EXPERIMENTS IN MODERN PHYSICS

Second Edition

Adrian C. Melissinos
UNIVERSITY OF ROCHESTER

Jim Napolitano
RENSSELAER POLYTECHNIC INSTITUTE

ACADEMIC PRESS
An imprint of Elsevier Science

Amsterdam Boston London New York Oxford Paris San Diego
San Francisco Singapore Sydney Tokyo

Senior Publishing Editor	Jeremy Hayhurst
Senior Project Manager	Julio Esperas
Editorial Coordinator	Nora Donaghy
Product Manager	Anne O'Mara
Cover Design	Dick Hannus
Copyeditor	Charles Lauder, Jr.
Composition	Cepha Imaging Pvt. Ltd.
Printer	The Maple-Vail Book Manufacturing Group

This book is printed on acid-free paper. ♾

Academic Press
An imprint of Elsevier Science
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.academicpress.com

Academic Press
84 Theobald's Road, London WC1X 8RR, UK
http://www.academicpress.com

Academic Press
200 Wheeler Road, Burlington, Massachusetts 01803, USA
http://www.harcourt-ap.com

Library of Congress Catalog Card Number: 2002117796

International Standard Book Number: 0-12-489851-3

PRINTED IN THE UNITED STATES OF AMERICA
03 04 05 06 07 08 9 8 7 6 5 4 3 2 1

Εἰς μνήμην τοῦ πατρός μου

Κωνσταντίνου Ι. Μελισσινοῦ

To the memory of my father

Jiusto Napolitano

Contents

Appendices

Preface

In the nearly forty years since the first edition of this book was published, the fundamental concepts are of course unchanged while many of the details are radically different. This new edition attempts to maintain the emphasis on the fundamental importance of experimental physics and laboratory technique, while updating the equipment and tools used to set up the experiments and to acquire and analyze the data.

As much as possible, this revision is in keeping with the style of the original text. The importance of experimental investigation and sound laboratory technique, as a way for students to connect advanced physics topics to measurements carried out with their own hand, is emphasized. If anything, this approach is even more important than it was forty years ago. Curricula have focused more and more on "interactive" techniques in the introductory sciences, and the advanced laboratory is a primary way to extend this approach to upper level courses.

We have incorporated many of the changes that have occurred in experimental techniques. Chapter 3 collects topics in basic laboratory electronics (including some simple experiments with elementary circuits), as well as the somewhat more advanced topics of OpAmps, lock-in amplifiers, and computer interfaces. Chapter 4 focuses on lasers and optical instruments. Data analysis and presentation is generally carried out with the program MATLAB; analysis programs are available from the authors. Throughout the book, we make use of computers and computer-controlled hardware, as well as various commercial software packages, as illustrative options for building such experiments. Also, a collection of exercises suitable for homework or examinations is included in Appendix G.

New experiments have been added and the material has been reorganized. A number of new experiments in condensed matter have been introduced in Chapter 2, including measurements of the resistivity of metals using eddy currents, the Hall effect in bismuth, electrical, and thermal properties of diodes, and high T_c superconductors. Chapter 3 includes new experiments on Johnson noise, and chaos. Chapters 4 and 5 are completely new and several experiments involving lasers are discussed. These include classical experiments on diffraction and interferometry as well as a measurement of the Faraday effect and of Berry's phase. Chapters 6 and 7 have been updated and an experiment on saturation absorption spectroscopy has been introduced. The material on nuclear physics and nuclear techniques has been reorganized into Chapters 8 and 9 and some new measurements, including cosmic ray experiments and muon decay have been added.

Space limitations have forced us to drop several experiments, and other material, from the first edition. We have eliminated experiments on the photoelectric effect, thermionic emission, the Hall effect in semiconductors, Rutherford scattering, and velocity and particle identification measurements. Some detailed discussions of experimental techniques, such as the prism spectrograph and vacuum pumping, have also been removed.

One of the most dramatic developments since the first edition has been the use of computers for data analysis and presentation. Indeed, today there are a multitude of both commercial and free programs that run on a variety of platforms, all of which would be suitable for the experiments we describe here. In this text, for many cases, we have chosen to use the program MATLAB (http://www.mathworks.com/) to illustrate the analyses. The student version is inexpensive and well documented, and provides some sophisticated routines for things such as nonlinear fitting and data presentation. (Appendix B gives a brief introduction to the program.) However, we emphasize that all of the necessary tools, including plotting, linear fitting, and so forth, are easily accessible through any number of programs.

This revision is built on advanced laboratory courses at the University of Rochester and at Rensselaer Polytechnic Institute, as well as laboratory components of upper level lecture courses. Our students take part in interactive courses at the introductory level, and they extend this exposure with this advanced laboratory material as they continue their education. In many cases, the experiments are developed, built, and debugged by students who have already gone through a dedicated advanced laboratory course. In most cases, the data presented were acquired by students. These students are listed collectively in Appendix A.

We are grateful to many of our colleagues for their help and support. In particular, A.C.M. thanks Todd Blalock, Glen Hallit, and Craig Spencer, who were in charge of the "senior lab" in recent years. He also thanks Judy Mack for cheerful and efficient typing of early versions of the manuscript. J.N. thanks Toh-Ming Lu for his support of this course, Peter Persans for his efforts to teach and extend the laboratory and for contributing his notes on using MATLAB, and Tom Shannon for his maintenance of the equipment.

A.C.M., J.N.
Rochester, New York
Troy, New York

Preface from the First Edition

It is generally accepted that training in the sciences, especially at the undergraduate level, is not complete without a fair amount of laboratory experience. This is particularly true in physics where the basic freshman and sophomore courses are supplemented by concurrent laboratory exercises.

At the junior and senior level, however, laboratory training becomes more important and forms the subject of an independent course. Rather than simple laboratory exercises, the students now perform complete experiments and one could list the aims of the course as follows:

(a) To teach the student the methods and procedures of experimental physics at an advanced level; and to give him confidence in his own ability to measure physical entities and relationships between them.

(b) To familiarize the student with modern research equipment and its use; also to make him aware of the most basic techniques presently used in widely varying fields of physics.

(c) To convince the student that the material he studied and covered in his lecture courses can indeed be tested experimentally; and to give him the satisfaction of doing so himself.

On the other hand the real professional training for students who will become experimental physicists takes place in graduate school during their thesis work; this is a period of intensive involvement in research but within a highly specialized field. It therefore appears that the best opportunity for a broad look at the general experimental methods of physics still remains in the junior and senior laboratory courses.

The present text is an outgrowth of such a laboratory course given by the author at the University of Rochester between 1959 and 1963. It consisted of a one-year course with two 3-hour meetings in the laboratory and two 1-hour lecture meetings weekly; the students had access to the laboratory at all times and, in general, worked during hours of their own choice well in excess of the scheduled periods. The students worked in pairs, which in most cases provides a highly motivating and successful relationship.

The material included in this course was selected from those experiments in atomic and nuclear physics that have laid the foundation and provided the evidence for modern quantum theory. The experiments were set up in such a fashion that they could be completed in a two- to four-week period of normal work taking into account the other demands on the student's time. A frequent tendency of students (especially the more enthusiastic ones) is to become involved in experiments that are "almost original" or in setting up new experiments; this, however, requires construction of their own equipment and can result in considerable "gadgeteering" as well as leading to extended involvement, which a senior cannot afford. We found this to be a common trap eventually leading to frustration and discouragement with a student having only a "progress report" or a marginal result to show for one term of work.

For these reasons we used, whenever possible, commercial equipment, and all experiments were carefully tested before being handed over to the student. The emphasis was on the "physics" of the experiment and the interpretation of the results obtained; clearly, to obtain correct results the student had to properly adjust, use, and understand his equipment. Furthermore, a time limit could be set so that eight to ten different experiments could be completed in one academic year. This variety not only brings the student in contact with a broader segment of physics and of techniques, it also gives him the opportunity of a "fresh start" several times throughout the course; and, most important, it keeps the student continuously interested in spite of any setback or difficulty he may encounter in one or more experiments.

The experiments described in the first four chapters of this text are, in general, easier than the ones discussed later; each can usually be completed in a one-week period, and at the University of Rochester are performed in the second term of the junior year. This leaves then the two terms of the senior year for the more advanced experiments described in the later chapters. The various experiments have been grouped according to the basic physical principle rather than the special technique. For each experiment the underlying theoretical ideas are first introduced, then the experimental apparatus is described in considerable detail and, finally, the results obtained by the students are given and discussed. In this respect we believe that this text is not a "laboratory manual"; instead we have aimed at a fairly coherent presentation of experimental physics in spite of the limited and occasionally random selection of the experiments. We feel that our approach is similar to that of G. P. Harnwell and J. J. Livingood in their classic text "Experimental Atomic Physics," which appeared originally in 1933.

The reader may occasionally be surprised by the great detail with which we describe apparatus or special procedures for analysis of data. We have done so to assist those who may wish to set up a similar laboratory and because these are the details the student has usually to find out by himself; but also we believe that only through such detail can one acquire the real flavor of experimental physics. We have placed special emphasis on numerical results and on simple calculations, emphasizing the use of the correct units.

Contrary to accepted practice we have included only a minimum number of references; instead, we have given a selected bibliography to each subject through which the interested reader may find all pertinent information. It is, however, expected that the student is familiar or is concurrently taking a course on modern physics. The usual mathematical level of calculus is considered as a prerequisite and is freely used throughout.

As mentioned before, modern commercial equipment is used whenever practicable; this is the same type of equipment as used in present-day research and frequently is the basis for a successful teaching laboratory. It is true, however, that similar equipment can be obtained from several manufacturers and that special apparatus is preferably built in one's own shop. We do have on file the prints of all such special equipment and we will be glad to supply them on request.

