0\,\mathrm{GHz}$.
(c) Substitute this value of τ into the relation $T_{MB} = (T_{ex} - T_{BG})(1 - e^{-\tau})$. If $|T_{BG}| \ll |T_{ex}|$ and $|\tau| \ll 1$, show that T_{MB} gives an accurate estimate of the column density in the lower level, N_0. Aside from questions of English usage, would you agree with the statement "Optically thin masers do not mase"?
(d) Evaluate the other extreme case, $T_{BG} \gg T_{ex}$, to show that the background radiation is amplified.

3. Use the large velocity gradient (LVG) relation for a two-level system (Eq. 16.40) to estimate the line temperature when $T_K \gg T_0, A \gg C$. In addition, $(A_{ji})/(3C_{ji}\tau_{ij}) \ll 1$. This is a hot, subthermally excited transition.

4. Repeat the above exercise for the case in which $A \ll C$, but with all other parameters unchanged. This is the case of a hot, thermalized gas. Compare these results with those of Problem 20.

5. In circumstellar envelopes, one assumes that spherical symmetry holds, and that density $n(r) = n_0 r^{-2}$. In addition, $r = (z^2 + p^2)^{1/2}$, where p is the projected distance and z the line-of-sight distance, and a constant velocity of expansion.
(a) Show that

$$\delta z = \Delta V/(dv_\parallel/dz) = p\frac{\Delta V}{V}(1 - (v_\parallel/V)^2)^{-3/2} .$$

(b) Take the abundance of a species to be a constant fraction of the abundance of H_2. Show that the optical depth for a given species at point p and for a given $v_\parallel$ is

$$\tau(p, v_\parallel) = \frac{\mu^2 f n_0 (J+1)}{1.67 \times 10^{14}\,T_{ex}} \frac{p}{V(2J+1)}(1 - (v_\parallel/V)^2)^{-1/2} .$$

(c) Take the beam to be much larger than the source. Then show that

$$T = 2\pi \int_0^{p_{max}} T_0(1 - e^{-\tau}) p \, dp \,.$$

(d) Assume that the line is optically thin. Show that the line profile is flat–topped. Then assume that the line is optically thick. Show that the profile is parabolic shaped.

6*. Bipolar outflows are common in pre-main sequence sources. This is a very elementary analysis of molecular line emission from well-defined bipolar outflows.
(a) Approximate the outflow as a cylinder of length l, width w, with constant density n, inclined at an angle i to the line of sight. Show that the functional description of the mass of the outflowing material is $(1/4)n(H_2)\pi l w^2$.
(b) If the observed velocity of the outflow is v_o, show that the age of the outflow is

$$\text{age} = l_o/(v_o \tan i) \,.$$

(c) Show that the total kinetic energy in the outflow is $(1/2)Mv_o^2 \sin i/\cos^2 i$.
(d) If we define the mechanical luminosity L as $\dot{E} = (2\times\text{kinetic energy in the outflow})/\text{age}$, show that $L = M(M_\odot) v^3/(\sin i \cos^3 i)$, where M is the mass of the outflow.

7*. For linear molecules, in principle one can determine both the kinetic temperature and the H_2 density if one can measure the "turn over" in the distribution of column densities from different transitions. One example is given by measurements of the CO molecule in Orion KL.
(a) Estimate the wavelengths, frequencies and Einstein A coefficients for the $J = 30 - 29$, $J = 16 - 15$ and $J = 6 - 5$ transitions, if CO is a rigid rotor molecule. Compare these to the value for the $J = 2 - 1$ transition. If the lines are optically thin, and $\langle \sigma v \rangle = 10^{-10}$ cm^3 s^{-1}, what are the critical densities?
(b) Determine the energies of the $J = 30$, $J = 16$ and $J = 6$ levels above the ground state. If the kinetic temperature of this outflow region is $\sim$2000 K, find the ratio of populations of the $J = 30$ to $J = 6$ levels, assuming LTE conditions. If the H_2 density, n, is $\sim$10^6 cm^{-3}, set A equal to the collision rate, $C = n \langle \sigma v \rangle$, to determine which of the transitions is sub-thermally excited, i.e. $A \gg C$.

8. The quantity $\sigma_g n_g$ equals $10^{-22} n_H$ cm^{-1}. The mean free path, λ, is equal to $1/\sigma_g n_g$. If the mean time between collisions, $t_{gas-grain} = \lambda/V$, where the expression for V is taken to be $\Delta V_{1/2}$.

9. From (Eq. 16.47), the free-fall time in years for a cloud under the influence of self gravity is $t_{ff} = 5 \times 10^7/\sqrt{n(H_2)}$, where $n(H_2)$ is the molecular hydrogen density in cm^{-3}. From this result and the result in the previous problem, find the density at which the free-fall time equals the average time for a molecule to strike a grain.

10. A typical giant molecular cloud (GMC) is thought to have a diameter of 30 pc, and total mass of $10^6 M_\odot$. Assume that GMC's have no small scale structure.
(a) Develop a general formula relating the H_2 density to the mass and radius of a uniform spherical cloud. Because the He/H number ratio is 0.1, the average molecular mass is 4.54×10^{-24} g.

(b) What is the density of the GMC? Find the column density of H_2 in this cloud. If the visual extinction is related to column density by $1^m = 10^{21}$ cm^{-2}, what is the extinction through the GMC?

(c) What is the FWHP width of a line if the cloud is in virial equilibrium? Use the simplest condition for virial equilibrium, as given in (Eq. 13.72).

(d) If the mass of the ISM in the galaxy from 2 kpc to 8.5 kpc is $3 \times 10^9 \, M_\odot$, and if there are $\sim$100 GMCs as described in part **(a)** in this part galaxy, how much of the total mass of the interstellar medium is in GMCs? If the thickness of the galaxy is 200 pc, how much of the volume is contained in GMCs?

(e) What is the H_2 column density through a GMC? If one visual magnitude is equivalent to a column density of 10^{21} cm^{-2} of H_2, what is A_v of a GMC?

11. (a) A well-established ion exchange reaction in diffuse molecular clouds is

$$H_2 + D^+ = HD + H^+ + \Delta E, \qquad\qquad (16.62)$$
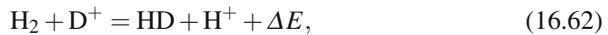
where the zero point energy difference between H_2 and HD is $\Delta E/k = 500$ K. It is thought that reaction (16.62) reaches equilibrium. Then one can relate the initial and final products by the Boltzmann relation

$$\frac{HD}{H_2} = \frac{D^+}{H^+} e^{\Delta E/kT} \ .$$

If the relevant temperature, T, is $T_k = 100$ K, what is the overabundance of HD?

(b) A similar reaction to that given in part **(a)** occurs for isotopes of carbon monoxide (CO) if the carbon ion, C^+, is present in the outer parts of molecular clouds:

$$^{12}CO + \, ^{13}C^+ = \, ^{13}CO + \, ^{12}C^+ + \Delta E \ .$$
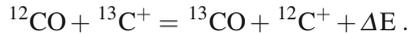
In this case, $\Delta E/k = 35$ K. Repeat the steps in part **(a)** for the case of CO.

Appendix A
Some Useful Vector Relations[1]

Let A, B, C, D be arbitrary vector fields assumed to be continuous and differentiable everywhere except at a finite number of points, and let ϕ and ψ be arbitrary scalar fields for which the same assumptions are adopted. If $A \cdot B$ is the scalar product and $A \times B$ the vector product then the following algebraic relations are true:

$$A \cdot (B \times C) = (A \times B) \cdot C = (A, B, C) = (B, C, A)$$
$$= (C, A, B) = -(A, C, B) = -(C, B, A)$$
$$= -(B, A, C), \tag{A.1}$$
$$A \times (B \times C) = (A \cdot C) B - (A \cdot B) C, \tag{A.2}$$
$$A \times (B \times C) + B \times (C \times A) + C \times (A \times B) = 0, \tag{A.3}$$
$$(A \times B) \cdot (C \times B) = A \cdot [B \times (C \times D)]$$
$$= (A \cdot C)(B \cdot D) - (A \cdot D)(B \cdot C), \tag{A.4}$$
$$(A \times B) \times (C \times D) = [(A \times B) \cdot D] C - [(A \times B) \cdot C] D. \tag{A.5}$$

Introducing the gradient of a scalar as $\nabla \phi$, ∇ considered as a differential operator obeys the following identities

$$\mathrm{grad}(\phi \, \psi) = \nabla(\phi \, \psi) = \phi \, \nabla \psi + \psi \, \nabla \phi, \tag{A.6}$$
$$\mathrm{div}(\phi A) = \nabla \cdot (\phi A) = A \cdot \nabla \phi + \phi \, \nabla \cdot A, \tag{A.7}$$
$$\mathrm{curl}(\phi A) = \mathrm{rot}(\phi A) = \nabla \times (\phi A) = \phi \, \nabla \times A - A \times \nabla \phi, \tag{A.8}$$
$$\mathrm{div}(A \times B) = \nabla \cdot (A \times B) = B \cdot (\nabla \times A) - A \cdot (\nabla \times B), \tag{A.9}$$
$$\mathrm{curl}(A \times B) = \mathrm{rot}(A \times B) = \nabla \times (A \times B),$$
$$= A \, (\nabla \cdot B) - B \, (\nabla \cdot A) + (B \cdot \nabla) A - (A \cdot \nabla) B, \tag{A.10}$$
$$\mathrm{grad}(A \cdot B) = \nabla(A \cdot B) = A \times (\nabla \times B) + B \times (\nabla \times A)$$
$$+ (B \cdot \nabla) A + (A \cdot \nabla) B. \tag{A.11}$$

---

[1] Mainly adopted from Panofsky, W., Phillips, M. (1962): *Classical Electricity and Magnetism* (Addison-Wesley, Reading MA).

Table A.1 The Vector components of ∇ in cylindrical and spherical polar coordinates

	Cartesian coord.	Cylindrical coord.	Spherical polar coord.	
Coordinates	x, y, z	r, ϕ, z	r, θ, ϕ	
Orthogonal line elements	dx, dy, dz	$dr, r\,d\phi, dz$	$dr, r\,d\vartheta, r\sin\vartheta\,d\phi$	
Gradient ∇	$\begin{aligned}(\nabla\psi)_x &= d\psi/dx\\(\nabla\psi)_y &= d\psi/dy\\(\nabla\psi)_z &= d\psi/dz\end{aligned}$	$\begin{aligned}(\nabla\psi)_r &= d\psi/dr\\(\nabla\psi)_\varphi &= \frac{1}{r}\frac{d\psi}{d\varphi}\\(\nabla\psi)_z &= d\psi/dz\end{aligned}$	$\begin{aligned}(\nabla\psi)_r &= d\psi/dr\\(\nabla\psi)_\theta &= \frac{1}{r}\frac{d\psi}{d\theta}\\(\nabla\psi)_\varphi &= \frac{1}{r\sin\theta}\frac{d\psi}{d\varphi}\end{aligned}$	(A.24)
Divergence $\nabla\cdot\mathbf{A}$	$\dfrac{\partial A_x}{\partial x} + \dfrac{\partial A_y}{\partial y} + \dfrac{\partial A_z}{\partial z}$	$\dfrac{1}{r}\left(\dfrac{\partial(rA_r)}{\partial r} + \dfrac{\partial A_\varphi}{\partial\varphi}\right) + \dfrac{\partial A_z}{\partial z}$	$\dfrac{1}{r^2}\dfrac{\partial(r^2 A_r)}{\partial r} + \dfrac{1}{r\sin\theta}\left(\dfrac{\partial(\sin\theta A_\theta)}{\partial\theta} + \dfrac{\partial A_\varphi}{\partial\varphi}\right)$	(A.25)
Components of curl $\mathbf{A} = \nabla\times\mathbf{A}$	$(\nabla\times\mathbf{A})_x = \dfrac{\partial A_z}{\partial y} - \dfrac{\partial A_y}{\partial z}$	$(\nabla\times\mathbf{A})_r = \dfrac{1}{r}\dfrac{\partial A_z}{\partial\varphi} - \dfrac{\partial A_\varphi}{\partial z}$	$(\nabla\times\mathbf{A})_r = \dfrac{1}{r\sin\theta}\left(\dfrac{\partial(\sin\theta A_\varphi)}{\partial\theta} - \dfrac{\partial A_\theta}{\partial\varphi}\right)$	
	$(\nabla\times\mathbf{A})_y = \dfrac{\partial A_x}{\partial z} - \dfrac{\partial A_z}{\partial x}$	$(\nabla\times\mathbf{A})_\varphi = \dfrac{\partial A_r}{\partial z} - \dfrac{\partial A_z}{\partial r}$	$(\nabla\times\mathbf{A})_\theta = \dfrac{1}{r\sin\theta}\dfrac{\partial A_r}{\partial\varphi} - \dfrac{1}{r}\dfrac{\partial(rA_\varphi)}{\partial r}$	(A.26)
	$(\nabla\times\mathbf{A})_z = \dfrac{\partial A_y}{\partial x} - \dfrac{\partial A_x}{\partial y}$	$(\nabla\times\mathbf{A})_z = \dfrac{1}{r}\left(\dfrac{\partial(rA_\varphi)}{\partial r} - \dfrac{\partial A_r}{\partial\varphi}\right)$	$(\nabla\times\mathbf{A})_\varphi = \dfrac{1}{r}\left(\dfrac{\partial(rA_\theta)}{\partial r} - \dfrac{\partial A_r}{\partial\theta}\right)$	
Laplacian of ψ div grad $\psi = \nabla^2\psi = \Delta\psi$	$\dfrac{\partial^2\psi}{\partial x^2} + \dfrac{\partial^2\psi}{\partial y^2} + \dfrac{\partial^2\psi}{\partial z^2}$	$\dfrac{1}{r}\dfrac{\partial}{\partial r}\left(r\dfrac{\partial\psi}{\partial r}\right) + \dfrac{1}{r^2}\dfrac{\partial^2\psi}{\partial\varphi^2} + \dfrac{\partial^2\psi}{\partial z^2}$	$\dfrac{1}{r}\dfrac{\partial}{\partial r}\left(r^2\dfrac{\partial\psi}{\partial r}\right) + \dfrac{1}{r^2\sin^2\theta}\left[\dfrac{\partial}{\partial\theta}\left(\sin\theta\dfrac{\partial\psi}{\partial\theta}\right) + \dfrac{1}{\sin\theta}\dfrac{\partial^2\psi}{\partial\varphi^2}\right]$	(A.27)

For ∇ some second-order formulae are useful

$$\nabla^2\phi = \nabla\cdot\nabla\phi = \Delta\phi \tag{A.12}$$
$$\nabla^2\boldsymbol{A} = \nabla(\nabla\cdot\boldsymbol{A}) - \nabla\times(\nabla\times\boldsymbol{A}) \tag{A.13}$$
$$\nabla\times\nabla\phi = 0 \tag{A.14}$$
$$\nabla\cdot(\nabla\times\boldsymbol{A}) = 0. \tag{A.15}$$

Relations for Special Functions. Let $\boldsymbol{r}$ be the radius vector from the origin to the point x, y, z. Then

$$\nabla\cdot\boldsymbol{r} = 3, \tag{A.16}$$
$$\nabla\times\boldsymbol{r} = 0, \tag{A.17}$$
$$\nabla r = \nabla|\boldsymbol{r}| = \boldsymbol{r}/|\boldsymbol{r}|, \tag{A.18}$$
$$\nabla(1/r) = -\boldsymbol{r}/r^3, \tag{A.19}$$
$$\nabla\cdot(\boldsymbol{r}/r^3) = -\nabla^2(1/r) = 4\pi\delta(r). \tag{A.20}$$

Integral Relations. Let a vector field $\boldsymbol{A}$ and its divergence $\nabla\cdot\boldsymbol{A}$ be continuous over a closed region V with the surface S, the surface element $\mathrm{d}\boldsymbol{S}$ being counted positive in the direction outward from the enclosed volume. Then Gauss theorem states

$$\oint_S \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{S} = \int_V (\nabla\cdot\boldsymbol{A})\,\mathrm{d}v, \tag{A.21}$$

while Stokes' theorem postulates

$$\oint_S \mathrm{d}\boldsymbol{S}\times\boldsymbol{A} = \int_V (\nabla\times\boldsymbol{A})\,\mathrm{d}v. \tag{A.22}$$

Green's theorem is

$$\int_V (\phi\,\nabla\cdot\nabla\psi - \psi\,\nabla\cdot\nabla\phi)\,\mathrm{d}v = \oint_S (\phi\,\nabla\psi - \psi\,\nabla\phi)\cdot\mathrm{d}\boldsymbol{S}. \tag{A.23}$$

Appendix B
The Fourier Transform[1]

Fourier transform:

$$F(s) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi s x}\, dx. \tag{B.1}$$

Inverse Fourier transform:

$$f(x) = \int_{-\infty}^{\infty} F(s)\, e^{i2\pi x s}\, ds. \tag{B.2}$$

Table B.1 Theorems for the Fourier transform

Theorem	$f(x)$	$F(s)$	
Similarity	$f(ax)$	$\dfrac{1}{\lvert a \rvert} F\left(\dfrac{s}{a}\right)$	(B.3)
Addition	$f(x) + g(x)$	$F(s) + G(s)$	(B.4)
Shift	$f(x-a)$	$e^{-i2\pi a s} F(s)$	(B.5)
Modulation	$f(x)\cos x$	$\dfrac{1}{2} F\left(s - \dfrac{\omega}{2\pi}\right) + \dfrac{1}{2} F\left(s + \dfrac{\omega}{2\pi}\right)$	(B.6)
Convolution	$f(x) \otimes g(x)$	$F(s)\, G(s)$	(B.7)
Autocorrelation	$f(x) \otimes f^*(-x)$	$\lvert F(s) \rvert^2$	(B.8)
Derivative	$f'(x)$	$i2\pi s\, F(s)$	(B.9)

Rayleigh theorem:

$$\int_{-\infty}^{\infty} \lvert f(x) \rvert^2\, dx = \int_{-\infty}^{\infty} \lvert F(s) \rvert^2\, ds. \tag{B.10}$$

Power theorem:

$$\int_{-\infty}^{\infty} f(x)\, g^*(x)\, dx = \int_{-\infty}^{\infty} F(s)\, G^*(s)\, ds. \tag{B.11}$$

---

[1] Adopted from Bracewell, R. (1965): *The Fourier Transform and its Applications* 1965 (Mc Graw Hill, New York).

Table B.2 A short list of Fourier transform pairs

$f(x)$	$F(s)$	$f(x)$	$F(s)$
$e^{-\pi x^2}$	$e^{-\pi s^2}$	$e^{-\lvert x\rvert}$	$\dfrac{2}{1+(2\pi s)^2}$
$xe^{-\pi x^2}$	$-ise^{-\pi s^2}$	$e^{-\lvert x\rvert}\dfrac{\sin x}{x}$	$\arctan\dfrac{1}{2\pi^2 s^2}$
1	$\delta(s)$	$\lvert x\rvert^{-1/2}$	$\lvert s\rvert^{-1/2}$
$\cos\pi x$	$II(s)$	$e^{-\lvert x\rvert}\cos\pi x$	$\dfrac{2}{1+(2\pi s)^2}\otimes II(s)$
$\sin\pi x$	$i\,{}^{I}_{I}(s)$	$\operatorname{sech}\pi x$	$\operatorname{sech}\pi x$
$III(x)$	$III(s)$	$\operatorname{sech}^2\pi x$	$2s\operatorname{cosech}\pi s$
$\operatorname{sinc}x$	$II(s)$	$H(x)$	$\dfrac{1}{2}\delta(s)-\dfrac{i}{2\pi s}$
$\operatorname{sinc}^2 x$	$II(s)\otimes II(s)$	$J_0(2\pi x)$	$\dfrac{II(s/2)}{\pi(1-s^2)^{1/2}}$
$\operatorname{sinc}^3 x$	$II(s)\otimes II(s)\otimes II(s)$	$J_1(2\pi x)/2x$	$(1-s^2)^{1/2}II\left(\dfrac{s}{2}\right)$

Where:

$$\operatorname{sinc}x = \frac{\sin\pi x}{\pi x},$$

$$II(x) = \begin{cases} 1 \text{ for } \lvert x\rvert < \dfrac{1}{2} \\[2mm] 0 \text{ for } \lvert x\rvert > \dfrac{1}{2} \end{cases},$$

$$H(x) = \begin{cases} 1 \text{ for } \lvert x\rvert > 0 \\ 0 \text{ for } \lvert x\rvert < 0 \end{cases},$$

$$III(x) = \sum_{n=-\infty}^{\infty} \delta(x-n),$$

$$II(x) = \tfrac{1}{2}\delta(x+1/2) + \tfrac{1}{2}\delta(x-1/2),$$

$${}^{I}_{I}(x) = \tfrac{1}{2}\delta(x+1/2) - \tfrac{1}{2}\delta(x-1/2).$$

Appendix C
The Van Vleck Clipping Correction: One Bit Quantization

One-bit quantization strongly influences the appearance of both the signal and the resulting ACF. We will determine the ACF $R_y(\tau)$ of some transformation $y(t) = \varphi[x(t)]$ chosen such that it can be easily implemented and $R_x(\tau)$ can be computed from $R_y(\tau)$. Let the stochastic process $y(t)$ be defined by (Fig. C.1)

$$y(t) = \begin{cases} 1 & \text{for } x(t) \gtrsim 0 \\ -1 & \text{for } x(t) < 0 \end{cases} . \tag{C.1}$$

$$R_y(\tau) = \frac{2}{\pi} \arcsin\left(\frac{R_x(\tau)}{R_x(0)}\right)$$
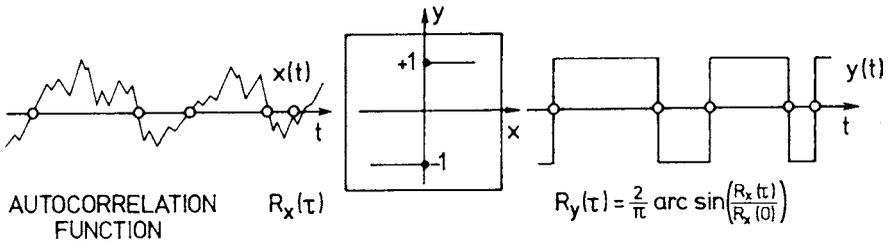
Fig. C.1 An illustration of one-bit clipping of a signal, $x(t)$, to produce the clipped signal, $y(t)$. The autocorrelation function of the input signal is recovered from that of $y(t)$ using the arcsin law

Then the ACF of y, $R_y(\tau)$ is given by

$$R_y(\tau) = E\{y(t+\tau)y(t)\} = P\{x(t+\tau)x(t) > 0\} - P\{x(t+\tau)x(t) < 0\} \tag{C.2}$$

where $P\{x(t+\tau)x(t) > 0\}$ is the probability of finding $x(t+\tau)x(t) > 0$. This can be computed using the joint probability density function $p(x_1, x_2; \tau)$. For a general stochastic process this is not known, but for a Gaussian process it is given by (C.5). The ACF of the Gaussian signal $x(t)$ is

$$E\{x(t+\tau)x(t)\} = R(\tau);$$

then the random variables $x(t+\tau)$ and $x(t)$ are jointly normal with the same variance $E\{x^2\} = R(0)$ and the correlation coefficient

$$r = \frac{E\{x(t+\tau)x(t)\}}{\sqrt{E\{x^2(t+\tau)\}E\{x^2(t)\}}} = \frac{R(\tau)}{R(0)} \tag{C.3}$$

so that their joint density distribution function is given by

$$p(x_1, x_2; \tau)$$

$$= \frac{1}{2\pi\sqrt{R^2(0) - R^2(\tau)}} \exp\left[-\frac{R(0)x_1^2 - 2R(\tau)x_1 x_2 + R(0)x_2^2}{2[R^2(0) - R^2(\tau)]}\right] \tag{C.4}$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - r^2}} \exp\left[-\frac{1}{2(1 - r^2)}\left(\frac{x_1^2}{\sigma_1^2} - \frac{2rx_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right)\right]. \tag{C.5}$$

Since the voltage distribution of the stochastic process describing both the signal and the noise in a receiver are accurately described by such a distribution, (C.5) is applicable.

Then we have,

$$P\{x(t+\tau)x(t) > 0\} = P\{[x(t+\tau) > 0] \wedge [x(t) > 0]\}$$
$$+ P\{[x(t+\tau) < 0] \wedge [x(t) < 0]\}$$
$$= P_{++} + P_{--} \tag{C.6}$$

and similarly

$$P\{x(t+\tau)x(t) < 0\} = P_{+-} + P_{-+} = 2P_{+-}, \tag{C.7}$$

where

$$P_{++} = \int_0^\infty\int_0^\infty p[x(t+\tau), x(t); \tau]\, dx(t+\tau)\, dx(t).$$

Substituting

$$x(t) = \sigma\cos\theta/\sigma_1,$$
$$x(t+\tau) = \sigma\sin\theta/\sigma_2,$$

and (C.5) for $p(x_1, x_2; \tau)$ we obtain

$$P_{++} = \frac{1}{2\pi\sigma^2(1 - r^2)^{1/2}} \int_0^{\pi/2}\int_0^\infty \exp\left[-\frac{z^2(1 - r\sin 2\theta)}{2\sigma^2(1 - r^2)}\right] z\, dz\, d\theta$$

$$= \frac{1}{4} + \frac{1}{2\pi}\arctan\frac{r}{\sqrt{1 - r^2}}$$

$$= \frac{1}{4} + \frac{1}{2\pi}\arcsin r. \tag{C.8}$$

Similarly

$$P_{--} = \frac{1}{4} + \frac{1}{2\pi} \arcsin r, \tag{C.9}$$

$$P_{+-} = \frac{1}{4} - \frac{1}{2\pi} \arcsin r. \tag{C.10}$$

Substituting (C.6, C.7, C.8, C.9 and C.10) into (C.2), we obtain with (C.3)

$$\boxed{R_y(\tau) = \frac{2}{\pi} \arcsin \frac{R_x(\tau)}{R_x(0)}} \tag{C.11}$$

or

$$\boxed{R_x(\tau) = R_x(0) \sin[\pi/2 R_y(\tau)]}. \tag{C.12}$$

$R_x(0)$ is the undelayed autocorrelation; the maximum value of the autocorrelated signal. (C.11) is known as the *arcsine law*, the inverse relation (C.12) is the *van Vleck clipping correction*. Thus if $R_y(\tau)$ can be measured, $R_x(\tau)$ can be easily computed.

Appendix D
The Reciprocity Theorem

Consider two antennas, 1 and 2. Let 1 be a transmitting antenna powered by the generator G while 2 is a receiving antenna that induces a certain current measured by M (Fig. D.1). The receiving antenna is assumed to be oriented such that M shows a maximum deflection and we will assume that no ohmic losses occur in 2.
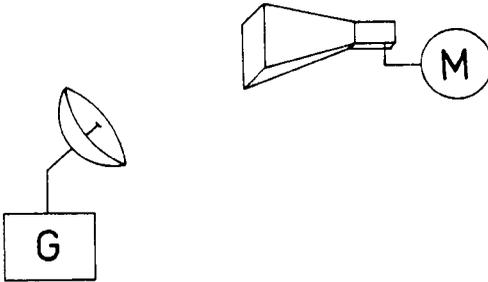
Fig. D.1 A sketch to illustrate the reciprocity theorem; G is the generator, transmitting a signal, and M is a meter for the measurement of the received signal

The reciprocity theorem now states that the current measurement on M remains the same even if we exchange generator G and meter M. Therefore it does not matter which antenna is transmitting and which is receiving. However, for the medium between 1 and 2, we must require that it has no preferred direction; that is, that its transmission properties are the same from 1 directed towards 2 as from 2 directed towards 1. There are some materials, such as certain ferrites in a magnetic field, that do not fulfill this prerequisite; these are used in direction-sensitive devices such as circulators or isolators. If such material is involved, an application of the reciprocity theorem needs special care.