The list of experiments in this text is not complete. For example, we have not included a discussion of "coherent scattering" (diffraction) experiments, of "electromagnetic spectrometers," and of "visual techniques" (bubble chamber, spark chamber, and nuclear emulsion) in spite of their successful performance by several students. We hope to be able to remedy these omissions in a future edition. We also realize that in some cases a better, or more educational, technique might be available for the experiments presented here. We would be grateful to our readers if they wish to indicate to us these alternatives.

In line with our original intention, all the data and results presented in this book were obtained by students of the "Senior Laboratory" of the University of Rochester and the appropriate credit is given in the text. The results presented here could not have been achieved without the support of the Physics Department of the University of Rochester; also major equipment was purchased through a grant from the United States Atomic Energy Commission and a matching funds grant from the National Science Foundation. As is always the case, whatever success this laboratory did enjoy is due to the combined efforts of many individuals, a large part of which was supplied by the participating students. It is a special pleasure to thank from here the graduate assistants during the 1959–1963 period, Drs. E. Griffin, J. Robbins, J. Mochel, and J. Reed, for their contributions to the laboratory. More than to anyone else the laboratory is indebted to Mr. F. L. Reynolds, who has been in charge of all technical matters and has kept the equipment in operating condition; I wish to express to him my personal gratitude for his friendship and for many helpful suggestions connected with this text. I also wish to acknowledge discussions with many of my colleagues in Rochester and, in particular, Dr. W. P. Alford, Dr. M. F. Kaplon, and Dr. R. E. Marshak.

In the preparation of the manuscript I benefited from the art work of Messrs. Yu-Chang Lee, W. Stinson, and J. Pinero; most of the manuscript was typed by Mrs. B. M. Marsh, and to all of them I express my appreciation for their excellent work. I am also indebted to the following of my colleagues for reading early parts of the manuscript and making many valuable suggestions and corrections: Dr. P. Baumeister on Chapter 2; Dr. T. Castner on Chapter 3; Dr. D. Cline on Chapter 5; Dr. R. Ellsworth on Chapter 6; Dr. L. Bradley on Chapter 7; Mr. C. Cook on Chapter 8; and Dr. J. Reed on Chapter 9. Still, however, the responsibility for all errors is mine and I would appreciate it if the readers could indicate them to me. Finally, I would like to thank my wife, Joyce, for her encouragement and assistance during the course of this work.

A.C.M.
Rochester, New York

CHAPTER 1

Experiments on Quantization

1.1. INTRODUCTION

A defining characteristic of present-day physics is that many of the quantities used to describe physical phenomena are quantized. That is, such quantities cannot take any one of a continuum of values, but are restricted to a set (perhaps an infinite set) of discrete values. Common examples are the intensity of radiation of the electromagnetic field, the energy of atomic systems, or the electric charge. Strong evidence for such quantization is obtained from experiments that will be described in this chapter:

(a) Millikan's experiment by which the charge on individual oil droplets is measured. The experiment shows that the charge is always an integer multiple of the smallest charge observed; this is identified with the charge of the electron.

(b) The Frank–Hertz experiment on the excitation by electron bombardment of atomic vapors. It is found that only for discrete bombarding energies is such excitation possible, and the first excited state of the mercury (Hg) atom is thus measured.

(c) A measurement of spectral lines in the visible. In particular the Balmer series of the hydrogen atom, as well as the more complicated spectra of sodium and mercury will be discussed.

All three experiments can be carried out with commercially available equipment from several manufacturers. For instance the Model AP-8210 "Millikan Oil Drop Apparatus" from PASCO Scientific (Roseville, CA) is a fully assembled system that yields excellent results. Two varieties of Millikan apparatus are available from Tel-Atomic, Incorporated (http://www.telatomic.com/). A Frank–Hertz tube with its oven can be obtained from ELWE Lehrsysteme (Creuilingen, Germany). Klinger Educational Products also offers a complete Frank–Hertz experimental setup. PASCO Scientific also markets a "Precision Student Spectrometer" Model SP-9268, which is fully equivalent to the spectrometer used to obtain the data described in Sections 1.5 and 1.6. Of course such an apparatus can also be built in-house, and we shall describe the apparatus and data-taking procedures in sufficient detail.

1.2. THE MILLIKAN OIL DROP EXPERIMENT

1.2.1. General

In 1909, R. Millikan reported a reliable method for measuring ionic charges. It consists of observing the motion of small oil droplets under the influence of an electric field. Usually the drops acquire a few electron charges and thus conventional fields impart to them velocities that permit isolation of a drop and continuous observation for a considerable length of time; further, the mass of the oil droplet remains almost constant (there is very slight evaporation) during these long observation times.

In principle, if we measure the force due to the electric field E,

$$F_e = qE = neE, \tag{1.1}$$

we can obtain ne; repeating this measurement for several (or the same) drops but with different values of the integer n, we can extract the charge of the electron e.

The electric force can be measured either by a null method — that is, by balancing the drop against the gravitational force — or, as will be described here, by observing the motion of the drop under the influence of both forces. Oil droplets in air, acted on by a constant force F, soon reach a terminal velocity given by Stokes' law,

$$F = 6\pi a\eta v, \tag{1.2}$$

where a is the radius of the (assumed spherical) droplet, η the viscosity of the air, and v the terminal velocity. To obtain the radius of the drop (needed in Eq. (1.2)) we observe the free fall of the drop; the gravitational force is

$$F_g = \frac{4}{3}\pi a^3(\rho - \sigma)g \tag{1.3}$$

with ρ and σ the density of air and oil and g the acceleration of gravity.

Schematically, as shown in Fig. 1.1, the apparatus consists of two parallel plates that can be alternatively charged to a constant potential $+V$, $-V$, or 0. The drop is then observed (with a telescope), and the time t it takes to travel through a distance d is measured. Let F_+ denote the force on a negatively charged drop with electric field up (time t_+, electric force aiding gravity) and F_- the force with electric field down (time t_-, electric force opposing gravity). Then

$$\begin{aligned} F_\pm &= \pm ne(V/s) - \frac{4}{3}\pi a^3(\rho - \sigma)g = 6\pi a\eta d\left(1/t_\pm^{(n)}\right) \\ F_0 &= -\frac{4}{3}\,\pi a^3(\rho - \sigma)g = 6\pi a\eta d(1/t_0), \end{aligned} \tag{1.4}$$

where the sign conventions hold if t is considered >0 when the drop moves up, and $t < 0$ when it is moving down (recall that e is negative).

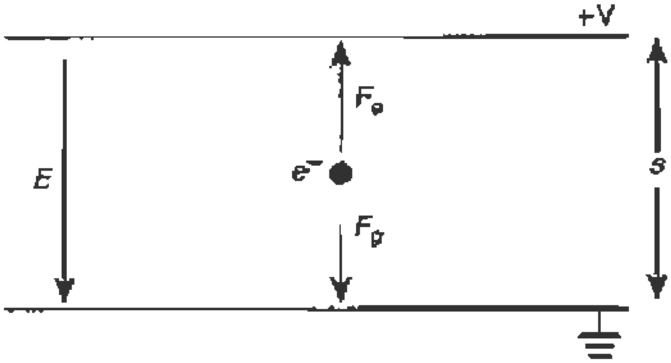


FIGURE 1.1 Forces on a charged oil drop between the plates of a Millikan apparatus.

A convenient method of analysis is to write Eq. (1.4) as

$$
\begin{aligned}
\frac{1}{t_{\pm}^{(n)}} &= \pm An - B \qquad & A &= \frac{Ve}{6s\pi a\eta d} \\
\frac{1}{t_0} &= -B \qquad & B &= \frac{2}{9}\frac{a^2(\rho - \sigma)g}{\eta d}
\end{aligned}
\tag{1.5}
$$

so that A and B can be easily determined.

Indeed a plot of $1/t_{\pm}^{(n)}$ against n reveals the linear relationship and the fact that only *integer* values of n appear, proving that the drop has acquired one, two, three, or more electric charges of value e, and never a fraction of that value. Thus we have clear evidence that the ionic charge picked up by the oil drops is *quantized*. Furthermore, the absolute value of this minimal electric charge is in good agreement with inferred measurements of the charge carried by the atomic electrons,[1] and therefore is accepted as the most accurate value of the charge of the electron.

1.2.2. The Experiment

The apparatus used in this laboratory (Fig. 1.2) consists of two parallel brass plates 1/4 in. thick and approximately 2 in. in diameter, placed in a lucite cylinder held apart by three ceramic spacers 4.7 mm long. This assembly is in turn enclosed in a cylindrical brass housing with provisions for electrical connections and containing two windows, one for illumination of the drops and one for observation. The top plate has a small hole in its center for the admission of the oil drops, which are produced by spraying oil with a regular atomizer.

To charge the plates, a 500-V DC power supply and a reversing switch are used, the plates are shunted by a 50-MΩ resistor to prevent them from remaining charged when the switch is open. For observations a 10-cm focal length microscope is used (Cenco 72925), while illumination is provided by a Mazda 1017-W lamp and condensing lens. To avoid convection currents inside the apparatus, a heat-absorbing filter (Corning infrared-absorbing) is placed in the illuminating beam.

The plates should be made perpendicular to the gravitational field by means of the three leveling screws at the base of the apparatus and a level

[1] As in e/m experiments, shot noise measurements, etc.