To prove the reciprocity theorem consider Maxwell's equations for the two systems 1 and 2:

$$\nabla \times \boldsymbol{H}_1 = \frac{4\pi}{c} \boldsymbol{J}_1 + \frac{\varepsilon}{c} \dot{\boldsymbol{E}}_1 , \quad \nabla \times \boldsymbol{H}_2 = \frac{4\pi}{c} \boldsymbol{J}_2 + \frac{\varepsilon}{c} \dot{\boldsymbol{E}}_2 ,$$

$$\nabla \times \boldsymbol{E}_1 = -\frac{\mu}{c} \dot{\boldsymbol{H}}_1 , \qquad \nabla \times \boldsymbol{E}_2 = -\frac{\mu}{c} \dot{\boldsymbol{H}}_2 .$$

Forming

$$\nabla \cdot (\boldsymbol{E}_1 \times \boldsymbol{H}_2) = \boldsymbol{H}_2 \cdot (\nabla \times \boldsymbol{E}_1) - \boldsymbol{E}_1 \cdot (\nabla \times \boldsymbol{H}_2)$$

$$= -\frac{\mu}{c} \dot{\boldsymbol{H}}_1 \cdot \boldsymbol{H}_2 - \frac{4\pi}{c} \boldsymbol{E}_1 \cdot \boldsymbol{J}_2 - \frac{\varepsilon}{c} \boldsymbol{E}_1 \cdot \dot{\boldsymbol{E}}_2$$

$$\nabla \cdot (\boldsymbol{E}_2 \times \boldsymbol{H}_1) = \boldsymbol{H}_1 \cdot (\nabla \times \boldsymbol{E}_2) - \boldsymbol{E}_2 \cdot (\nabla \times \boldsymbol{H}_1)$$

$$= -\frac{\mu}{c} \boldsymbol{H}_1 \cdot \dot{\boldsymbol{H}}_2 - \frac{4\pi}{c} \boldsymbol{E}_2 \cdot \boldsymbol{J}_1 - \frac{\varepsilon}{c} \dot{\boldsymbol{E}}_1 \cdot \boldsymbol{E}_2$$

according to (A.9), the difference of these two equations is

$$\nabla \cdot (\boldsymbol{E}_1 \times \boldsymbol{H}_2 - \boldsymbol{E}_2 \times \boldsymbol{H}_1) = \frac{\mu}{c} (\boldsymbol{H}_1 \cdot \dot{\boldsymbol{H}}_2 - \dot{\boldsymbol{H}}_1 \cdot \boldsymbol{H}_2)$$

$$- \frac{4\pi}{c} (\boldsymbol{E}_1 \cdot \boldsymbol{J}_2 - \boldsymbol{E}_2 \cdot \boldsymbol{J}_1)$$

$$- \frac{\varepsilon}{c} (\boldsymbol{E}_1 \cdot \dot{\boldsymbol{E}}_2 - \dot{\boldsymbol{E}}_1 \cdot \boldsymbol{E}_2) .$$

We now consider harmonic waves as in (2.35)

$$\dot{\boldsymbol{H}} = -\mathrm{i}\,\omega \boldsymbol{H} , \qquad \dot{\boldsymbol{E}} = -\mathrm{i}\,\omega \boldsymbol{E} ,$$

and thus

$$\boldsymbol{H}_1 \cdot \dot{\boldsymbol{H}}_2 - \dot{\boldsymbol{H}}_1 \cdot \boldsymbol{H}_2 = 0 , \qquad \boldsymbol{E}_1 \cdot \dot{\boldsymbol{E}}_2 - \dot{\boldsymbol{E}}_1 \cdot \boldsymbol{E}_2 = 0 ,$$

so that

$$\nabla \cdot (\boldsymbol{E}_1 \times \boldsymbol{H}_2 - \boldsymbol{E}_2 \times \boldsymbol{H}_1) = \frac{4\pi}{c} (\boldsymbol{E}_2 \cdot \boldsymbol{J}_1 - \boldsymbol{E}_1 \cdot \boldsymbol{J}_2) .$$

But according to the Gauss theorem [see (A.21)]

$$\int_V \nabla \cdot (\boldsymbol{E}_1 \times \boldsymbol{H}_2 - \boldsymbol{E}_2 \times \boldsymbol{H}_1) \, \mathrm{d}V = \oint_S (\boldsymbol{E}_1 \times \boldsymbol{H}_2 - \boldsymbol{E}_2 \times \boldsymbol{H}_1) \cdot \mathrm{d}\boldsymbol{S} .$$

If V is a sphere, the radius of which tends towards ∞, then $\boldsymbol{E}$ and $\boldsymbol{H}$ tend towards zero at the surface of this sphere and we suppose that the surface integral on the right-hand side vanishes. For a rigorous proof of this, we would have to show that $|\boldsymbol{E}_1 \times \boldsymbol{H}_2 - \boldsymbol{E}_2 \times \boldsymbol{H}_1|$ is indeed decreasing faster than $1/r^2$. This can be done considering that $\boldsymbol{E} \perp \boldsymbol{H}$ for spherical waves; we will not give the details here, but adopting this we find

$$\int_V (\boldsymbol{E}_2 \cdot \boldsymbol{J}_1 - \boldsymbol{E}_1 \cdot \boldsymbol{J}_2) \, \mathrm{d}V = 0 . \qquad (\text{D.1})$$

If the two antennas 1 and 2 are contained in different space regions V_1 and V_2, then

$$\int\limits_{V_1} \boldsymbol{E}_2 \cdot \boldsymbol{J}_1 \, \mathrm{d}V = \int\limits_{V_2} \boldsymbol{E}_1 \cdot \boldsymbol{J}_2 \, \mathrm{d}V \tag{D.2}$$

since $\boldsymbol{J}_1 = 0$ in V_2 and $\boldsymbol{J}_2 = 0$ in V_1; the distance between the two antennas is arbitrary.

If an antenna is contained in an infinitesimal cylinder with the cross section q and the length $\mathrm{d}l$

$$\mathrm{d}V = q \, \mathrm{d}l \,,$$

the total current in the antenna is

$$I = q \, |\boldsymbol{J}|$$

and the voltage

$$U = E \, \mathrm{d}l$$

and thus

$$\boxed{U_2 I_1 = U_1 I_2} \quad . \tag{D.3}$$

Here U_1 is the voltage induced by antenna 2 in antenna 1 and I_1 is the total current in antenna 1; U_2 and I_2 are similar quantities for antenna 2. Equation (D.3) is a quantitative formulation of the reciprocity theorem.

Appendix E
The Hankel Transform[1]

Hankel transform:

$$F(q) = 2\pi \int_0^\infty f(r) J_0(2\pi q r)\, r\, dr.$$ (E.1)

Inverse Hankel transform:

$$f(r) = 2\pi \int_0^\infty F(q) J_0(2\pi q r)\, q\, dq.$$ (E.2)

Table E.1 Theorems for the Hankel transform

Theorem	$f(r)$	$F(q)$
Similarity	$f(ar)$	$\dfrac{1}{a^2} F\left(\dfrac{q}{a}\right)$
Addition	$f(r) + g(r)$	$F(q) + G(q)$
Shift	shift of origin destroys circular symmetry	
Convolution	$\displaystyle\int_0^\infty \int_0^{2\pi} f(r')\, g(R)\, r'\, dr'\, d\theta$ $R^2 = r^2 + r'^2 - 2rr' \cos\theta$	$F(q)\, G(q)$

Rayleigh theorem:

$$\int_0^\infty |f(r)|^2\, r\, dr = \int_0^\infty |F(q)|^2\, q\, dq.$$ (E.3)

Power theorem:

---

[1] Adopted from Bracewell, R. (1965): *The Fourier Transform and its Applications* 1965 (Mc Graw Hill, New York).

$$\int\limits_0^\infty f(r)\,g^*(r)\,r\,\mathrm{d}r = \int\limits_0^\infty F(q)\,G^*(q)\,q\,\mathrm{d}q. \tag{E.4}$$

Table E.2 Some Hankel transforms

$f(r)$	$F(q)$
$II\left(\dfrac{r}{2a}\right)$	$\dfrac{a}{q}J_1(2\pi a q)$
$\dfrac{1}{r}\sin(2\pi a r)$	$\dfrac{II(q/2a)}{(a^2-q^2)^{1/2}}$
$\dfrac{1}{2}\delta(r-a)$	$\pi a J_0(2\pi a q)$
$\mathrm{e}^{-\pi r^2}$	$\mathrm{e}^{-\pi q^2}$
$(a^2+r^2)^{-1/2}$	$\dfrac{1}{q}\mathrm{e}^{-2\pi a q}$
$(a^2+r^2)^{-3/2}$	$\dfrac{2\pi}{a}\mathrm{e}^{-2\pi a q}$
$(a^2+r^2)^{-1}$	$2\pi K_0(2\pi a q)$
$2a^2\,(a^2+r^2)^{-2}$	$4\pi^2 a q K_1(2\pi a q)$
$(a^2-r^2)II\left(\dfrac{r}{2a}\right)$	$\dfrac{a^2}{\pi q^2}J_2(2\pi a q)$
$\dfrac{1}{r}$	$\dfrac{1}{q}$
e^{-ar}	$2\pi a\,(4\pi^2 q^2+a^2)^{-3/2}$
$\dfrac{1}{r}\mathrm{e}^{-ar}$	$2\pi(4\pi^2 q^2+a^2)^{-1/2}$
$\dfrac{1}{\pi r}\delta(r)$	1
$\dfrac{1}{2a^2}\left[II\left(\dfrac{r}{2a}\right)\otimes II\left(\dfrac{r}{2a}\right)\right]$	$\dfrac{1}{2a^2}\,|J_1(2\pi a q)|^2$
$r^2\mathrm{e}^{-\pi r^2}$	$\dfrac{1}{\pi}\left(\dfrac{1}{\pi}-q^2\right)\mathrm{e}^{-\pi q^2}$

All definitions given in this Table are as in Appendix B.

Appendix F
A List of Calibration Radio Sources

The usual method to determine the flux of radio sources is to measure the ratio of the response of the antenna / receiver combination to the source and to that of a well-known calibrator. In order to avoid as much as possible the effects of the varying telescope efficiencies it is desirable to have a set of calibrators reasonably well distributed over the sky, such that radio source and calibration source can be measured at nearly equal zenith angles.

The list of calibration sources given in Table F.1 is based on a compilation originally given by Baars, J. W. M. et al. (1977): Astron. Astrophys. **61**, 99 and updated by Ott, M. et al. (1994): Astron. Astrophys. **284**, 331. Values in brackets are taken from the compilation of Baars et al.

In Table F.2 a list of secondary calibrators at submillimeter wavelengths is given. This data are taken from G. Sandell (1994): M. N. R. A. S. **271**, 75. The data were obtained with the James Clerk Maxwell Telescope (JCMT) on Mauna Kea in 1989–1991 using the UKT14 bolometer at the Nasmyth focus. The flux density scales are calibrated relative to Mars and Uranus. The standard deviations of the data show that accuracies of a few percent are reached.

Table F.1 A list of calibration radio sources at microwave frequencies

Source	RA (1950.0) [^{h} ^{m} ^{s}]	Dec (1950.0) [° ′ ″]	b^{II} [°]	$S_{1408}^{(a)}$ [Jy]	S_{2695} [Jy]	S_{4750} [Jy]	S_{10550} [Jy]	S_{23780} [Jy]	Ident.$^{(b)}$	Ang.Size (″×″)
3C 48	01 34 49.8	+32 54 20	−29	16.27	9.49	5.72	2.65	1.11	QSO	1.5 × 1.5
3C 123	04 33 55.2	+29 34 14	−12	47.47	27.62	16.56	7.27	3.12	GAL	23 × 5
3C 147	05 38 43.5	+49 49 42	+10	21.86	13.04	7.92	3.68	1.70	QSO	1 × 1
3C 161	06 24 43.1	−05 51 14	−8	18.58	11.13	6.75	2.94	–	GAL	3 × 3
3C 218	09 15 41.5	−11 53 06	+25	42.65	23.19	13.65	6.53	–	GAL	47 × 15
3C 227	09 45 07.8	+07 39 09	+42	7.61	4.39	2.96	(1.34)	(0.73)	GAL	200 × 50
3C 249.1	11 00 25.0	+77 15 11	+39	2.28	1.33	0.82	(0.39)	–	QSO	15 × 15
3C 274	12 28 17.7	+12 39 55	+74	203.0	121.7	77.3	40.2	(20.0)	GAL	150 × 250
3C 286	13 28 49.7	+30 45 58	+81	14.68	10.52	7.60	4.45	2.38	QSO	1.5 × 1.5
3C 295	14 09 33.5	+52 26 13	+61	22.22	12.24	6.74	2.58	0.92	GAL	5 × 1
3C 348	16 48 40.1	+05 04 28	+29	46.73	23.37	13.11	4.83	–	GAL	170 × 2.5
3C 353	17 17 54.6	−00 55 55		56.25	34.26	23.49	10.71	–	GAL	210 × 60
DR 21	20 37 14.2	+42 09 07	+1	–	14.99	19.07	(20.8)	(19.0)	HII	20 × 20
NGC 7027	21 05 09.4	+42 02 03	−3	1.34	3.62	5.47	6.13	5.39	PN	7 × 10

$^{a)}$ Where S_{1408} is the flux density in Jy at 1408 MHz.
$^{b)}$ Where QSO denotes a Quasi-Stellar Object and GAL denotes a galaxy.

Table F.2 Calibrators at submillimeter wavelengths

Source[a]	R.A.(1950) [h m s]	Dec.(1950) [° ′ ″]	2 mm [Jy]	1.3 mm [Jy]	1.1 mm [Jy]	850 μm [Jy]	800 μm [Jy]	800 μm [Jy]	750 μm [Jy]	450 μm [Jy]	350 μm [Jy]	Ang.Size ("×")
W3(OH)	02 23 16.45	+61 38 57.2	5.5	10.2	12.8	25.5	30.9	26.5	39.9	222	498	14 × 10
G343.0	16 54 43.77	−42 47 32.4	2.0	6.2	8.7	19.8	27.3	23.3	37.1	215	419	10 × 15
NGC 6334I	17 17 32.34	−35 44 03.1	8.8	21.9	30.0	65.3	80.8	71.7	109.1	600	1263	12 × 6
G5.89	17 57 26.75	−24 03 56.0	9.5	13.8	16.3	32.6	36.2	32.0	45.1	259	571	9 × 10
G10.62	18 07 30.66	−19 56 29.1	7.1	13.2	127.2	37.0	43.7	38.1		356	788	12 × 10
G34.3	18 50 46.17	+01 11 12.6	12.5	24.2	31.3	62.5	73.9	63.4	103.1	511	1125	6 × 8
G45.1	19 11 00.42	+10 45 42.9	1.8	3.2	4.0		10.3		13.2	78	187	9 × 10
K3-50	19 59 50.09	+33 24 19.4	6.5	8.5	9.2	15.8	17.9	16.7	21.9	110	241	11 × 6
ON-1	20 08 09.91	+31 22 42.3	1.6	3.6	4.7	10.8	15.4	12.5		143	271	11 × 17
W75N	20 36 50.00	+42 26 58.0	2.9	8.5	11.6	27.3	33.2	27.9	44.6	283	672	9 × 11
NGC 7538	23 11 36.62	+61 11 49.6	4.9	8.2	9.9	19.0	22.2	18.3	27.8	157	301	20 × 18
GL 490	03 23 39.22	+58 36 36.6	0.9	2.3	2.9	5.9	6.7	6.2	8.2	38	45	< 5 × 9
L1551	04 28 40.24	+18 01 42.1	0.7	1.7	2.4		6.2	5.7	8.2	41	95	5 × 10
NGC 2071IR	05 44 30.7	+00 20 45.0	1.5	4.2	4.8	11.2	12.4	10.4		79	177	11 × 13
16293-2422	16 29 21.05	−24 22 15.5	2.4	6.1	8.4	17.3	21.1	18.5	28.1	119	235	9 × 11
VY CMa	07 20 54.7	−25 40 12.4	0.35	0.66	0.90	2.2	2.2		2.1	9.7	19	< 5
OH231.8+4.2	07 39 59.0	−14 35 41.0	0.2	0.8	1.2	2.2	2.6		3.0	13	20	< 5
IRC+10216	09 45 14.89	+13 30 40.8	1.0	2.3	2.7	6.5	5.9	5.4		19	30	9 × 11
IRC+10216	09 45 14.89	+13 30 40.8	0.6	1.8	1.9	4.9	4.5	4.0		16		9 × 11
CRL 2688	21 00 19.9	+36 29 45.0	0.7	2.1	2.7	5.6	6.5		7.0	24	46	< 5
NGC 7027	21 05 09.4	+42 02 03.0	4.4	3.8	3.7		4.8			7.1	13	10

[a] All are galactic sources.

Appendix G
The Mutual Coherence Function and van Cittert-Zernike Theorem

G.1 The Mutual Coherence Function

It was shown in Chap. 3 that four parameters are needed to describe the full state of polarization of a quasi-monochromatic electromagnetic wave field, even at a single point. Therefore, when the wave fields at two separate points are to be compared, a multitude of correlations can be formed. We will simplify the discussion by considering only wave fields with a single state of polarization, so that only one scalar quantity is needed to describe the wave field at a single point. This quantity could be any of the four Stokes parameters or any other component of an orthogonal representation of the wave field. We now consider the distribution of this one parameter over the whole wave field.

Considering the simplest wave field that can be imagined, namely, that of a plane monochromatic wave, the field intensity at a position, P_2, for all time can be calculated from that at another position, P_1, once the phase difference for these two points has been determined. An arbitrary polychromatic wave field is in some respects the other extreme; If P_1 and P_2 are not too close, even a full knowledge of the field and its time variation at P_1 has no relation to the field at P_2. The monochromatic plane wave field is said to be fully *coherent*; the second example is that of an *incoherent* wave field. Other wave fields will have properties between these two extremes. Obviously a measure of this coherence is needed that can be determined in some practical way, even for fields where the instantaneous values of the strengths at a chosen point cannot be measured. A useful measure of coherence must be based on time averages; we will only consider stationary fields. Therefore the *mutual coherence function* of the (complex) wave field $U(P_1, t_1)$ and $U(P_2, t_2)$ will be defined as

$$
\begin{aligned}
\Gamma(P_1, P_2, \tau) &= \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} U(P_1, t) U^*(P_2, t + \tau) \, \mathrm{d}t \\
&= \langle U(P_1, t) U^*(P_2, t + \tau) \rangle
\end{aligned}
\tag{G.1}
$$

Where the $\langle \cdots \rangle$ brackets are used to indicate time averaging, as introduced in (3.46). We will assume that this limit exists. The intensity of the wave (3.48) is a special case of this definition

$$I(P) = \Gamma(P,P,0) = \langle U(P,t)U^*(P,t)\rangle \,. \tag{G.2}$$

For a plane monochromatic wave field propagating in the z direction, Γ is easily computed. Using a complex representation,

$$U(P,t) = U_0 \, \mathrm{e}^{\mathrm{i}(kz-\omega t)},$$

where

$$P = (x,y,z), \quad k = 2\pi/\lambda = \text{const}, \quad \omega = 2\pi v = \text{const},$$

then we have

$$\Gamma(P_1,P_2,\tau) = |U_0|^2 \, \mathrm{e}^{\mathrm{i}[k(z_1-z_2)+\omega\tau]}, \tag{G.3}$$

where τ is the time delay. The mutual coherence function of the travelling monochromatic wave field is periodic with a constant amplitude and a wavelength equal to that of the original wave field. The coherence function does not propagate; it is a standing wave with a phase such that, for $\tau = 0, \Gamma = \Gamma_{\max}$ for $z_1 = z_2$. It is often useful to normalize Γ by referring it to a wave field of intensity I. Thus

$$\gamma(P_1,P_2,\tau) = \frac{\Gamma(P_1,P_2,\tau)}{\sqrt{I(P_1)I(P_2)}}\,. \tag{G.4}$$

For this complex coherence, we always have

$$|\gamma(P_1,P_2,\tau)| \leqq 1\,. \tag{G.5}$$

G.2   The Coherence Function of Extended Sources: The van Cittert-Zernike Theorem

Taking (G.3) and (G.4) together we see that for a monochromatic plane wave the complex degree of coherence has the constant amplitude 1, while an arbitrary polychromatic wave field will result in $\gamma = 0$ for $P_1 \neq P_2$. Which properties of a wave field result in a partial loss of coherence? We will gradually reduce the restrictions imposed on the wave field in order to see the effect this has on the degree of coherence. In the following we will assume that the radiation is: (1) spatially incoherent, (2) well described by a single component of the two vector fields $\boldsymbol{E}$ and $\boldsymbol{B}$, (3) stationary in time, and (4) radiated by very distant sources. A wave field that is only slightly more complex than a monochromatic plane wave is formed by the (incoherent) superposition of two such wave fields with identical wavelengths (and frequencies) but propagating in different directions (Fig. G.1):

$$\begin{aligned} U_a &= U_{0a}\, \mathrm{e}^{\mathrm{i}(k\boldsymbol{s}_a\cdot\boldsymbol{x}-\omega t)}, \\ U_b &= U_{0b}\, \mathrm{e}^{\mathrm{i}(k\boldsymbol{s}_b\cdot\boldsymbol{x}-\omega t)}. \end{aligned} \tag{G.6}$$
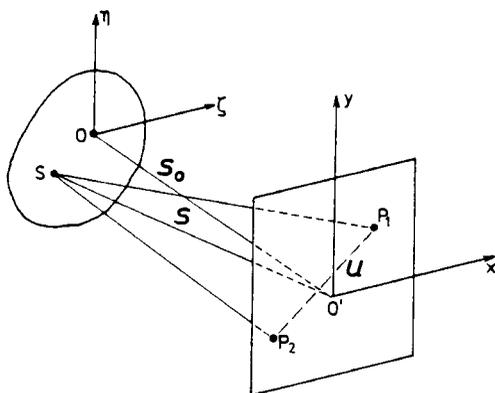
Fig. G.1 The coordinate systems and designations used in the discussion of the van Cittert-Zernike theorem

Here s_a and s_b are unit vectors describing the propagation direction, and both $k = 2\pi/\lambda$ and $\omega = 2\pi\nu$ are assumed to be equal for both waves. The total wave field is then formed by

$$U = U_a + U_b$$

and the mutual coherence function (G.1) is

$$\begin{aligned}\langle U(P_1,t_1)U^*(P_2,t_2)\rangle &= \langle\{U_a(P_1,t_1)+U_b(P_1,t_1)\}\{U_a(P_2,t_2)+U_b(P_2,t_2)\}^*\rangle\\ &= \langle U_a(P_1,t_1)U_a^*(P_2,t_2)\rangle + \langle U_b(P_1,t_1)U_b^*(P_2,t_2)\rangle\\ &\quad + \langle U_a(P_1,t_1)U_b^*(P_2,t_2)\rangle\\ &\quad + \langle U_b(P_1,t_1)U_a^*(P_2,t_2)\rangle.\end{aligned} \qquad (G.7)$$

If we assume the two wave fields U_a and U_b are *incoherent*, we require that the field strengths U_a and U_b are uncorrelated even when measured at the same point, so that

$$\langle U_a(P_1,t_1)U_b^*(P_2,t_2)\rangle = \langle U_b(P_1,t_1)U_a^*(P_2,t_2)\rangle \equiv 0. \qquad (G.8)$$

Such incoherence is not possible for strictly monochromatic waves of identical polarization consisting of a wave train of infinite duration and length. Equation (G.8) can be correct only if the wave is made up of sections of finite duration between which arbitrary phase jumps occur. Then the waves are not strictly monochromatic but have a finite, although small bandwidth. Substituting (G.8) into (G.7) we obtain

$$\begin{aligned}\Gamma(P_1,P_2,\tau) &= \langle U(P_1,t)U^*(P_2,t+\tau)\rangle\\ &= \langle U_a(P_1,t)U_a^*(P_2,t+\tau)\rangle + \langle U_b(P_1,t)U_b^*(P_2,t+\tau)\rangle\end{aligned}$$

or, using (G.6),

$$\Gamma(P_1,P_2,\tau) = |U_{0a}|^2\, e^{\mathrm{i}(k s_a \cdot \boldsymbol{u}+\omega\tau)} + |U_{0b}|^2\, e^{\mathrm{i}(k s_b \cdot \boldsymbol{u}+\omega\tau)}\,, \tag{G.9}$$

where

$$\boldsymbol{u} = \boldsymbol{x}_1 - \boldsymbol{x}_2\,. \tag{G.10}$$

Thus only the difference of the two positions P_1 and P_2 enters into the problem. For the case of two waves of equal amplitude,

$$|U_{0a}| = |U_{0b}| = |U_0|\,,$$

(G.9) can be simplified using the identities

$$s_a = \frac{1}{2}(s_a + s_b) + \frac{1}{2}(s_a - s_b)\,,$$

$$s_b = \frac{1}{2}(s_a + s_b) - \frac{1}{2}(s_a - s_b)\,,$$

resulting in

$$\Gamma(\boldsymbol{u},\tau) = 2|U_0|^2 \cos\left(\frac{k}{2}(s_a - s_b)\cdot\boldsymbol{u}\right)\, e^{\mathrm{i}\left(\frac{k}{2}(s_a+s_b)\cdot\boldsymbol{u}+\omega\tau\right)}\,, \tag{G.11}$$

or, if normalized,

$$\gamma(\boldsymbol{u},\tau) = \cos\left(\frac{k}{2}(s_a - s_b)\boldsymbol{u}\right)\, e^{\mathrm{i}\left(\frac{k}{2}(s_a+s_b)\cdot\boldsymbol{u}+\omega\tau\right)}\,. \tag{G.12}$$

For two waves propagating in directions that differ only slightly, $|s_a - s_b|/2$ is a small quantity, while $(s_a + s_b)/2$ differs only little from either s_a or s_b. The normalized coherence function (G.12) therefore is similar to that of a single plane wave, but with an amplitude that varies slowly with position. We will have a complete loss of coherence for

$$\frac{k}{2}(s_a - s_b)\cdot\boldsymbol{u} = (2n+1)\frac{\pi}{2}\,, \qquad n = 0,1,2,\dots\,. \tag{G.13}$$

This principle of superposition of simple monochromatic plane waves can be extended to an arbitrary number of plane waves, and the result will be a simple generalization of (G.9) if we again assume these fields to be mutually incoherent. The signals at P_1 and P_2 are then the sum of the components $U_n(P,t)$,

$$U(P,t) = \sum_n U_n(P,t)\,, \tag{G.14}$$

and, if the different waves are incoherent, then

$$\langle U_m(P_1,t)\, U_n^*(P_2,t+\tau)\rangle = 0 \qquad \text{for all}\quad m \neq n\,, \tag{G.15}$$

while

$$\langle U_n(P_1,t)\, U_n^*(P_2,t+\tau)\rangle = |U_{0n}|^2\, e^{i(k s_n \cdot u + \omega\tau)}, \qquad (G.16)$$

so that

$$\Gamma(\boldsymbol{u},\tau) = \langle U(P_1,t)\, U^*(P_2,t+\tau)\rangle = \sum_n |U_{0n}|^2\, e^{i(k s_n \cdot u + \omega\tau)}.$$

Or, if we go to the limit $n \to \infty$

$$\boxed{\Gamma(\boldsymbol{u},\tau) = \iint I(\boldsymbol{s})\, e^{i(k s \cdot u + \omega\tau)}\, d\boldsymbol{s}} \qquad , \qquad (G.17)$$

where

$$\boxed{I(\boldsymbol{s}) = \iint U(\boldsymbol{s}+\boldsymbol{\sigma})\, U^*(\boldsymbol{s}+\boldsymbol{\sigma})\, d\boldsymbol{\sigma}} \qquad (G.18)$$

is the total intensity at the position P if the integral is taken over the angular extent of those positions $\boldsymbol{s}+\boldsymbol{\sigma}$ that contribute to the radiation field propagating into the direction $\boldsymbol{u}$, and the generalization of (G.14) is

$$\boxed{U(\boldsymbol{x},t) = \iint U(\boldsymbol{s})\, e^{i(k s \cdot x - \omega t)}\, d\boldsymbol{s}} \qquad . \qquad (G.19)$$

Equation (G.17) is the monochromatic version of the van Cittert-Zernike theorem. This theorem specifies how the mutual coherence function of an arbitrary monochromatic wave field (G.19), built up from plane waves is related to the intensity distribution (G.18). The proof given here is simplified and abbreviated; a more extensive version can be found in Born and Wolf (1964).