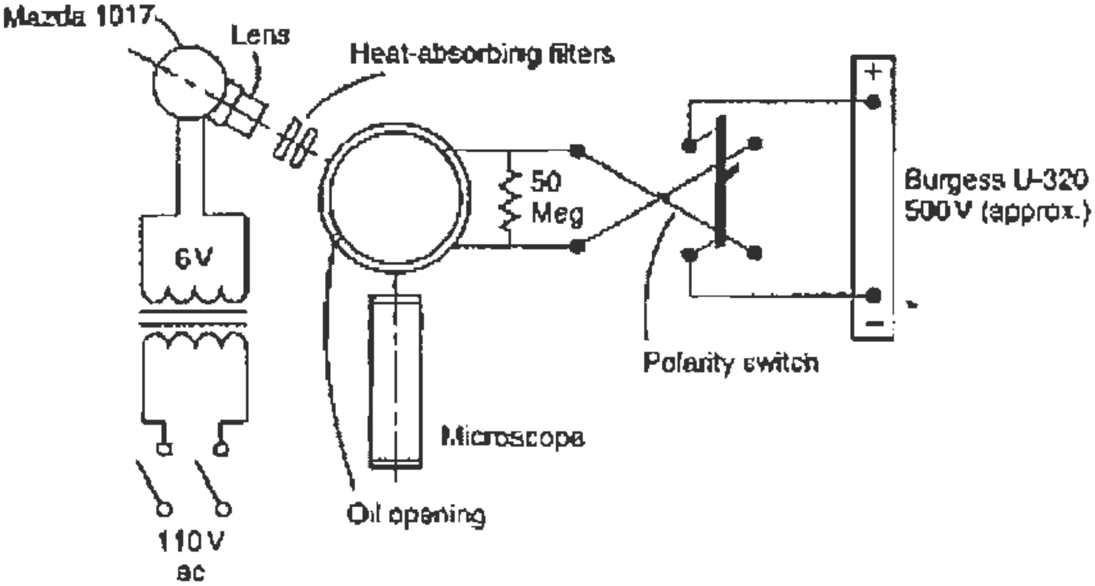


FIGURE 1.2 Millikan oil drop experiment schematic of the apparatus.

placed on the top plate. Being a cosine error, the deviation introduced by an angular displacement of the gravitational component from perpendicular by 8° is 1%. A value for the plate spacing *s* may be obtained by using the stage micrometer. The micrometer should be focused on a wire inserted in the oil hole in the center of the top plate, and the cross hair of the micrometer should be moved along the length of the wire. Several measurements should be taken and their results averaged.

The velocities are determined by measuring with a stopwatch the time required for the droplet to cover a specified number of divisions of the microscope scale. Care must be taken to avoid drafts and vibrations in the vicinity of the apparatus: for that reason and because of Brownian motion, the drop may wander or be displaced out of the field of the microscope. It may then be necessary to *reposition* the microscope between measurements on a single drop. Moreover, the drop should be kept in focus to avoid parallax errors.

Both the microscope and the light source may be adjusted by viewing a small wire inserted in the oil hole. The light should be adjusted so that the focal point is somewhat ahead or behind the wire and the wire is more or less evenly illuminated. To light the scale, a small light is placed next to the slit just ahead of the eyepiece of the microscope. The actual distance to which a scale division corresponds may be found by using a microscope slide on which a subdivided millimeter scale has been scratched.[2] The eyepiece focus of the microscope should not be changed during a run, since moving the eyepiece changes the effective distance of the scale. (To bring the drop back into focus the entire microscope should be moved.)

It is important to be *sparing* in the amount of oil sprayed into the chamber. In addition to gumming up the interior more quickly, large quantities create so many particles in the microscope field that without excessive eyestrain it is virtually impossible to single out and follow a single droplet.

Under the influence of gravity, droplets will fall at various limiting speeds. If the plates are charged, some of the drops will move down more rapidly, whereas others will reverse their direction of motion since in the process of spraying some drops become positively charged and others negatively charged. By concentrating on one drop that can be controlled by the field, and manipulating the sign of the electric field so that this particular drop is retained, it is possible to remove all other drops from the field. The limiting velocity is reached very quickly and the measurement should be started near the top or bottom of the plate. Measurement should be completed before the drop has reached a point in its travel where application of the reverse potential is insufficient to save the drop from being "gobbled up."

The density in air of the oil used was[3] 0.883±0.003 g/cm^3. It is desirable to take measurements in the shortest possible time since, as previously mentioned, the mass of the drop changes through evaporation.

It is also important to make measurements on as many different charges on the same or different drops as possible. Thus after four or five measurements of $t_+^{(n)}$, $t_-^{(n)}$, and t_0 have been taken, the charge on the drop must be changed; this is accomplished by bringing close to one of the windows a radioactive source (10 to 100 μCi of Co^{60} will do).[4] The droplet should be brought close to the top plate and allowed to fall with the *field off*; on its way down it will sweep up a few ions created by the source. This can be checked by occasionally turning the field on to see whether the charge has changed; rarely will a drop pick up any charge when the field is on.

The power supply voltage should be checked with a 1% digital multimeter (DMM); microscope calibration should be checked before and after

[2] Note that the focal length of the microscope must not be changed, but instead the slide should be brought into the focal plane.

[3] This may be found by a simple measurement.

[4] Ci ≡ Curie = 3.7×10^{10} disintegrations per second.

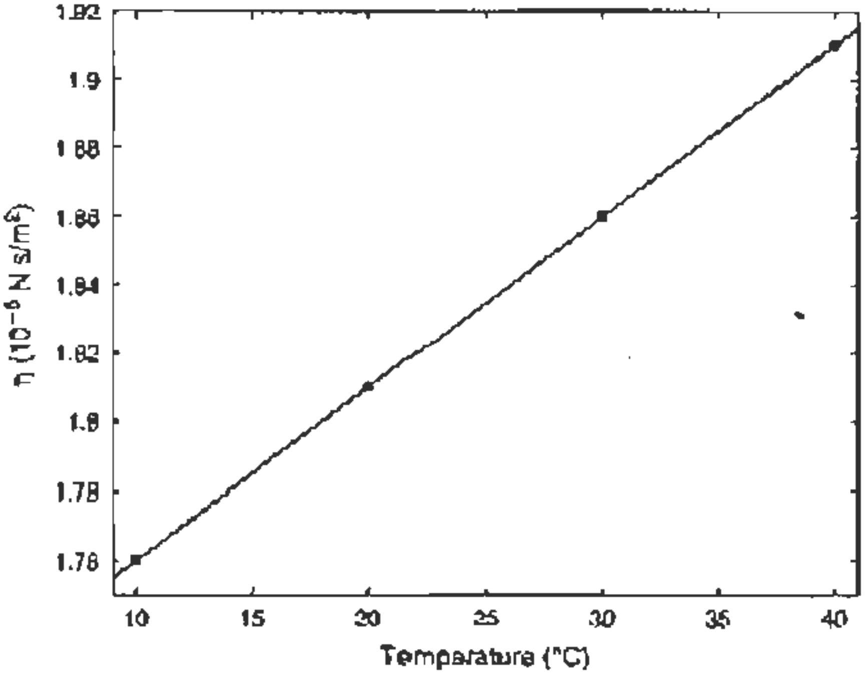


FIGURE 1.3 Viscosity of dry air as a function of temperature. The data points are taken from D. Pnueli and C. Gutfinger, *Fluid Mechanics*, Cambridge Univ. Press, Cambridge, UK, 1992, Table B-1. These points are fitted to a second-order polynomial to interpolate to the temperature in the laboratory.

the measurements. The same holds true for air temperature and pressure, which are needed for a correction to Stokes' law.

Indeed, when the diameter of the drop is comparable to the mean free path in air, the viscosity η in Eq. (1.2) should be replaced by[5]

$$\eta(T) = \eta_0(T)\left[1 + \frac{b}{aP}\right]^{-1}, \tag{1.6}$$

where $\eta_0(T)$ is the viscosity of air as a function of T (Fig. 1.3), $b = 6.17 \times 10^{-6}$, P is the air pressure in centimeters of mercury, and a is the radius of the drop in meters (on the order of 10^{-6} m). In analyzing the data it is convenient to calculate a_0 by letting $\eta = \eta_0(T)$ in the second of

[5] This formula, alternatively parameterized with $b/P = A\ell$, where ℓ is the mean-free path of the air molecules, was the subject of much research by Millikan and many others. See, for example, R. A. Millikan, *Phys. Rev.* 22, 1 (1923). Our value for b is taken from Y. Ishida, *Phys. Rev.* 21, 550 (1923), Table I.

Eqs. (1.5); a_0 is then inserted in Eq. (1.6) to obtain $\eta(T)$ and thus a more accurate value for a.

1.2.3. Analysis of the Data

Table 1.1 is a sample of data obtained by a student. Two drops were used and several charges were measured; for each charge six measurements were performed and averaged, with the results shown in Fig. 1.4. The drop radius a was determined from the average values of $1/\bar{t}_0$. The viscosity η uses the correction from Eq. (1.6). Values of n that give consistent values for $A = [(1/t_+) - (1/t_-)]/2n$ were identified. The pertinent parameters for these data were

Distance of fall	$d = 7.63 \times 10^{-4}$ m
Temperature	$T = 25°\mathrm{C}$
Pressure	$P = 76.01$ cm Hg
Density	$\rho' = \rho - \sigma = 882$ kg/m^3
Potential	$V = 500$ V
Plate separation	$s = 4.71 \times 10^{-3}$ m

TABLE 1.1 Data from the Millikan Oil Drop Experiment

$\bar{t}_0$	$\bar{t}_-^{(n)}$	$\bar{t}_+^{(n)}$	n	$\dfrac{(1/t_+) - (1/t_-)}{2n}$
		Drop 1		
−27.9	+8.69	−5.65	1	−0.146
−29.6	+1.36	−1.18	5	−0.158
−28.2	+3.66	−3.00	2	−0.152
−29.3	+0.75	−0.716	9	−0.152
−29.4	+2.35	−1.97	3	−0.155
$\Rightarrow a = 4.66 \times 10^{-7}$ m			$\eta = 1.58 \times 10^{-5}\ \mathrm{N \cdot s/m^2}$	
		Drop 2		
−24.22	+3.98	−3.071	2	−0.144
−25.75	+9.73	−5.65	1	−0.140
−25.4	+2.5	−2.12	3	−0.145
−25.22	+9.67	−5.42	1	−0.144
−25.22	+4.1	−3.07	2	−0.143
−24.4	+1.73	−1.73	4	−0.144
−24.4	+9.95	−6.02	1	−0.133
$\Rightarrow a = 5.04 \times 10^{-7}$ m			$\eta = 1.60 \times 10^{-5}\ \mathrm{N \cdot s/m^2}$	