Provided that $\Gamma(\boldsymbol{u},\tau)$ can be measured and (G.17) can be solved for $I(\boldsymbol{s})$, so we can measure $I(\boldsymbol{s})$. As stated earlier the possible angular resolution using $I(\boldsymbol{s})$ depends on the size of the telescope used. For (G.17) the difference in the positions where measurements are made, $|\boldsymbol{u}|$ is introduced. Since it is possible to measure $\Gamma(\boldsymbol{u},\tau)$ for values of $|\boldsymbol{u}|$ much larger than the largest single telescope diameters possible, the resolution of $I(\boldsymbol{s})$ obtainable from the inversion of (G.17) is much greater than that which can be achieved by using telescopes that form direct images $I(\boldsymbol{s})$.

The principle, the methods and the limitations of the recovery of $I(\boldsymbol{s})$ from $\Gamma(\boldsymbol{u},\tau)$ will be discussed in Chapter 9. Both the theory and the practical details of it are fairly complicated, so this can only be considered an introduction. For detailed presentation, specialized books (given in the references) should be consulted.

Bibliography

Chapter 1

a) General

Burke, B.F., Graham-Smith, F. (1996): *An Introduction to Radio Astronomy* (Cambridge University Press, Cambridge)

Condon, J., Ransom, S. (2007): *Essential Radio Astronomy* http: /www.cv.nrao.edu/course/534

Gurvits, L., Frey, S., Rawlings, S. eds. (2005): *Radio Astronomy from Karl Jansky to Microjansky* (EDP Sciences, Paris)

Kraus, J.D. (1986): *Radio Astronomy*, 2nd ed. (Cygnus-Quasar, Powell, Ohio)

Leighton, R.B. (1960): *Principles of Modern Physics* (McGraw-Hill, New York)

Sullivan, W.T. ed. (1984): *The Early Years of Radio Astronomy: Reflections 50 Years After Jansky's Discovery* (Cambridge University Press, Cambridge)

Verschuur, G.L., Kellermann, K. I. eds. (1988): *Galactic and Extragalactic Radio Astronomy* 2nd ed. (Springer-Verlag, Heidelberg)

b) Special

Mihalas, D. (1978): *Stellar Atmospheres* (Freeman, San Francisco), Chaps. 1, 2

ITU Handbook on Radio Astronomy (http://www.itu.int/publ/R-HDB-22/en)

Reif, F. (1965): *Statistical and Thermal Physics* (McGraw Hill, New York)

Rybicki, G.B., Lightman, A.P. (1979): *Radiative Processes in Astrophysics* (Wiley, New York)

Chapter 2

Becker, R., Sauter, F. (1962): *Theorie der Elektrizität*, Vol. I (Teubner, Stuttgart)

Jackson, J.D. (1975): *Classical Electrodynamics*, 2nd ed. (Wiley, New York)

Kühn, R. (1964): *Mikrowellen-Antennen* (VEB Verlag Technik, Berlin)

Manchester, R.N., Taylor, J.H. (1977): *Pulsars* (Freeman, San Francisco, CA)

Panofsky, W., Phillips, M. (1962): *Classical Electricity and Magnetism* (Addison-Wesley, Reading, MA)

Sommerfeld, A. (1959): *Elektrodynamik* (Akademische Verlags-Gesellschaft, Leipzig)

Sommerfeld, A. (1964): *Optik* (Akademische Verlags-Gesellschaft, Leipzig)

Chapter 3

Born, M., Wolf, E. (1965): *Principles of Optics* (Pergamon, Oxford)

Sommerfeld, A. (1964): *Optik* (Akademische Verlags-Gesellschaft, Leipzig)

Spitzer, Jr., L. (1978): *Physical Processes in the Interstellar Medium* (Wiley, New York)

Wave propagation in a plasma is treated in most texts on plasma physics. Some well-known text-books are:

Chen, F.F. (1974): *Introduction to Plasma Physics* (Plenum, New York)

Krall, N.A., Trivelpiece, A.W. (1973): *Principles of Plasma Physics* (McGraw Hill, New York)

Spitzer, Jr., L. (1962): *Physics of Fully Ionized Gas*, 2nd ed. (Wiley, New York)

Wiener, N. (1949): *Extrapolation, Interpolation and Smoothing of Stationary Time Series* (MIT Cambridge, MA)

Recent references for polarization measurements in the centimeter, millimeter and sub-mm ranges are:

Cortes, P.C., Crutcher, R.M., Shepherd, D.S., Bronfman, L. (2008): *ApJ*, **676**, 464

Crutcher, R.M. (2008): *Astrophys. & Sp. Sci.* **313**, 141

Han, J.-L., Wielebinski, R. (2002): *Chinese J. Astron.* **2**, 293

Kronberg, P.P. (1994): *Rep. Prog. Phys.* **57**, 325

Laing, R.A., Bridle, A.H., Parma, P., Feretti, L., Giovannini, G., Murgia, M., Perley, R. (2008): *MNRAS*, **386**, 657

Thum, C., Wiesemeyer, H., Paubert, G., Navarro, S., Morris, D. (2008): accepted by PASP (2008arXiv0806.1666T)

Chapter 4

a) General

This chapter draws material from many heterogeneous sources. Thus, there is no single source that covers all aspects. But the following books contain most of the material.

Hachenberg, O., Vowinkel, B. (1982): *Technische Grundlagen der Radioastronomie* (Bibliographisches Institut, Mannheim) Chaps. 2, 3

Kingston, H., Blake, C. (1968): "Receivers", in *Radar Astronomy*, Evans, J. Hagfors, T. ed. (McGraw Hill, New York) pp. 465–496

Meeks, M.L. ed. (1976): *Astrophysics Part B: Radiotelescopes, Methods of Experimental Physics*, Vol. 12 B (Academic Press, New York) Chap. 3, p. 201

Rieke, G.H. (2002): *Detection of Light; From the Ultraviolet to the Submillimeter*, 2nd ed. (Cambridge University Press, Cambridge)

Stanimirovic, S., Altschuler, D., Goldsmith, P., Salter, C., eds. (2002): *ASP Conference Proceedings*, **278**, "Single-Dish Radio Astronomy: Techniques and Applications"

van Duzer, T., Turner, C.W. (1999): *Principles of Superconductive Devices and Circuits*, 2nd ed. (Prentice Hall:Upper Saddle River, NJ)

Discussions of Fourier Transforms and an introduction to noise analysis are found in:

Bracewell, R.N. (1986): *The Fourier Transform and its Application*, 2nd ed. (McGraw Hill, New York)

Davenport, W.B., Root, W.L. (1987): *An Introduction to the Theory of Random Signals and Noise* (Wiley-IEEE Press, New York)

Dicke, R.H. (1946): *Rev. Sci. Instrum.* **17**, 268

Jennison, R.C. (1961): *Fourier Transforms and Convolutions for the Experimentalist* (Pergamon, New York)

Lahti, B.T. (1965): *Signals, Systems and Communication* (Wiley, New York)

Middleton, D. (1996): *An Introduction to Statistical Communication Theory* (Wiley-Interscience-IEEE Press, New Jersey)

Papoulis, A., Pillai, S.U. (2002): *Probability, Random Variables and Stochastic Processes,* 4th ed. (McGraw Hill, New York)

Khinchin, A.I. (1949): *Mathematical Foundations of Statistical Mechanics* (Dover, New York)

Rice, S.O. (1954): "Mathematical Analysis of Random Noise", in *Selected Papers on Noise and Stochastical Processes*, Wax, N. ed. (Dover, New York)

Wiener, N. (1949): *Extrapolation, Interpolation and Smoothing of Stationary Time Series* (MIT Cambridge, MA)

b) Special

Abramowitz, M., Stegun, J.A. (1964): *Handbook of Mathematical Functions* (National Bureau of Standards, Washington, DC)

Allan, D.W. (1966): *proc. IEEE* **54**, 221

Blachman, N.M. (1966): *Noise and its Effect on Communication* (McGraw Hill, New York)

Blackman, R.B., Tukey, J.W. (1958): *The Measurement of Power Spectra* (Dover, New York)

Chang, I.C. (1976): *IEEE Trans. Sonics Ultrason.* SU **23**, 2

Recoquillon, C., Baudry, A., Begueret, Gauffre, S., Montignac, G. (2005): *The ALMA 3-bit, 4 Gsample/s, 2-4 GHz Input Bandwidth, Flash Analog-to-Digital Converter*, ALMA Memo No. 532

Whittaker, E.T., Watson, G.N. (1963): *A Course of Modern Analysis* (Cambridge University Press, Cambridge)

Chapter 5

Backer, D.C. (1988): "Pulsars", in *Galactic and Extragalactic Radio Astronomy*, Verschuur, G.L., Kellermann, K.I. eds. 2nd ed. (Springer, Heidelberg) p. 480

Baker, A.J., Glenn, J., Harris, A.I., Mangum, J.G., Yun, M.S. eds. (2007): *From Z Machines to ALMA: (Sub)millimeter Spectroscopy of Galaxies* **Conf. Ser. 75** (Astron. Soc of Pacific, San Francisco)

Ball, J.A. (1975): *Computations in Radio-Frequency Spectroscopy in Methods in Computational Physics*, Vol. 14, *Radio Astronomy*, Alder, B. ed. (Academic Press, New York)

Born, M., Wolf, E. (1965): *Principles of Optics* (Pergamon, Oxford)

Cole, T.W., Ables, J.G. (1974): *A. & A.* **34**, 149

Cooley, J.W., Tukey, J.W. (1965): *Math. Comp* **19**, 297

Cooper, B.F.C. (1976): "Autocorrelation Spectrometers", in *Astrophysics Part B: Radiotelescopes, Methods of Experimental Physics*, Vol. 12 B, Meeks, M.L. ed. (Academic Press, New York), p. 280

D'Addario, L. (1989): "Cross Correlators", in *Synthesis Imaging in Radio Astronomy*, Perley, R.A., Schwab, F.R., Bridle, A.H. eds. (ASP Conf. Ser. Vol. 6) p. 54

Darlington, S. (1964): *Bell Syst. Tech. J.* **43**, 339

Gershenzon, E.M., Gol'tsman, G.N., Gogidgze, I.G., Gusev, Y.P., Elant'ev, A.I., Karasik, B.S., Semenov, A.D. (1990): *Soviet Phys. Superconductivity* **3** (10), 1582

Goldsmith, P.F. ed. (1988): *Instrumentation and Techniques for Radio Astronomy* (IEEE Press, New York)

Griffin, M.J., Holland, W.S. (1988): *Int. J. Infrared & Millimeter Waves* **9**, 861

Hankins, T.H., Rickett, B.J. (1975): "Pulsar Signal Processing", in *Methods in Computational Physics*, Alder, B. et al. ed. (Academic Press, New York)

Hartogh, P., Osterschek, K. (1998): *SPIE* **2583**, 282

Harris, A.I., Zmuidzinas, J. (2001): *Review of Scientific Instruments* **72**, 1531

Hewish, A., Bell, S.J., Pilkington, J.D.H., Scott, P.F., Collins, R.A. (1968): *Nature* **217**, 709

Holland, W.S. et al. (1999): *MNRAS* **303**, 659

Jones, R.C. (1953): *J. Opt. Soc. America* **43**, 1

Kawamura, J. et al. (2002): *A. & A.* **394**, 271

Kerr, A.R., Feldman, M.J., Pan, S.-K. (1996): MMA Memo 161 *Receiver Noise Temperature, the Quantum Noise Limit and Zero-Point Fluctuations*, NRAO, Charlottesville, VA, USA

Lorimer, D., Kramer, M. (2004): *Handbook of Pulsar Astronomy*, (Cambridge University Press, Cambridge, UK)

Lyne, A.G., Graham–Smith, F. (2006): *Pulsar Astronomy* 3rd ed. (Cambridge Univ. Press, Cambridge, UK)

Manchester, R.N., Taylor, J.H. (1977): *Pulsars*, (Freeman, San Francisco, CA)