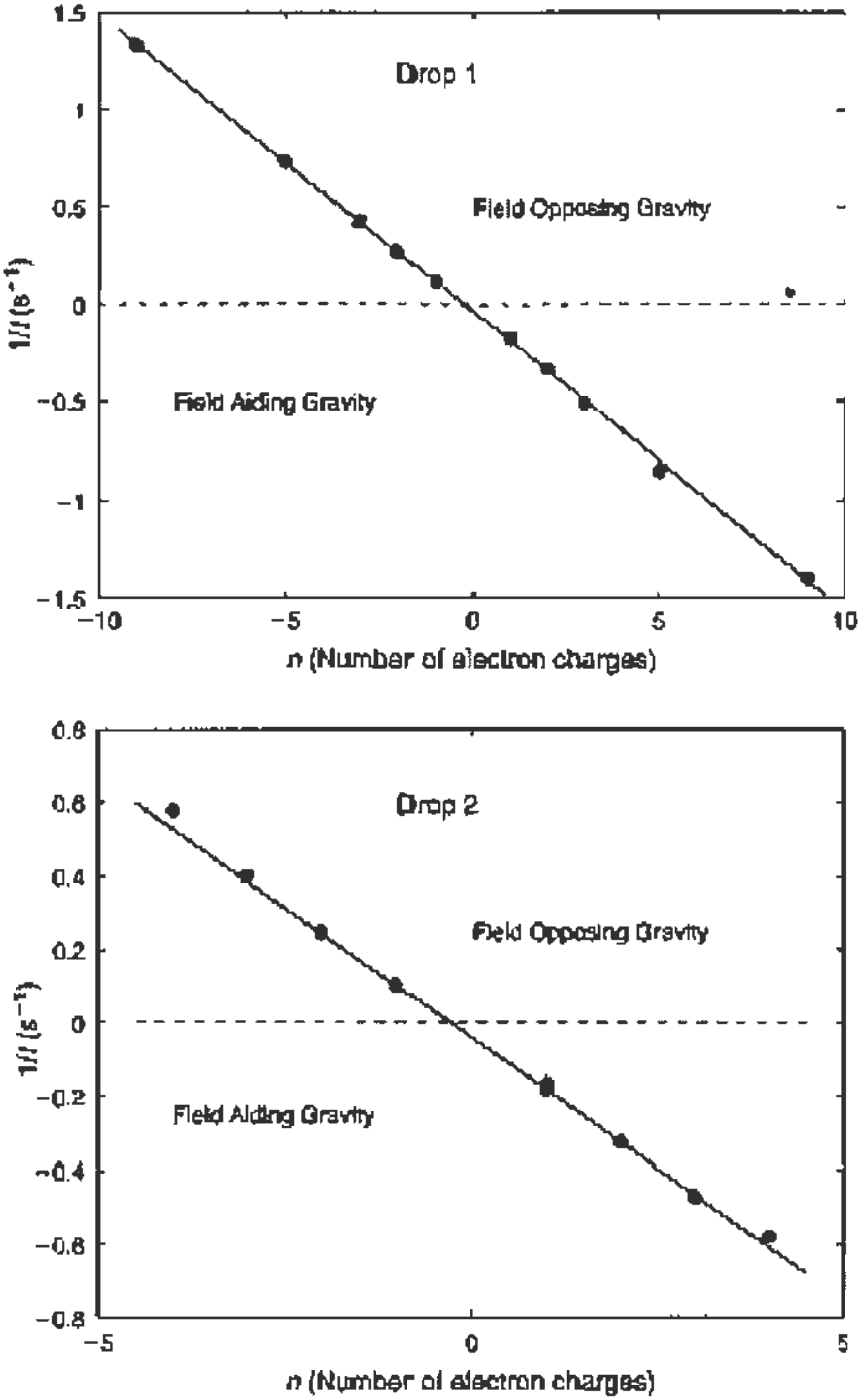


FIGURE 1.4 Plots of $1/t_+$ and $1/t_-$ versus n where n is an integer. Negative values of n are used to represent the data taken with the electric field pointing downward (i.e., t_+). The data are from Table 1.1.

Averaging the appropriate columns in Table 1.1 (See Eq. (1.5)) we find that

$$A_1 = -0.1526 \pm 0.0046\ \mathrm{s}^{-1} \qquad B_1 = 0.0346 \pm 0.0009\ \mathrm{s}^{-1}$$

$$|e| = (1.52 \pm 0.05) \times 10^{-19}\ \mathrm{C}$$

$$A_2 = -0.1419 \pm 0.0042\ \mathrm{s}^{-1} \qquad B_2 = 0.0401 \pm 0.0010\ \mathrm{s}^{-1}$$

$$|e| = (1.55 \pm 0.05) \times 10^{-19}\ \mathrm{C},$$

where the values of e are calculated using the value of A and the drop radius as obtained from the value of B. They are in good agreement[6] with the accepted value

$$|e| = 1.602 \times 10^{-19}\ \mathrm{C}.$$

Errors on A and B are simply taken to be the standard deviation of the set of measurements. (See Chapter 10.) The data are plotted in Fig. 1.4 along with the straight lines predicted by Eq. (1.5) using the values of A and B derived above.

The realization that the elementary (hadronic) particles are composites of quarks that have electric charge of $\frac{1}{3}$ or $\frac{2}{3}$ of the electron's charge led to a revival of the Millikan experiment. Automated versions[7] of the experiment have been built and run for a long time without revealing any such fractional charges.

1.3. THE FRANK–HERTZ EXPERIMENT

1.3.1. General

From the early spectroscopic work it was clear that atoms emitted radiation at discrete frequencies; from Bohr's model the frequency of the radiation ν is related to the change in energy levels through $\Delta E = h\nu$. Further experiments demonstrated that the absorption of radiation by atomic vapors also occurred only for discrete frequencies.

[6] It is seen that in this special case (partly because of the low voltage), the diameter of the drops is so small that the correction to the Stokes equation, i.e., Eq. (1.6), is considerable (about 7%).

[7] See, for example, N. Mar *et al.*, *Phys. Rev. D* 53, 6017 (1996).

It is then to be expected that the transfer of energy to atomic electrons by any mechanism should always be in discrete amounts[8] and related to the atomic spectrum through the equation given above. One such mechanism of energy transfer is by the inelastic scattering of electrons from the entire atom. If the atom that is bombarded does not become ionized, and since little energy is needed for momentum balance, almost the entire kinetic energy of the bombarding electron can be transferred to the atomic system.

J. Frank and G. Hertz in 1914 set out to verify these considerations, namely that (a) it is possible to excite atoms by low-energy electron bombardment, (b) that the energy transferred from the electrons to the atoms always had discrete values, and (c) that the values so obtained for the energy levels were in agreement with the spectroscopic results.

The necessary apparatus consists of an electron-emitting filament and an adequate structure for accelerating the electrons to a desired (variable) potential. The accelerated electrons are allowed to bombard the atomic vapor under investigation, and the excitation of the atoms is studied as a function of accelerating potential.

For detecting the excitation of the atoms in the vapor it is possible to observe, for example, the radiation emitted when the atoms return to the ground state, the change in absorption of a given spectral line, or some other related phenomenon; however, a much more sensitive technique consists of observing the electron beam itself. Indeed, if the electrons have been accelerated to a potential just *equal* to the energy of the first excited level, some of them will excite atoms of the vapor and as a consequence will lose almost all their energy; if a small retarding potential exists before the collector region, electrons that have scattered inelastically will be unable to overcome it and thus will not reach the anode.

These conditions are created in the experimental arrangement by using two grids between the cathode and collector. When the potentials are distributed as in Fig. 1.5a, the beam is accelerated between the cathode and grid 1; then it is allowed to drift in the interaction region between the two grids and finally must overcome the retarding potential between grid 2 and the anode. When the threshold for exciting the first level is reached, a sharp decrease in electron current is observed, proportional to the number of collisions that have occurred (product of the atomic density and cross section). When the threshold of the next level is reached, a further dip in the collector current will be observed. These current decreases (dips) are superimposed

[8] When they remain bound after the collision.

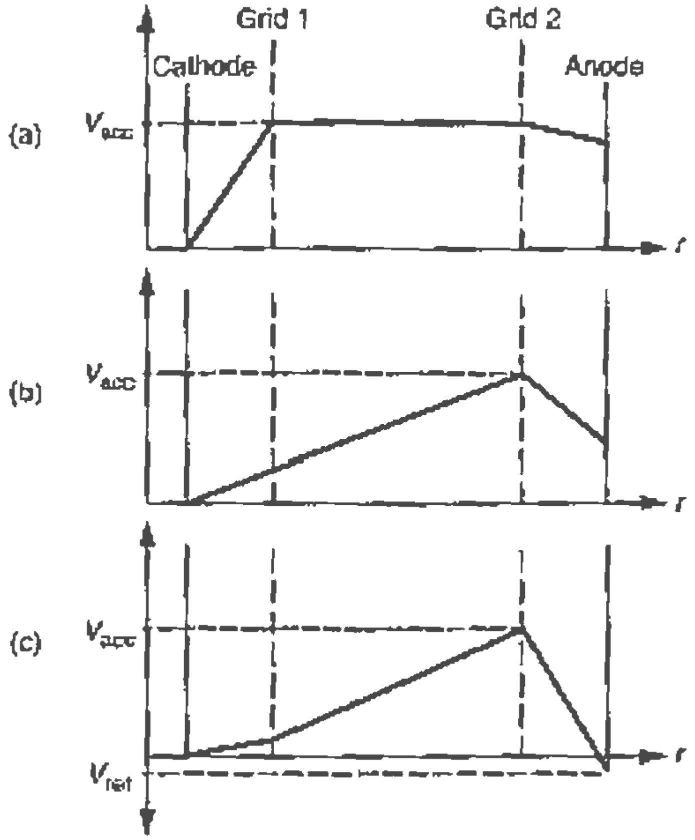


FIGURE 1.5 Different configurations of the potential in a Frank–Hertz arrangement: (a) For observation of a single excitation, (b) for observation of a multiple excitation, and (c) for measuring the ionization potential.

on a monotonically rising curve; indeed the number of electrons reaching the anode depends on V_{acc}, inasmuch as it reduces space charge effects and elastic scattering in the dense vapor. In addition, the dips are not perfectly sharp because of the distribution of velocities of the thermionically emitted electrons, and the energy dependence of the excitation cross section.