Mather, J.C. (1982): *Appl. Optics* **21**, 1125

Neshioka, S., Richards, P.L., Woody, D.P. (1978): *Appl. Optics* **17**, 1563

Oliver, B.M (1965): *Proc. IEEE* **53**, 436

Phillips, T.G., Woody, D.P. (1982): *Ann. Rev. A. & A.* **20**, 285

Phinney, E.S., Kulkarni, S.R. (1994): *Ann. Rev. A. & A.* **32**, 591

Tofani, G. (2005): in *Radio Astronomy from Karl Jansky to Microjansky*, Gurvits, L., Frey, S. Rawlings, S., eds. (EDP Sciences, Paris), p. 405

Stacey, G. et al. (2002): "Direct Detection Spectroscopy in the 350 µm Window: SPIFI on the JCMT" in *Proc. of Infrared & Submillimeter Space Astronomy*, (Giard, M. et al. eds.) *EAS Pub. Ser.* **4**, 419

Weinreb, S. (1963): *A Digital Spectral Analysis Technique and its Application to Radio Astronomy* MIT Tech. Rep. 412 (MIT, Cambridge, MA)

Chapter 6

Abramowitz, M., Stegun, I.A. (1964): *Handbook of Mathematical Functions* (National Bureau of Standards, Washington, DC)

Baars, J.W.M. (2007): *The Parabolic Reflector Antenna in Radio Astronomy and Communication* Astrophysics Space Science Library (Springer-Verlag, Heidelberg)

Gradshteyn, I.S., Ryzhik, I.M. (1965): *Tables of Integrals, Series and Products* (Academic Press, New York)

Jenkins, F.A., White, H.E. (2001): *Fundamentals of Optics*, 4th ed. (McGraw-Hill, New York)

Korn, G.A., Korn, T.M. (2000): *Mathematical Handbook for Scientists and Engineers* (Dover, New York)

Kraus, J.D. (1988): *Antennas*, 2nd ed. (McGraw Hill, New York)

Kühn, R. (1964): *Mikrowellen-Antennen* (VEB Verlag Technik, Berlin)

Pawsey, J.L., Bracewell, R.N. (1954): *Radio Astronomy* (Oxford University Press, Oxford)

Rossi, B. (1957): *Optics* (Addison_wesley, Reading, Mass.)

Silver, S. ed. (1949): *Microwave Antenna Theory and Design* (McGraw Hill, New York)

Slater, J., Frank, N. (1933): *Introduction to Theoretical Physics* (McGraw Hill, New York)

Stratton, J.A. (2007): *Electromagnetic Theory* (Wiley-Interscience-IEEE Press, New Jersey)

Chapter 7

a) General

Baars, J.W.M. (2007): *The Parabolic Reflector Antenna in Radio Astronomy and Communication* Astrophysics Space Science Library (Springer-Verlag, Heidelberg)

Christiansen, W.N., Högbom, J.A. (1985): *Radiotelescopes*, 2nd ed. (Cambridge University Press, Cambridge)

Heilmann, A. (1970): *Antennen I-III* (Bibliographisches Institut, Mannheim)

Kraus, J.D. (1986): *Radio Astronomy*, 2nd ed. (Cygnus-Quasar Books, Powell, Ohio)

Kühn, R. (1964): *Mikrowellen-Antennen* (VEB Verlag Technik, Berlin)

Meeks, M.L. ed. (1976): *Astrophysics Part B: Radiotelescopes, Methods of Experimental Physics*, Vol. 12 B (Academic Press, New York)

Meinke, H., Gundlach, F.N. eds. (1986): *Taschenbuch der Hochfrequenztechnik*, 4th ed. (Springer, Berlin, Heidelberg)

Rush, W.V.T., Potter, P.D. (1970): *Analysis of Reflector Antennas* (Academic Press, New York)

Silver, S. ed. (1949): *Microwave Antenna Theory and Design* (McGraw Hill, New York)

Stutzman, W.L., Thiele, G.A. (1981): *Antenna Theory and Design* (Wiley, New York)

b) Special

Bao, V.T. (1969): *Proc. IEEE* **116**, 195

Bracewell, R.N. (1962): "Radio Astronomy Techniques", in *Handbuch der Physik*, Vol. 54, Flügge, S. ed. (Springer, Berlin, Heidelberg) p. 42

Encrenaz, P.J., Penzias, A.A., Wilson, R.W. (1970): *Ap. J.* **160**, 1185

Goldsmith, P.F. (1994): *Quasioptical Systems: Gaussian Beam Quasioptical Propagation and Applications* (Wiley-IEEE Press, New York)

Höglund, B. (1967): *Acta Polytech. Scand. PH.* **48**

Keller, J.B. (1962): *J. Opt. Soc. Am.* **52**, 116

Love, A.W. ed. (1976): *Electromagnetic Horn Antennas* (IEEE Press, New York)

Mittra, R. ed. (1975): *Numerical and Asymptotic Techniques in Electromagnetics*, Topics Appl. Phys., Vol. 3 (Springer, Berlin, Heidelberg)

Napier, P.J. (1989): "The Primary Antenna Elements", in *Synthesis Imaging in Radio Astronmomy*, Perley, A. et al. ed. (ASP Conf. Ser. Vol. 9) p. 39

Nash, R.T. (1964): *IEEE Trans. Antennas Propag.* **12**, 918

Pearson, T.J., Readhead, A.C.S. (1984): *Ann. Rev. A. & A.* **22**, 97

Ricardi, L.J. (1977): *Proc. IEEE* **65**, 356

Rush, W.V.T., Sörensen, O. (1975): *IEEE Trans. Antennas Propag.* **23**, 414

Ruze, J. (1952): *Nuovo Cimento* **IX**, Suppl., 364

Ruze, J. (1966): *Proc. IEEE* **54**, 633

Stark, A.A., et al. (1992): *Ap. J. Supp.* **79**, 77

Tofani, G. (2005): in *Radio Astronomy from Karl Jansky to Microjansky*, Gurvits, L., Frey, S., Rawlings, S. eds. (EDP Sciences, Paris), p. 405

Chapter 8

Altenhoff, W.J. (1985): "The Solar System: (Sub)mm continuum observations" in ESO Conf. & Workshop Proc. No. 22, p. 591

Altenhoff, W.J., Baars, J.W.M., Downes, D., Wink, J. (1987): *A. & A.* **184**, 381

Bania, T.M., Rood, R.T., Wilson, T.L. (1994): "The Frequency Baseline Structure in the 100 m Telescope at 3.6 cm", MPIfR Electr. Rep. Nr. 75

Cernicharo, J. (1985): IRAM internal rept. (the ATM program)

Crovisier, J. (1978): *A. & A.* **70**, 43

Condon, J.J. (1974): *Ap. J.* **188**, 279.

Downes, D. (1989): "Radio Telescopes: Basic Concepts" in *Diffraction-Limited Imaging with Very Large Telescopes*, Alloin, D.M., Mariotti, J.M. NATO ASI Ser. Vol. 274 ed. (Kluwer, Dordrecht) p. 53

Emerson, D.T., Klein, U., Haslam, C.G.T. (1979): *A. & A.* **76**, 92

Johnstone, D. et al. (2000): *Ap. J. Supp.* **131**, 505

Kalberla, P.M.W., et al. (1980): *A. & A.* **82**, 274 & **106**, 190

Kalberla, P.M.W., Mebold, U., Reif, K. (1982): *A. & A.* **106**, 190

Kutner, M. L., Ulich, B. L. (1981): *Ap. J.* **250**, 341

Mauersberger, R. et al. (1989): *Astron. Astophys. Suppl.* **79**, 217

Motte, F., Bontemps, S., Schneider, N., Schilke, P., Menten, K.M., Broguierè, D. (2007) *A. & A.* **476**, 1243

Ryle, M. (1968): *Ann. Rev. A. & A.* **6**, 249

Schoenberg, E. (1929): "Theoretische Photometrie" in *Hdb. d. Astrophys.* Bd. II/1, Bottlinger, K.F. et al. ed. (Springer, Berlin)

Schwab, F.R. (1984): *Astron. J.* **89**, 1076

Schwarz, U.J. (1978): *A. & A.* **65**, 345

Serabyn, E., Weisstein, D.C., Lis, D.C., Pardo, J.R. (1998): *Appl. Optics* **37**, 12

Chapter 9

a) General

Born, M., Wolf, E. (1965): *Principles of Optics* (Pergamon, Oxford) Chap. X

Christansen, W.N., Högbom, J.A. (1985): *Radiotelescopes*, 2nd ed. (Cambridge University Press, Cambridge) Chaps. 5–7

Clark, B.G. (1979): "Digital Processing Methods for Aperture Synthesis Observations", in *Image Formation from Coherence Functions in Radio Astronomy*, van Schoonefeld, C. ed. (Reidel, Dordrecht) p. 113

Dutrey, A. ed. (2000): *IRAM Millimeter Interferometry Summer School 2* (IRAM, Grenoble, France)

Fomalont, E.B., Wright, M.C.H. (1974): "Interferometry and Aperture Synthesis" in *Galactic and Extragalactic Radio Astronomy*, Verschuur, G.L., Kellermann, K.I. ed. (Springer, New York, Heidelberg, Berlin) p. 256

Jenkins, F.A., White, H.E. (2001): *Fundamentals of Optics*, 4th ed. (McGraw-Hill, New York) Chap. 13

Mensa, D.L. (1991): *High Resolution Radar Cross-Section Imaging* (Artech House, Boston)

Perley, R.A., Schwab, F.R., Bridle, A.H., eds. (1989): *Synthesis Imaging in Radio Astronomy* (ASP Conf. Ser. Vol. 6)

Taylor, G.B., Carilli, C.L., Perley, R.A. eds. (1999): *Synthesis Imaging in Radio Astronomy II* (ASP Conf. Ser. Vol. 180)

Thompson, A.R., Moran, J.M., Swenson, G.W. (2001): *Interferometry and Synthesis in Radio Astronomy* 2nd ed. (Wiley, New York)

Steel, W.H. (1967): *Interferometry* (Cambridge University Press, Cambridge)

b) Special

Clark, B.G. (1980): *A. & A.* **89**, 377

Cohen, M.H. (1969): *Ann. Rev. A. & A.* **7**, 619

Cornwell, T.J., Evans, K.F. (1985): *A. & A.*, 77

Cornwell, T., Fomalont, E.B. (1989): "Self Calibration" in *Synthesis Imaging in Radio Astronomy*, Perley, R.A. et al. ed. (ASP Conf. Ser. Vol. 6) p. 185

Gull, S.F., Daniell, G.J. (1978): *Nature* **272**, 686

Hall, P.J. ed. (2005): *The Square Kilometer Aray: An Engineering Perspective* (Springer, Dordrecht)

Hamaker, J.P., Sullivan, J.D., Noordam, J.E. (1977): *J. Opt. Soc. Am.* **67**, 1122

Högbom, J. (1974): *A. & A. Suppl.* **15**, 417

Holdaway, M.A., Carilli, C., Laing, R.A. (2004): "Finding Fast Switching Calibrators for ALMA" *ALMA Memo 493*

Narayan, R., Nityananda, R. (1986): *Ann. Rev. A & A* **24**, 127 (MEM)

Pearson, T.J., Readhead, A.C.S. (1984): *Ann. Rev A. & A.* **22**, 97

Schwarz, U.J. (1977): "The Method CLEAN – Use, Misuse and Variations", in *Image Formation from Coherence Functions in Radio Astronomy*, van Schoonefeld, C. ed. (Reidel, Dordrecht) p. 261

Sramek, R.A. (1982): "Map Plane – UV Plane Relationships", in *Synthesis Mapping*, Thompson, A.R., D'Addario, L.R. ed. (NRAO, Green Bank) Chap. 2

Steel, W.H. (1967): *Interferometry* (Cambridge University Press, Cambridge) p. 40ff

Thompson, A.R. (1982): "Introduction and Basic Theory", in *Synthesis Mapping*, Thompson, A.R., D'Addario, L.R. ed. (NRAO, Green Bank) Chap. 1

Thompson, A.R. (1989): "The Interferometer in Practice", in *Synthesis Imaging in Radio Astronomy*, Perley, R.A., Schwab, F.R., Bridle, A.H. eds. (ASP Conf. Ser. Vol. 6) p. 11

Walker, R.C. (1999): "Very Long Baseline Interferometry" in *Synthesis Imaging in Radio Astronomy II*, Ed. G.B. Taylor, C. Carilli, R.A. Perley ASP Conference Series **180**, p. 433

Wernecke, S.J., D'Addario, L.R. (1976): *IEEE Trans. Computers* **C-26**, 351

Chapter 10

a) General

Bekefi, G. (1966): *Radiation Processes in Plasmas* (Wiley, New York)

Ginzburg, V.L., Syrovatskii, S.I. (1965): *Ann. Rev. A. & A.* **3**, 297

Ginzburg, V.L., Syrovatskii, S.I. (1969): *Ann. Rev. A. & A.* **7**, 375

Ginzburg, V.L. (1979): *Theoretical Physics and Astrophysics* (Pergamon Press, Oxford)

Green, R.M. (1985): *Spherical Astronomy* (Cambridge University Press, Cambridge) Chaps. 15 and 16

Jackson, J.D. (1975): *Classical Electrodynamics*, 2nd ed. (Wiley, New York)

Krügel, E. (2002): *The Physics of Interstellar Dust* (Inst. of Physics Press: Bristol UK)

Landau, L.D., Lifschitz, E.M. (1967): *Lehrbuch der Theoretischen Physik*, Vol. 2 (Akademie, Berlin)

Lang, K. (1974): *Astrophysical Formulae* (Springer, New York, Heidelberg, Berlin) Chap. I

Longair, M.S. (1981): *High Energy Astrophysics* (Cambridge University Press, Cambridge) Chaps. 3 and 18

Pacholczyk, A.G. (1970): *Radio Astrophysics* (Freeman, San Francisco, CA)

Pacholczyk, A.G. (1977): *Radio Galaxies* (Pergamon, Oxford)

Panofsky, W.K., Phillips, M. (1962): *Classical Electricity and Magnetism* (Addison-Wesley, Reading, MA)

Rybicki, G.B., Lightman, A.P. (1979): *Radiative Processes in Astrophysics* (Wiley, New York)

Tucker, W.H. (1975): *Radiation Processes in Astrophysics* (MIT, Cambridge, MA)

Unsöld, A. (1955): *Physik der Sternatmosphären*, 2nd ed. (Springer, Berlin, Heidelberg)

b) Special

Alfvén, H., Herlofson, N. (1950): *Phys. Rev.* **78**, 616 (Letter)

Altenhoff, W.J. et al. (1960): *Veröff. Sternwarte Bonn*, Nr. 59

Bennett, A.S. (1962): *Mem. Roy. Astron. Soc.* **68**, 163

Marrone, D.P., Moran, J.M., Zhao, J.-H., Rao, R. (2007): *Ap. J.* **654**, L57

Chrysostomou, A., Aitken, D.K., Jenness, T., Davis, C.J., Hough, J.H., Curran, R., Tamura, M. (2002): *A. & A.* **385**, 1014

Hildebrand, R. (1983): *Quarterly J. Roy. Astron. Soc.* **24**, 267

Hirschfield, J.L., Baldwin, D.E., Brown, S.C. (1961): *Phys. Fluids* **4**, 198

Kiepenheuer, K.O. (1950): *Phys. Rev.* **79**, 138 (Letter)

Kramers, H.A. (1923): *Philos. Mag.* **46**, 836

Meyer, P. (1969): *Ann. Rev. A. & A.* **7**, 1

Mezger, P.G., Wink, J.E., Zylka, R. (1990): *A. & A.* **228**, 95 (Appendix)

Oort, J.H., Walraven, T. (1956): *Bull. Astron. Inst. Netherlands* **12**, 285

Oster, L. (1959): *Z. f. Ap.* **47**, 169

Oster, L. (1961): *Rev. Mod. Phys.* **33**, 525

Scheuer, P.A.G. (1967): "Radiation" in *Plasma Astrophysics*, Sturrock, P.A. ed. Proc. Int. Sch. Phys. "Enrico Fermi", 39, (Academic Press, New York) p. 39

Schott, G.A. (1907): *Ann. Phys. 4. Folge* **24**, 635

Schott, G.A. (1912): *Electromagnetic Radiation* (Cambridge University Press, Cambridge)

Schwinger, J. (1949): *Phys. Rev.* **75**, 1912

Westfold, K.C. (1959): *Ap. J.* **130**, 241

Chapter 11

a) General

Backer, D.C. (1988): "Pulsars", in *Galactic and Extragalactic Radio Astronomy*, 2nd ed. Verschuur, G.L., Kellermann, K.I. eds. (Springer, Berlin) p. 480

Backer, D.C., Hellings, R.W. (1986): *Ann. Rev. A. & A.* **24**, 537

Bastian, T.S., Benz, A.O., Gary, D.E. (1998): *Ann. Rev. A. & A.* **36**, 131

Chevalier, R.A. (1994): "Supernovae and the Interstellar Medium", in *Supernovae*, Bludman, S.A. et al. eds. Les Houches Session 54 (North-Holland, Amsterdam)

Dulk, G.A. (1985): *Ann. Rev. A. & A.* **23**, 169

Faranaroff, B.L., Riley, J.M. (1974): *Month. Not. RAS* **167**, 31p

Green, R.M. (1985): *Spherical Astronomy* (Cambridge U.P., Cambridge)

Krüger, A. (1979): *Introduction to Solar Radio Astronomy and Radio Physics* (Reidel, Dordrecht)

Kundu, M.R. (1965): *Solar Radio Astronomy* (Interscience, New York)

Landau, L.D., Lifschitz, E.M. (1967): *Lehrbuch der Theoretischen Physik VI, Hydrodynamik* (Akademie, Berlin)

Lyne, A.G., Graham–Smith, F. (2006): *Pulsar Astronomy*, 3rd ed. (Cambridge University Press, Cambridge, UK)

McLean, D.J., Labrum, N.R. eds. (1985): *Solar Radiophysics* (Cambridge U.P., Cambridge)

Rees, M.J., Stoneham, R.J. (1982): *Supernovae: A Survey of Current Research* (Reidel, Dordrecht)

Reynolds, S.P. (1988): "Supernova Remnants" in *Galactic and Extragalactic Radio Astronomy*, 2nd ed. Verschuur, G.L. Kellermann, K.I. eds. (Springer, Berlin) p. 439

Scheuer, P.A.G. (1967): "Radio Galaxies and Quasi-Stellar Sources", in *Plasma Astrophysics*, Sturrock, P.A. ed. Proc. Enrico Fermi, 39, (Academic Press, New York) p. 262

Spitzer, L. (1978): *Physical Processes in the Interstellar Medium* (Wiley, New York)

Taylor, J.H., Stinebring D.R. (1986): *Ann. Rev. A. & A.* **24**, 285

Will, C.M. (1985): *Theory and Experiment in Gravitational Physics* (Cambridge U.P., Cambridge)

Woltjer, L. (1970): "Supernovae and the Interstellar Medium", in *Proc. IAU Symposium* No. 39, Habing, H.J. ed. (Reidel, Dordrecht) p. 229

Woltjer, L. (1972): *Ann. Rev. A. & A.* **10**, 129

Zheleznyakov, V.V. (1970): *Radio Emission of the Sun and Planets* (Oxford U.P., Oxford)

b) *Special*

Backer, D.C., Kulkarni, S.R., Heiles, C., Davis, M.M., Goss, W.M. 1982 *Nature* **300**, 615

Begelman, M.C., Blandford, R.D., Rees, M.J. (1984): *Rev. Mod. Phys.* **56**, 259

Carilli, C.L., Perley, R.A., Dreher, J.W., Lahey, J.P. (1991): *Ap. J.* **383**, 554

Chevalier, R.A. (1974): *Ap. J.* **188**, 501

Chevalier, R.A. (1975): *Ap. J.* **198**, 355

Chevalier, R.A. (1982): *Ap. J.* **259**, 302

Gary, D.E., Keller, C.U. eds. (2004): *Solar and Space Weather Radiophysics: Current Results and Future Developments* (Kluwer, Dordrecht)

Goodger, J.L., Hardcastle, M.J., Croston, J.H., Kassim, N.E., Perley, R.A. (2008) *Mon. Not. R. Astron. Soc.,* **386**, 337

Hewish, A., Bell, S.J., Pilkington, J.D.H., Scott, P.F., Collins, R.A. (1968): *Nature* **217**, 709

Hulse, R.A., Taylor, J.H. (1974): *Ap. J.* **195**, L51

Jaeger, J.K., Westfold, K. (1949): *Austral. J. Sci. Res.* **2**, 322

Kuzmin, A.D. et al. (1998): *A. & A. Suppl.* **127**, 355

Lorimer, D.R., Lyne, A.G., Anderson, B. (1995): *Month. Not. RAS* **275**, L15

Panagia, N., Felli, M. (1975): *A. & A.* **39**, 1

Panagia, N., Walmsley, C.M. (1978): *A. & A.* **70**, 411

Radhakrishnan, V., Cooke, D.J. (1969): *Astophys. Lett* **3**, 225

Scheuer, P.A.G., Redhead, A.C.S. (1979): *Nature* **277**, 182

Shklovsky, J.S. (1960): *Sov. Astron. AJ.* **4**, 243

Smerd, S.F. (1950): *Aust. J. Sci. Res.* **3 A**, 34

Taylor, J.H., Cordes, J.M. (1993): *Ap. J* **411**, 674

Taylor, J.H., Manchester, R.N., Lyne, A.G. (1993): *Ap. J Supp. Ser.* **88**, 529

Weiler, K.W., et al. (1986): *Ap. J.* **301**, 790

Wilson, T.L., Codella, C., Filges, L., Reich, W., Reich, P. (1997): *A. & A.* **327**, 1177

Wilson, T.L., Pauls, T.A. (1984): *A. & A.* **138**, 225
Woltjer, L. (1972): *Ann. Rev. A. & A.* **10**, 129
Zheleznyakov, V.V. (1970): *Radio Emission of the Sun and Planets* (Oxford U.P., Oxford)

Chapter 12

Einstein, A. (1916): *Verh. Dtsch. Phys. Ges.* **18**, 318
Spitzer, L. (1978): *Physical Processes in the Interstellar Medium* (Wiley, New York)

Chapter 13

a) General

Binney, J., Tremaine, S. (1987): *Galactic Dynamics* (Princeton University Press, Princeton)
Burton, W.B. (1988): "The Structure of Our Galaxy Derived from Observations of Neutral Hydrogen", in *Galactic and Extra-Galactic Radio Astronomy*, 2nd ed. Verschuur, G.L. Kellermann, K.I. eds. (Springer, New York, Heidelberg, Berlin) p. 295
Dickey, J.M., Lockman, F.J. (1990): *Ann. Rev. A. & A.* **28**, 215
Genzel, R. (1991): *Physics and Chemistry of Molecular Clouds in Galactic Interstellar Medium*, Burton, W.B. et al. ed. (Springer, Berlin)
Hartmann, D., Burton, W.B. (1997): *Atlas of Galactic Neutral Hydrogen*, (Cambridge Univ. Press, Cambridge)
Kalberla, P.M.W., Burton, W.B., Hartmann, D., Arnal, E.M., Bajaja, E., Morras, R., Pöppel, W.G.L. (2005): *A. & A.* **440**, 775
Kerr, F.J. (1968): "Radio Line Emission and Absorption by the Interstellar Gas", in *Stars & Stellar System VII*, Middlehurst, B.M. Aller, L.H. eds. (University of Chicago Press, Chicago, IL) p. 575
Kulkarni, S.R., Heiles, C. (1988): "Neutral Hydrogen and the Diffuse Interstellar Medium" in *Galactic and Extra-Galactic Radio Astronomy*, 2nd ed. Verschuur, G.L. Kellermann, K.I. eds. (Springer, New York, Heidelberg, Berlin) p. 95
Rogers, A.E.E., Dudevoir, K.A., Bania, T.M. (2007): *Astron. J.* **133**, 1625

b) Special

Bernstein, G. et al. (1994): *Astronom. J.* **107**, 1962
Burton, W.B., Liszt, H.S. (1978): *Ap. J.* **225**, 815
Celnik, W., Rohlfs, K., Braunsfurth, E. (1979): *A. & A.* **76**, 24
Elwert, G. (1959): *Ergeb. Exakten Naturwiss.* **32**, 1
Field, G.B. (1958): *Proc. IRE* **46**, 240
Gunn, J.E., Knapp, G.R., Tremaine, S.D. (1979): Astron. J. **84**, 1181
Kahn, F.D. (1955): "Gasdynamics of Cosmic Clouds", in *Proc. IAU Symposium Nr. 2* (North-Holland, Amsterdam)
Peters, H.E. et al. (1965): *Appl. Phys. Lett.* **7**, 34
Purcell, E.M., Field, G.B. (1956): *Ap. J.* **124**, 542
Tully, R.B., Fisher, J.R. (1976): *A. & A.* **54**, 661

Chapter 14

a) General

Chaisson, E.J. (1976): "Gaseous Nebulae and their Interstellar Environment", in *Frontiers of Astrophysics*, Avrett E.H. ed. (Harvard University Press, Cambridge, MA)

Dupree, A.K., Goldberg, L. (1970): *Ann. Rev. A. & A.* **8**, 231

Gordon, M.A. (1988): "HII Regions and Radio Recombination Lines", in *Galactic and Extra-Galactic Radio Astronomy*, 2nd ed. Verschuur, G.L. Kellermann, K.I. eds. (Springer, New York, Heidelberg, Berlin) p. 37

Gordon, M.A., Sorochenko, R.L. (2002): "Radio Recombination Lines, Their Physics and Astronomical Applications", *Astrophysics and Space Science Library*. **282**, (Kluwer Academic Publications: Dordrecht)

Lang, K.L. (1974): *Astrophysical Formulae* (Springer, New York, Heidelberg, Berlin) Chap. 2

Osterbrock, D.E. (1989): *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (Freeman, San Francisco, CA)

Rybicki, G.B., Lightman, A.P. (1979): *Radiative Processes in Astrophysics* (Wiley, New York)

Seaton, M.J. (1980): "Theory of Recombination Lines", in *Radio Recombination Lines*, Shaver, P.A. ed. (Reidel, Dordrecht) p. 3

Shaver, P.A. ed. (1980): *Radio Recombination Lines* (Reidel, Dordrecht)

Spitzer, L. (1978): *Physical Processes in the Interstellar Medium* (Wiley, New York)

b) Special

Balser, D.S. et al. (1994): *Ap. J.* **430**, 667

Brocklehurst, M. (1970): *Mon. Not. R. Astron. Soc.* **148**, 417

Brocklehurst, M., Leeman, S. (1971): *Astrophys. Lett.* **9**, 139

Brocklehurst, M., Seaton, M.J. (1972): *Mon. Not. R. Astron. Soc.* **157**, 179

Brown, R.L., Lockman, F.J., Knapp, G.R. (1978): *Ann. Rev. A. & A.* **16**, 445

Dupree, A.K. (1969): *Ap. J.* **158**, 491

Goldberg, L. (1968): "Theoretical Intensities of Recombination Lines", in *Interstellar Ionized Hydrogen*, Terzian, Y. ed. (Benjamin, New York) p. 373

Griem, H.R. (1967): *Ap. J.* **148**, 547

Höglund, B., Mezger, P.G. (1965): *Science* **130**, 339

Kardashev, N.S. (1959): *Astron. Zh.* **36**, 838 [English transl.: *Sov. Astron. A.J.* **3**, 813 (1960)]