An alternate distribution of potentials is shown in Fig. 1.5b, where V_{acc} is applied at grid 2 so that an electron can gain further energy after a collision in the space between the two grids. In this case when V_{acc} reaches the first excitation potential, inelastic collisions are again possible and the decrease in electron current is observed at the anode; when, however, V_{acc} reaches a value twice that of the first excitation potential, it is possible for an electron to excite an atom halfway between the grids, lose all its energy, and then gain anew enough energy to excite a second atom and reach grid 2 with practically zero energy. Thus it is not able to overcome the retarding potential to reach the anode, giving rise to a second dip in the current.

The advantage of this setup is that the current dips are much more pronounced, and it is easy to obtain fivefold or even larger multiplicity in the excitation of the first level. However, it is practically impossible to observe the excitation of higher levels. As before, a slight retarding potential is applied between grid 2 and the anode, and an accelerating potential between the cathode and grid 1, sufficient to overcome space charge effects and to provide adequate electron current. It is evident that the density of the atomic vapor through which the electron beam passes greatly affects the observed results. Low densities result in large electron currents but very small dips; in contrast, high density has as a consequence weaker currents but proportionally larger dips. When mercury vapor is used, adjustment of the tube temperature provides control of the density.

Another important point is that in principle the experiment must be performed with a monatomic gas; since if a molecular vapor is bombarded, it is possible for the electrons to transfer energy to the molecular energy levels which form almost a continuum. Some of the preferred elements for the Frank–Hertz experiment are mercury, neon, and argon.

The same apparatus can be used for the measurement of the ionization potential—that is, the energy required to remove an electron completely from the atom. In this case, instead of observing the bombarding electron beam, it is easier to detect the ions that are formed. The distribution of potentials is as shown in Fig. 1.5c, where the anode is made slightly negative with respect to the cathode; no electrons can then reach the anode, which becomes an ion collector. The accelerating potential is increased until a sharp rise in the ion current measured at the anode is observed.

In both types of measurements the values obtained for the accelerating potential have to be corrected for the contact potential difference (cpd) between cathode and anode.[9] If in the excitation experiment the same level has been observed two or more times, however, the potential difference between adjacent peaks is an exact measure of the excitation energy, since the contact potential difference shifts the whole voltage scale. Once the excitation energy has been found the contact potential difference is given by the difference between this true value and the first peak; in turn the

[9] Briefly this is because the "work function" for the metal of which the anode is made is usually higher than that of the cathode. The work function is a measure of the "ionization potential" of the metal, that is, of the energy needed to extract an electron from it.

contact potential difference so found can be used to correct the ionization potential measurement.

1.3.2. The Experiment

In this laboratory a mercury-filled tube made by the Leybold Company (55580) was used, the electrode configuration is shown in Fig. 1.6, and the circuit diagrams for the measurement of excitation and of ionization potential are given in Figs. 1.7a and 1.7b, respectively.

As seen in the circuit diagram, grid 1 is operated in the neighborhood of 1.5 V, and the retarding potential is of the same order. The anode currents are on the order of 10^{-9} A and are measured either with a Keithley 610B electrometer or with a high-input impedance digital multimeter, for instance, Hewlett–Packard 34401A; adequate shielding of the leads is required to eliminate AC pickup and induced voltages. The diagram of Fig. 1.7a uses the distribution of potentials shown in Fig. 1.5b, and the accelerating voltage can be measured with a DMM in steps of 0.1 V.

The Frank–Hertz tube is placed in a small oven, which is heated by line voltage through a variac; it should be operated in the vicinity of 200°C for the excitation curve and between 100 and 150°C for the ionization curve. To measure the temperature a copper-constantan thermocouple should be

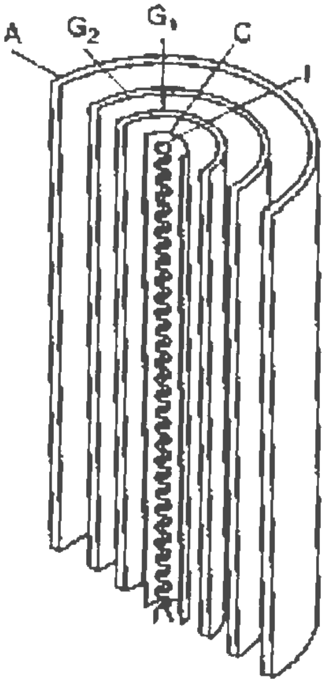


FIGURE 1.6 Sketch of a cylindrical Frank–Hertz tube.

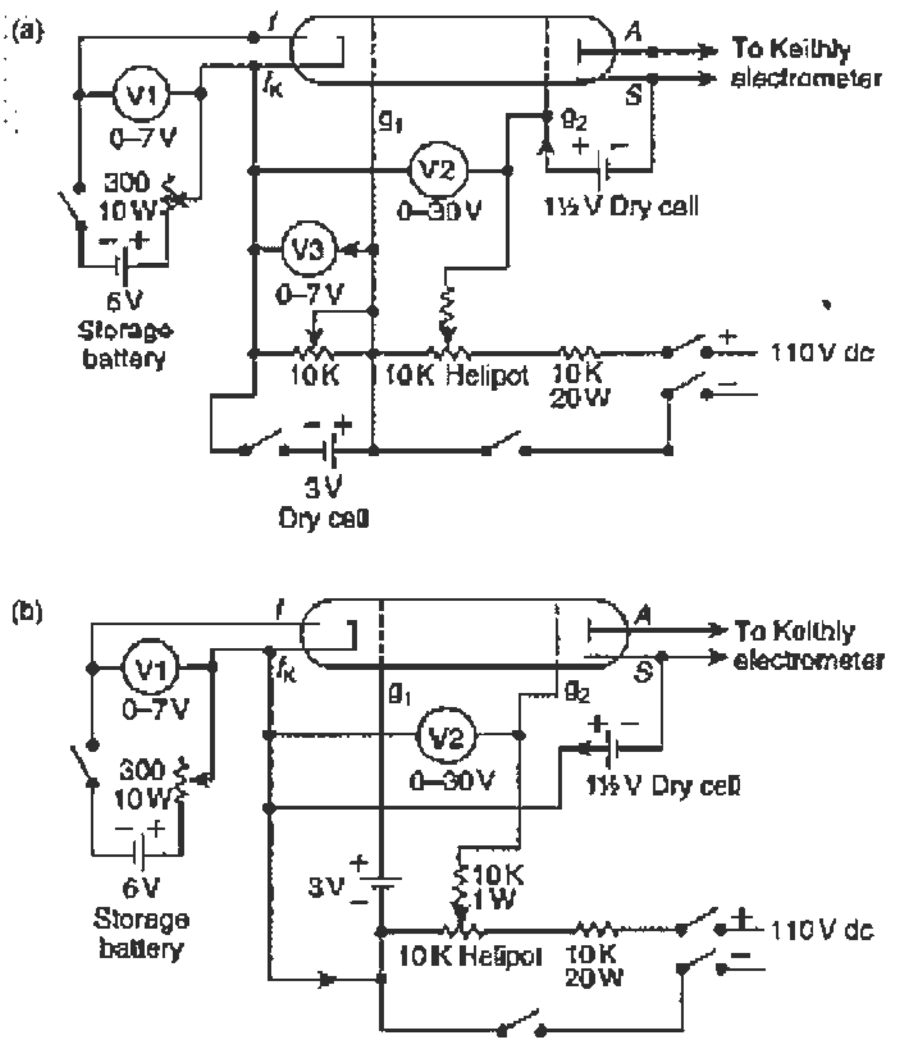


FIGURE 1.7 Wiring diagram for the Frank–Hertz experiment (a) for observation of excitation, and (b) for observation of ionization.

inserted through the small hole of the furnace. The junction should be positioned on the side of the tube near the electrodes. The other junction is immersed in a thermos of ice and water bath. The potential developed across the thermocouple is measured with a DMM; Fig. 1.8 gives a calibration curve for the copper-constantan thermocouple.

The resolution and definition of both the excitation and ionization curves is a function of atom density (temperature) and electron beam density (filament and grid 1 voltage) and the experimenter must find the optimum conditions. However, for large beam densities a discharge occurs, which, obviously, should be avoided.

A suggested adjustment procedure is to set grid 2 at 30 V and then advance grid 1 until the discharge sets in, as evidenced by the immediate

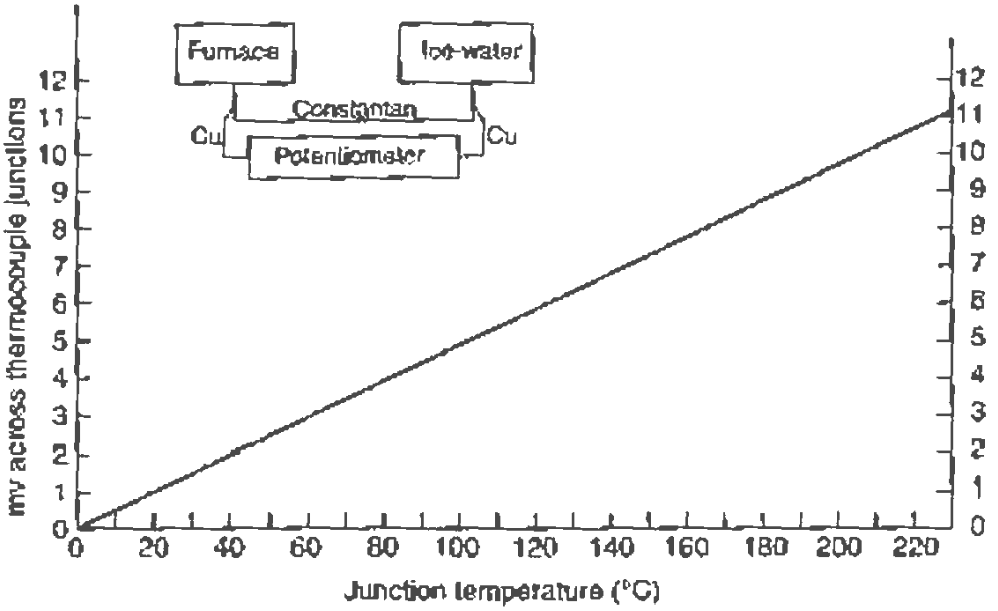


FIGURE 1.8 Calibration of copper-constantan thermocouple using ice standard.

build-up of the anode current. Grid 2 should then be quickly returned to 0 V and grid 1 set slightly below the discharge voltage; a reasonable filament voltage is between 4 and 6 V. To determine whether the tube is overheated it can be taken out of the oven for about 30 s; the collector current will then increase and maxima may appear if such is the case. If the tube is too cool, the emission current will be large, and the maxima, particularly those of higher order, will be washed out.

It is possible to use an oscilloscope for a simultaneous display of the electron or ion current against accelerating potential. The sweep generator (sawtooth) output is fed to the accelerating grid, while it synchronously drives the horizontal sweep; the output of the electrometer is fed to the vertical input. An excitation curve and an ionization curve obtained by a student in this fashion are shown in Fig. 1.9. Alternately a simple ramp circuit can be built to drive the accelerating grid and the digitized output of the electrometer read directly into a computer.

1.3.3. Analysis of the Data

Two sets of data obtained by a student for the excitation potential are shown in Fig. 1.10; both curves were obtained at a temperature of 195°C and with +1 V on grid 1. The filament voltage was 2.5 V for curve C and 1.85 V for

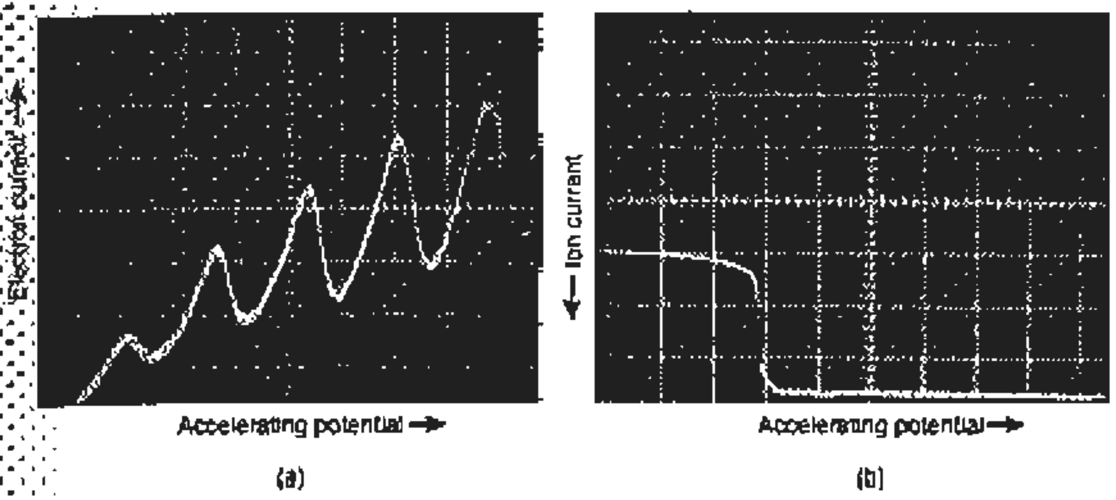


FIGURE 1.9 Oscilloscope display of a Frank–Hertz experiment: (a) Beam current vs accelerating potential. (b) Ion current vs accelerating potential.

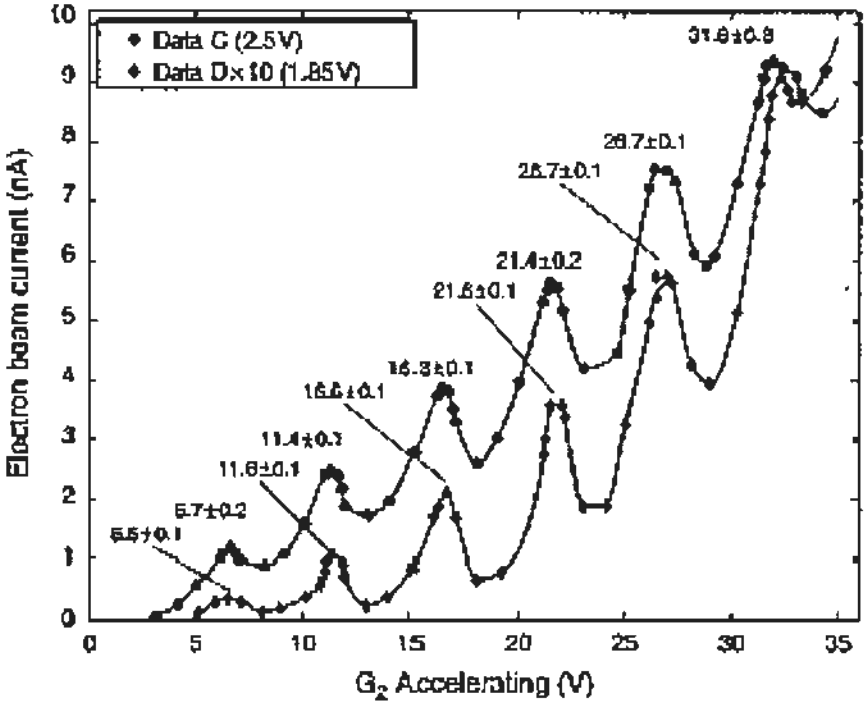


FIGURE 1.10 Plot of beam current versus accelerating voltage in Frank–Hertz experiment. Data for curve *C* (upper points) are obtained with the filament set at 2.5 V while data for curve *D* (lower points) are obtained with filament at 1.85 V.

curve *D* with the consequent decrease of the electron current by a *whole decade.*

Readings are taken for 1-V changes on grid 2 with smaller steps in the vicinity of the peak. A significant decrease in electron (collector) current is noticed every time the potential on grid 2 is increased by approximately 5 V, thereby indicating that energy is transferred from the beam in bundles ("quanta") of 5 eV only. Indeed, a prominent line in the spectrum of mercury exists at 253.7 nm, corresponding to $1237.8/253.7 = 4.86$ eV, arising from the transition of the $6s6p\,^3P_1$ excited state to the $6s6s\,^1S_0$ ground state. Our interpretation is that the electrons in the beam excite the mercury atom from the ground state to the $^3P_1$ state, thereby losing 4.86 eV in the process.

The location of the peaks is indicated in Fig. 1.10 and was measured in this case with a DMM. The average value obtained for the spacing between peaks is

$$5.02 \pm 0.1\ \text{V},$$

to be compared with the accepted spectroscopic value for the energy level difference (as already mentioned) of 4.86 eV.

Using the value found for the spacing between peaks and the location of the first peak, we obtain the contact potential

$$(6.65 \pm 0.15) - (5.02 \pm 0.1) = 1.63 \pm 0.18\ \text{V}.$$

As discussed in Section 1.3.1, with the configuration of potentials used (Fig. 1.5b) it is more probable that the same energy level will be excited twice rather than that several different levels will be excited; indeed, this is the way in which the data in Fig. 1.10 have been interpreted. This is not surprising if one considers the excitation probabilities for the energy levels lying closest to the ground state of mercury. It is possible, however, by using different grid and voltage configurations (for example, Fig. 1.5a) and improved resolution, to observe the excitations to other levels, namely, 6^3P_2, 6^3P_0, and 6^1P_1.

For the ionization potential, data obtained by a student are shown in Fig. 1.11. A word of caution is to be added to the interpretation of such ionization curves, which seem strongly dependent on filament voltage and vapor pressure; indeed, the very sharp increase observed in ion current is due to an avalanche (regenerative effect) of the ejected electrons ionizing more atoms, the thus-ejected electrons ionizing still more atoms, and so on. This avalanche does not necessarily occur as soon as the ionization threshold is crossed. If the vapor is too dense, the ions recombine before reaching the anode, thus masking the effect until complete breakdown sets in.

The curve shown was taken at a temperature of 155°C with a filament voltage of 2.6 V. If, then, the onset of ion current is taken to be at 11.4 ± 0.2 V,

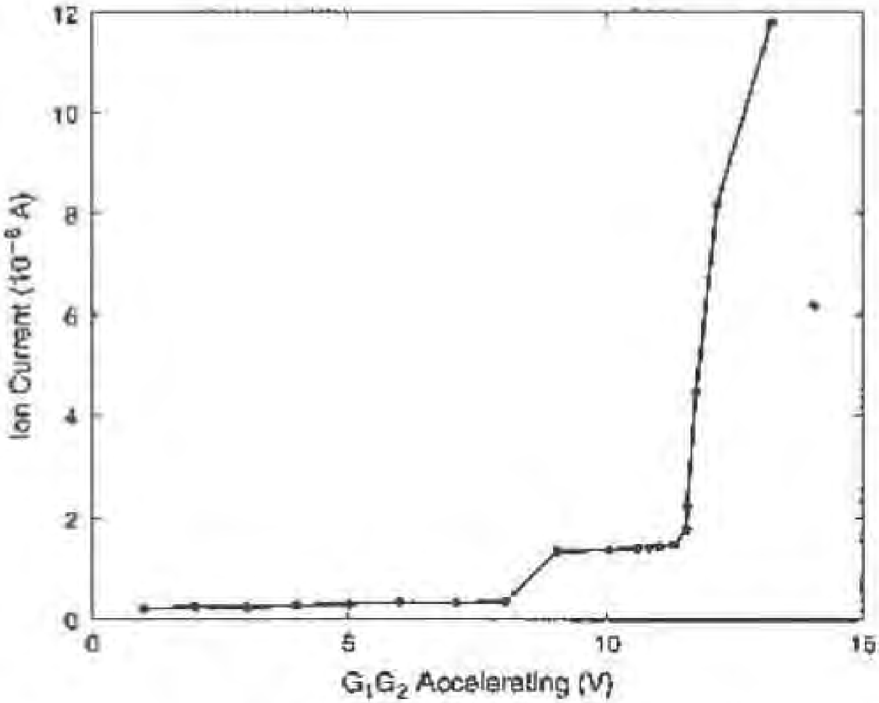


FIGURE 1.11 Ion current versus accelerating voltage in the Frank–Hertz experiment. The "knee" at 8 V is due to the photoelectric effect.

and using the value for the contract potential previously determined (from the excitation curve), 1.63 ± 0.18 V, the ionization potential is obtained as

$$(11.4 \pm 0.2) - (1.63 \pm 0.18) = 9.77 \pm 0.25 \text{ eV}$$

only in fair agreement with the accepted value of 10.39 eV.