Kantharia, N.G., Anantharamaiah, K.R., Payne, H.E. (1998): *Ap. J.* **506**, 758

Kurucz, R.L. (1979): *Ap. J. Supp.* **40**, 1

Martin-Pintado, J. et al. (1994): *A. & A.* **286**, 890

Martin-Hernandez, N.L., Vermeij, R., Tielens, A.G.G.M., van der Hulst, J.M., Peeters, E. (2002): *A. & A.* **389**, 286

Mehringer, D.M. et al. (1993): *Ap. J.* **412**, 684

Menzel, D.H. (1937): *Ap. J.* **85**, 330

Menzel, D.H., Pekeris, C.L. (1935): *Mon. Not. R. Astron. Soc.* **96**, 77

Mezger, P.G. (1980): "Helium Recombination Lines", in *Radio Recombination Lines*, Shaver, P.A. ed. (Reidel, Dordrecht)

Pankonin, V. (1980): "The Partially Ionized Medium Adjacent to HII Regions", in *Radio Recombination Lines*, Shaver, P.A. ed. (Reidel, Dordrecht) p. 111

Panagia, N. (1973): *Astron. J.* **78**, 929

Panagia, N., Walmsley, C.M. (1978): *A. & A.* **70**, 411

Pengelly, R.M., Seaton, M.J. (1964): *Month. Not. Roy. Astr. Soc.* **127**, 165
Rubin, R.H. (1985): *Ap. J. Supp.* **57**, 349
Strömgren, B. (1939): *Ap. J.* **89**, 529
Thum, C., Martin-Pintado, J., Quirrenbach, A., Matthews, H.E. (1998): *A. & A.* **333**, L63
Walmsley, C.M. (1990): *A. & A. Suppl.* **82**, 201
Wink, J.E., Wilson, T.L., Bieging, J.H. (1983): *A. & A.* **127**, 211

Chapter 15

a) *General*

Bingel, W.A. (1969): *Theory of Molecular Spectra* (Wiley, New York)
Flygare, W.H. (1978): *Molecular Structure and Dynamics* (Prentice-Hall, New York)
Hellwege, K.H. (1974): *Einführung in die Physik der Molekülen*, Heidelberger Taschenbücher, Bd. 146 (Springer Verlag, Berlin)
Gordy, W., Cook, R.L. (1970): *Microwave Molecular Spectra* 2nd ed. (Wiley, New York)
Herzberg, G. (1950): *Molecular Spectra and Molecular Structure*, Vol. I (Van Nostrand, New York)
Herzberg, L., Herzberg, G. (1960): "Molecular Spectra", in *Fundamental Formulas of Physics*, Menzel, D.H. ed. (Dover, New York) p. 465
Kroto, H.W. (1992): "Molecular Rotation Spectra" (Dover, New York)
Townes, C.H., Schawlow, A.H. (1975): "Microwave Spectroscopy" (Dover, New York)

b) *Special*

Anderson, T., Herbst, E., DeLucia, F.J. (1992): *ApJS* **82**, 405
Barrett, A.H. (1964): *IEEE Transactions on Military Electronics Special Issue on Radio Astronomy*, October 1964
Bauder, A. (1979): *J. Phys. Chem Ref Data* **8**, 583
Butner, H., Charnley, S.B., Ceccarelli, C., Rodgers, S.D., Pardo, J.R., Parise, B., Cernicharo, J., Davis, G.R. (2007): *ApJ* **659**, L137
Cernicharo, J., Polehampton, E., Goicoechea, J. (2007)): *ApJ* **657**, L21
DeLucia, F.J., Helminger, P., Kirchoff, W.H. (1972): *J. Phys. Chem Ref Data* **3**, 21
Gerlich, D., Windisch, F., Hlavenka, P., Plasil, R., Glosik, J. (2006): *Phil. Trans. R. Soc. A* **364**, 3007
Harmony, M.D., Lurie, V.W., Kuzokski, R.L., Schwendeman, R.H., Ramsey, D.A., Lovas, F.J. (1978): *J. Phys. Chem Ref Data* **8**, 619
Lees, R.M. (1973): *ApJ* **184**, 763
Lovas, F.J., Lutz, H., Dreizler, H. (1979): *J. Phys. Chem Ref Data* **8**, 1051
Lovas, F.J. (1992): *J. Chem phys. Ref Data* **21**, 181
Mangum, J.G., Wootten, A. (1993): *Ap. J. Suppl.* **89**, 123
Oka, T., Geballe, T.R., Goto, M., Usada, T., McCall, B.J. (2005): *ApJ* **632**, 882
Plummer, G.M., Herbst, E., DeLucia, F.J. (1987): *ApJ* **318**, 873
Tiemann, E. (1974): *J. Phys. Chem Ref Data* **3**, 259
Vastel, et al. (2006): *ApJ* **645**, 1198
Wilson, T.L., Henkel, C., Huettemeister, S. (2006): *A& A* **460**, 533
Wilson, T.L., Walmsley, C.M., Baudry, A. (1990) *A& A* **231**, 159

Chapter 16

a) General

Cook, A.H. (1977): *Celestial Masers* (Cambridge University Press, Cambridge)

Elitzur, M. (1992): *Astronomical Masers*, Astrophys. & Space Sci. Library Vol. 170, (Kluwer, Dordrecht)

Genzel, R. (1991): "Physics and Chemistry of Molecular Clouds" in *Galactic Interstellar Medium*, Burton, W.B. et al. eds. (Springer Verlag, Berlin) p. 275

Herbst, E. (1994): "Interstellar Chemistry in the Last Two Decades", in *The Structure and Content of Molecular Clouds*, Wilson, T.L., Johnston, K.J. eds. (Springer, Berlin) p. 29

Levy, E.H., Lunine, J.I. eds. (1993): *Protostars & Planets III* (University of Arizona Press, Tucson)

Lequeux, J. (2005): *The Interstellar Medium* (Springer Verlag, Berlin)

Mannings, V., Boss, A.P., Russell, S.S. eds. (2000): *Protostars and Planets IV* (Univ of Arizona Press: Tucson)

Reipurth, B., Jewett, D., Keil, K. eds. (2007): *Protostars and Planets V* (Univ of Arizona Press: Tucson) **PPV**

Sparke, L., Gallagher, J.S. III (2000): *Galaxies in the Universe: An Introduction* (Cambridge Univ. Press, Cambridge, UK)

Solomon, P.M., vanden Bout, P.A. (2005): *Ann. Rev. A. & A.* **43**, 677

Solomon, P.M., Sanders, D.B. (1980): "Giant Molecular Clouds as a Dominant Component of the Interstellar Medium of the Galaxy", in *Giant Molecular Clouds in the Galaxy*, Solomon, P.M., Edmunds, M.G. eds. (Pergamon, Oxford) p. 41

Spitzer, L., Jenkins, E.B. (1975): *Ann. Rev. A. & A.* **13**, 133

Spitzer, L. (1978): *Physical Processes in the Interstellar Medium* (Wiley, New York)

Stahler, S.W., Palla, F. (2005): *The Formation of Stars* (Wiley-VCH, New York)

Tielens, A.G.G.M. (2005): *The Physics and Chemistry of the Interstellar Medium* (Cambridge Univ. Press, Cambridge, UK)

Townes, C.H.: "Introduction to Radio and IR Studies of Molecular Clouds" in *The Structure and Content of Molecular Clouds*, Eds. T.L. Wilson, K.J. Johnston (Springer, Berlin) p. 1

Turner, B.E. (1988): "Molecules as Probes of the Interstellar Medium and of Star Formation" in *Galactic and Extra-Galactic Radio Astronomy*, 2nd ed. Verschuur, G.L., Kellermann, K.I. eds. (Springer, New York, Heidelberg, Berlin) p. 154

Wilson, T.L., Johnston, K.J. eds. (1994): *The Structure and Content of Molecular Clouds* (Springer, Berlin)

b) Special

Boboltz, D.A., Simonetti, J.H., Dennison, B., Diamond, P.J., Uphoff, J.A. (1998): *ApJ* **509**, 256

Churchwell, E., Hollis, J.M. (1983): *ApJ* **272**, 591

Combes, F., Wiklind, T. in *Highly Redshifted Radio Lines* (1998): *ASP Conference Series* **156**, Carilli, C.L. et al. eds. Astronomical Society of the Pacific: San Francisco

Crutcher, R.M. (2008): *Astrophys & Sp. Sci.* **313**, 141

Dame, T.M. (1993): *AIP Conf. Proc.* **278**, (Holt, S.S., Verter, F. eds.), p. 267

Deguchi, S. (1981): *ApJ* **249**, 145

Dishoeck, E.F. van, Blake, G.A., Draine, B.T., Lunine, J.I. (1993): "Protostellar and Protoplanetary Matter" in *Protostars and Planets III*, Levy, E.H., Lunine, J.I. eds. University of Arizona Press: Tucson, p. 163

Dishoeck, E.F. van, Blake, G.A. (1998): *Ann Rev A. & A.* **36**, 317

Evans, N.J. II (1999): *Ann Rev A. & A.* **37**, 313

Fich, M., Tremaine, S. (1991): *Ann Rev A. & A.* **29**, 409

Frerking, M.A., Langer, W.D., Wilson, R.W. (1982): *Ap. J.* **262**, 590

Galli, D., Walmsley, M., Goncalves, J. (2002): *A. & A.* **394**,275

Goldreich, P., Kylafis, N.D. (1982): *ApJ* **253**, 606

Gordon, V.D., McCarthy, M.C., Apponi. A.J., Thaddeus, P. (2002): *AP J Suppl* **138**, 297

Genzel, R., Stutzki, J. (1989): *Ann Rev A. & A.* **27**, 41

Herbst, E. (1999): "Molecules in Space and Molecular Spectroscopy", in *Molecular Chemistry & Physics in Space*, Wall, W.F., Carraminana, A., Carrasco, L., Goldsmith, P.F. eds. (Kluwer, Dordrecht, 1999), p. 329

Herbst, E. (2001): *Chem. Soc. Rev.* **30**, 168

Hildebrand, R. (1983): *Quarterly J. Roy. Astr. Soc.* **24**, 267

Hollenbach, D. (1988): *Astron Lett. Commun*. **26**, 191

Hollenbach, D.J., Tielens, A.G.G.M. (1999): *Rev. Mod. Phys.* **71**, 173

Hollis, J.M., Lovas, F.J., Jewell, P.R., Coudert, L.H. (2002): *ApJ* **571**, L59

Irvine, W.M. et al. (1985): "The Chemical State of Dense Interstellar Matter", in *Protostellar and Planets II* Black, D.C., Matthews, M.S. eds. (University Arizona Press) p. 579

Ivison, R.J. et al. (2002): *MNRAS* **337**, 11

Johnson, D.R., Lovas, F.J., Kirchoff, W.H. (1972): *J. Phys. Chem Ref Data* **1**, 1011

Kaufman, M.J., Wolfire, M.G., Hollenbach, D.J., Luhman (1999): *ApJ* **527**, 795

Kramer, C. et al. (1998): *A. & A.* **329**, L33

Lada, C.J., et al. (1994): *Ap. J.* **429**, 694

Lafferty, W.J., Lovas, F.J. (1978): *J. Phys. Chem Ref Data* **7**, 441

Langer, W.R. et al. (2002): "Chemical Evolution of Protostellar Matter", in *Protostars and Planets IV* Mannings, V. Boss, A.P., Russell, S.S. eds. (University Arizona Press), p. 20

Linke, R., Goldsmith, P.F. (1980): *Ap. J.* **187**, L67

Lovas, F.J. (1982): *J. Phys. Chem. Ref. Data* **11**

Lovas, F.J. (1992): *J. Chem Phys. Ref Data* **21**, 181

Maiolino, R. et al. (2007): *A. & A.* **472**, L33

Mauersberger, R. (1983): Diplomarbeit, Bonn University

Megeath, S.T., Wilson, T.L. (1997): *AJ* **114**, 1106

Menten, K.M. (1994): "Dense Molecular Gas in Star-Forming Regions – The Importance of Sub-millimeter Observations", in *The Structure and Content of Molecular Clouds*, Wilson, T.L., Johnston, K.J. eds. (Springer, Berlin) p. 150

Moran, J.M. (1976): "Radio Observations of Galactic Masers", in *Frontiers of Astrophysics*, Avrett, E.H. ed. (Harvard University Press, Cambridge, MA)

Mezger, P.G., Wink, J.E., Zylka, R. (1990): *A. & A.* **229**, 95

Olano, C.A., Walmsley, C.M., Wilson, T.L. (1988): *A. & A.* **196**, 194

Ossenkopf, V., Henning, T. (1994): *A. & A.* **291**, 943

Remijan et al. (2007): *Bull. Am. Astron. Soc.* **39**, 963

Schilke, P. (1989) unpublished Diplomarbeit University of Bonn

Schilke, P., Groesbeck, T.D., Blake, G.A., Phillips, T.G. (1997): *Ap J Suppl* **108**, 301

Shu, F.H., Adams, F.C., Lizano, S. (1987): *Ann. Rev. A. & A.* **25**, 23

Störzer, H., Stutzki, J., Sternberg, A. (1996): *A. & A.* **310**, 592

Watson, W.D. (1994): "Interpretations for Observations of Astronomical Masers" in *Structure and Content of Molecular Clouds* Wilson, T.L., Johnston, K.J. eds. (Springer Verlag, Berlin)

Zadelhoff, van G.J. et al. (2002): *A. & A.* **395**, 373

Index

ASTRONOMY AND ASTROPHYSICS LIBRARY

Series Editors: G. Börner · A. Burkert · W. B. Burton · M. A. Dopita
A. Eckart · T. Encrenaz · E. K. Grebel · B. Leibundgut
J. Lequeux · A. Maeder · V. Trimble