An additional feature of the curve Fig. 1.11 is a "knee" in the ion current, setting in at approximately 8 V; the observation of this "knee" as well is strongly dependent on the temperature and current density, but can be consistently reproduced over a considerable range of these parameters. In order to understand this behavior we remember that the arrival of ions at the anode is equivalent to the departure of electrons; indeed, the observed behavior is due to a photoelectric effect produced at the anode, by short-wavelength light quanta (the electrons are further accelerated by grid 2). When the electron beam reaches 8 V, it can excite the 6^1P_1 level (lying at 6.7 eV above the ground state, plus 1.63 V for contact potential difference), so the mercury atoms radiate the ultraviolet line at 184.9 nm when returning to the ground state. These quanta are very efficient in ejecting photoelectrons from the anode, and the cylindrical geometry of the anode is most favorable for this process.

1.4. THE HYDROGEN SPECTRUM

The hydrogen atom is the simplest quantum-mechanical system. It consists of an electron bound, due to the Coulomb force, to a proton. It is characteristic of bound quantum-mechanical systems that their total energy cannot have any value, but that the system is found in one of a discrete set of energy levels, or states. Transitions of the system between these states may occur. Such transitions must satisfy the basic conservation laws of electric charge, energy, momentum, angular momentum, and the other relevant symmetries of nature.

Transition from a higher energy state to a state with less energy can occur for an isolated system, and the larger the probability for this transition, the shorter the "lifetime" of that excited state. During such spontaneous transitions of a quantum-mechanical system to a lower energy state, a quantum of radiation, or one or more particles, can be emitted, which will carry away the energy lost by the system (after recoil effects have been taken into account). In the presence of a radiation field the quantum-mechanical system can either gain energy from the field and change into a state with higher energy, or lose energy to the field and revert to a lower energy state. For all quantum-mechanical systems there exists a lowest energy state, called the *ground state*.

By observing the quanta of radiation, or the particles emitted during such transitions, we gain information on the energy levels involved. The typical example is optical spectroscopy, which consists of the accurate determination of the energy of the light quanta emitted by atoms. Infrared spectroscopy deals mainly with the quanta emitted by molecules, nuclear spectroscopy with the quanta emitted in nuclear transitions, and so on. In nuclei, however, the separation between energy levels is much larger, so that the emitted quanta of electromagnetic radiation lie in the gamma ray region; thus different techniques are employed for detection and measurement of their energy. It is also very common for nuclei to decay from one energy state to another by the emission of an electron and neutrino (beta decay) and for certain heavier nuclei by the emission of a helium nucleus (alpha particle). Similar processes take place in the interactions or decay of the elementary particles.

The idea of energy levels and their structure for the hydrogen atom was first introduced by Niels Bohr in 1913. However, a complete theoretical interpretation had to wait until the introduction of the Schrödinger equation in 1926. Even then, for theory to agree with observation it is necessary to include additional small effects such as the fine and hyperfine structure, relativistic motion, and other higher order corrections. These corrections are derived using the theory of quantum electrodynamics (QED) so that today we can theoretically calculate the energy levels of the hydrogen atom to the amazing accuracy of 1 part in 10^{11}.

We will use the Bohr theory to predict the hydrogen energy levels, because it is so simple, even though it assigns the incorrect angular momentum to the states. The postulates of the Bohr theory are (a) that the electron is bound in a circular orbit around the nucleus such that the angular momentum is quantized in integral units of Planck's constant (divided by 2π); namely, $pr = mvr = n(h/2\pi) = n\hbar$; and (b) that the electron in this orbit does not radiate energy, unless a transition to a different orbit occurs. We can then calculate the radii of these orbits and the total energy of the system, potential plus kinetic energy of the electron. The attractive force between the electron (charge $-e$) and the proton (charge $+e$) or a nucleus (of charge $+Ze$) is the Coulomb force, which is set equal to the centripetal force.

The total mechanical energy of the electron is

$$\begin{aligned} E &= T + V \\ &= \frac{1}{2}mv^2 - \frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r}. \end{aligned} \tag{1.7}$$

Here m, v, and $-e$ are the electron's mass, velocity, and electric charge, $+Ze$ is the charge on the nucleus, and r is the "orbital radius" of the electron.[10] The potential energy, of course, is just the attractive Coulomb potential between the electron and the nucleus. We can relate the velocity v to the other variables by using $F = ma$, where F is the Coulomb force and a is the centripetal acceleration. That is

$$\frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r^2} = m\frac{v^2}{r},$$

which implies that

$$v^2 = \frac{1}{m}\frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r}. \tag{1.8}$$

[10] We assume that the nucleus is infinitely heavy.

If we introduce this result into Eq. (1.7) we obtain

$$E = \frac{1}{2}\frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r} - \frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r} = -\frac{1}{2}\frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r} = -\frac{1}{2}|V|. \qquad (1.9)$$

At this point we can impose the Bohr quantization condition

$$r = n\frac{\hbar}{mv} \qquad (1.10)$$

to eliminate v in Eq. (1.8). Here n is the principal *quantum number*. We obtain

$$\frac{n^2\hbar^2}{m^2r^2} = \frac{1}{m}\frac{1}{4\pi\varepsilon_0}\frac{Ze^2}{r}$$

or

$$\frac{1}{r} = \frac{m}{n^2\hbar^2}\frac{1}{4\pi\varepsilon_0}Ze^2. \qquad (1.11)$$

Inserting this result in Eq. (1.9) we find for the total energy

$$E_n = -\left[\frac{mZ^2e^4}{2(4\pi\varepsilon_0)^2\hbar^2}\right]\frac{1}{n^2}. \qquad (1.12)$$

For the hydrogen atom where $Z = 1$, the expression in brackets in Eq. (1.12) equals 13.6 eV. This is the energy required to take an electron in the ground state ($n = 1$) and separate it from the nucleus completely ($E = 0$). We refer to it as the binding energy of the hydrogen atom. It is customary to introduce the Rydberg constant (wave number) through

$$E_n = -hcR_\infty\frac{1}{n^2}, \qquad (1.13)$$

where

$$R_\infty = 10973731.534\,\mathrm{m}^{-1}$$

and thus

$$E_1 = -13.6057\,\mathrm{eV}.$$

Furthermore, from Eq. (1.11) we can write for the radius of the orbits in hydrogen

$$r_n = n^2a_\infty$$

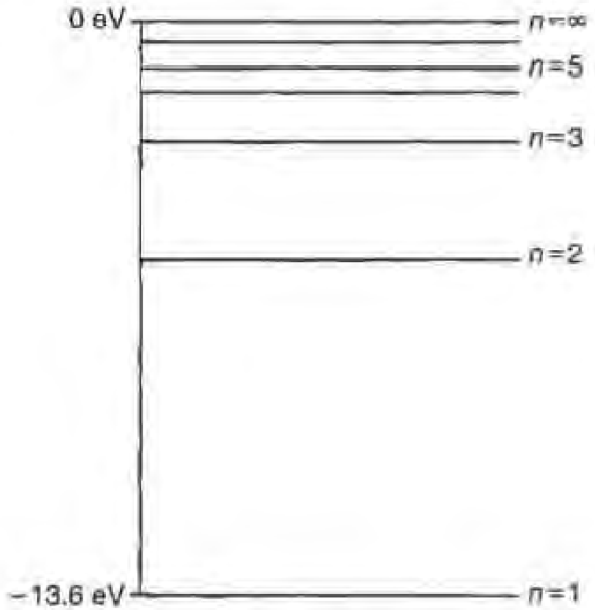


FIGURE 1.12 Energy-level diagram of the hydrogen atom according to the simple Bohr theory.

with

$$a_\infty = \frac{\hbar^2}{m}\frac{4\pi\epsilon_0}{e^2} = 0.5291772 \times 10^{-10}\ \mathrm{m},$$

called the Bohr radius.

The energy levels of the hydrogen atom that we derived can be represented by Fig. 1.12. However, the lines observed in the spectrum correspond to transitions between these levels; this is shown in Fig. 1.13, where arrows have been drawn for all possible transitions. The energy of a line is given by

$$\Delta E_{if} = hcR_\infty\left(\frac{1}{n_f^2}-\frac{1}{n_i^2}\right), \tag{1.14}$$

where the subscripts i and f stand for initial and final state, respectively.

Since the frequency of the radiation is connected to the energy of each quantum through

$$E = h\nu$$

one finds that

$$\frac{1}{\lambda} = \frac{\nu}{c} = \frac{E}{hc}$$

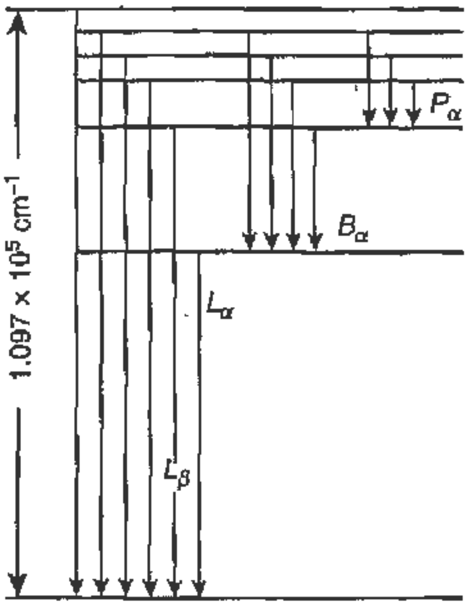


FIGURE 1.13 Transitions between the energy levels of a hydrogen atom. The lines L_α, L_β, etc., belong to the Lyman series, B_α, B_β, etc., to the Balmer series, and P_α, P_β, etc., to the Paschen series, and so forth.

and

$$\frac{1}{\lambda_{if}} = R_\infty \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right). \tag{1.15}$$

Indeed, the simple expression of Eq. (1.15) is verified by experiment to a high degree of accuracy.

From Eq. (1.14) (or from Fig. 1.13) we note that the spectral lines of hydrogen will form groups depending on the final state of the transition, and that within these groups many common regularities will exist; for example, in the notation of Fig. 1.13

$$\nu(L_\beta) - \nu(L_\alpha) = \nu(B_\alpha).$$

If $n_f = 1$, then

$$\lambda_{i1} = 91.1 \left(\frac{n_i^2}{n_i^2 - 1} \right) \text{nm} \qquad n_i \geq 2$$

and all lines fall in the far ultraviolet; they form the (so-called) Lyman series. Correspondingly if $n_f = 2$, then

$$\lambda_{i2} = 364.4 \left(\frac{n_i^2}{n_i^2 - 4} \right) \text{nm} \qquad n_i \geq 3$$

and all lines fall in the visible part of the spectrum, forming the Balmer series. For $n_f = 3$ the series is named after Paschen and falls in the infrared.

1.5. EXPERIMENT ON THE HYDROGEN SPECTRUM

1.5.1. General

To measure the frequency of the radiation emitted by atoms one can use either a grating or a prism to disperse the different wavelengths. When using a prism, one exploits the variation, with wavelength, of the refractive index of certain media. Prism spectrometers are limited to wavelength regions for which they are able to transmit the radiation; for example, in the infrared, special fluoride or sodium chloride prisms and lenses are used. In the ultraviolet, the optical elements are made of quartz. Also, the sensitivity of the detectors varies with wavelength, so that different types are used in each case (thermopile, photographic emulsion, phototube, etc.).

In this laboratory a small constant-deviation prism spectrograph and a 2-in. reflection grating spectrometer were used. We will consider in detail a measurement of the hydrogen spectrum with the grating, since an absolute value for the wavelengths can be obtained and visual detection is used. A brief discussion of prism spectrographs is given in Section 1.5.4.

From Fig. 1.14, it is evident that the path difference between rays 1 and 2 after reflection is

$$BD - AC = CB \sin\theta_r - CB \sin\theta_i,$$

where CB is the grating spacing d. The angles θ_i and θ_r are both taken as positive when they lie on opposite sides of the normal. Since for constructive interference the path difference must be a multiple of the wavelength, we obtain the condition

$$n\lambda = d\ (\sin\theta_r - \sin\theta_i). \tag{1.16}$$

It can be shown[11] that the resolution of the grating is given by

$$\frac{\lambda}{\Delta\lambda} = nN,$$

where n is the order of diffraction and N the total number of rulings. The same considerations apply to a transmission grating.

[11] See Chapter 5, Section 5.5.

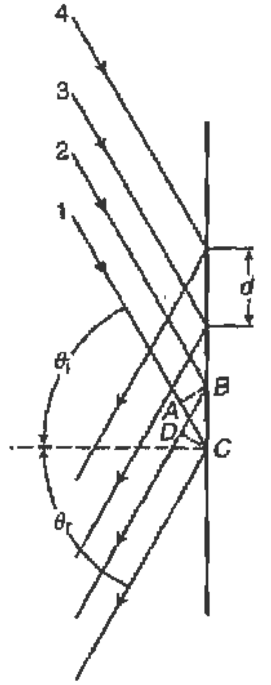


FIGURE 1.14 Schematic diagram of a reflection grating. A parallel beam of radiation is incident along the rays 1 through 4 at an angle θ_i, with respect to the normal; the reflected radiation is observed at an angle θ_r. The spacing between the grooves of the grating is d.

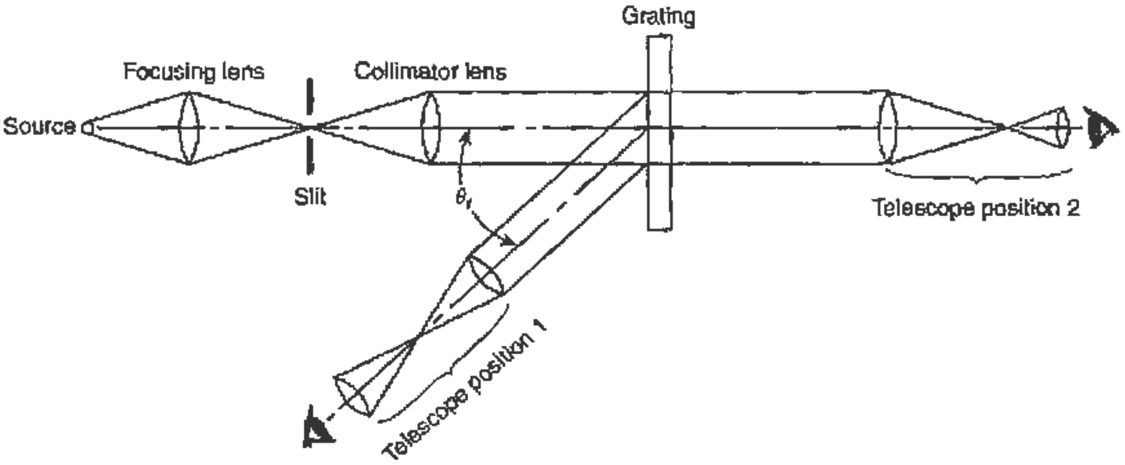


FIGURE 1.15 Diagrammatic arrangement of a grating spectrometer.

The grating is mounted on a goniometer table in the general arrangement shown in Fig. 1.15. A slit and collimating lens are used to form a beam of parallel light from the source, and a telescope mounted on a rotating arm is used for viewing the diffracted lines. It is obviously necessary to ensure parallelism of the incident and reflected beams, normality of the grating, and so on. A suggested alignment procedure is as follows:

(a) The viewing telescope is focused for parallel rays (on some distant object).

(b) Then with the grating removed, the slit is viewed with the telescope (in position 2) to ascertain that the slit is aligned and in focus; in this way the collimator lens is adjusted.

(c) The source and source lens are placed in position and the alignment and focusing are again checked. The cross hairs are aligned with the slit.

(d) This position of the telescope is carefully noted since it represents the 0° position. The readings on the scale should be made to one minute of a degree by using the vernier and a flashlight.

(e) From now on one may have to work in dark, or by draping the apparatus with a black cloth.

(f) The grating is placed in position and aligned for normal incidence ($\theta_i = 0$). This can be done by "autocollimation"; a strong light is focused onto the slit and a cardboard mask with a narrow slit is placed on the collimator lens. The grating is then adjusted until the reflected image of the cardboard slit coincides with the slit itself.

(g) Finally, the lines of the grating should be made parallel to the slit (hence the cross hairs); this can be done by viewing one edge of the grating with the telescope in position 1.

With any reasonable grating it is possible to observe the visible lines of the spectrum in several orders; thus we expect the measurements for λ/d to be self-consistent, since

$$\sin\theta_{m+1} - \sin\theta_m = (m+1)\,\frac{\lambda}{d} - m\,\frac{\lambda}{d} = \frac{\lambda}{d} \qquad (1.17)$$

independently of angle of incidence θ_i, or order.[12] The grating spacing d is usually stated by the manufacturer; for example, the grating in this laboratory had rulings on the order of 7000 to the inch ($d = 3.629 \times 10^{-6}$ m). However, d can be obtained by using one or more standard lines of known wavelength.

[12] Provided that both θ_m and θ_{m+1} are taken on the same side of the normal.

The following data were obtained by a student using the grating spectrometer. The source was a low-pressure hydrogen discharge tube (Cenco type 87210) operated at a few thousand volts; a 5-kV transformer and variac were used to provide the variable voltage. The useful life of these discharge tubes is limited because of the appearance of strong molecular bands after some hours of operation.

1.5.2. Determination of *d*

To obtain the grating spacing d, sodium (Na) was used as a standard, and measurement on three lines (for the shorter wavelength of the doublet) gave the results shown in Table 1.2. Since for all the above measurements θ_i is the same, it follows that

$$d^{-1}(n_k\lambda_k) + \sin\theta_\text{i} = \sin\theta_k$$

and a least-squares fit to the linear relation $\beta x + \alpha = y$ can be made; we have

$$\frac{1}{d} = \frac{N\sum(n_k\lambda_k \sin\theta_k) - \sum(\sin\theta_k)\sum(n_k\lambda_k)}{N\sum(n_k\lambda_k)^2 - \left[\sum(n_k\lambda_k)\right]^2}, \tag{1.18}$$

TABLE 1.2 Diffraction Angles from a Sodium Source

λ in nm	Order n	θ_n	$\theta_\text{i} = 19°12'$
615.43	1	29°42′	
	2	41°27′	
	3	55°58′	
589.00	1	29°14′	
	2	40°21′	
	3	53°49′	
	4	75°15′	
568.27	2	39°32′	
	3	52°12′	
	4	70°48′	

where the sums are over k, $k = 1, 2, \ldots, N$ and N is the total number of measurements. From the data of Table 1.2 we obtain[13]

$$\frac{1}{d} = 2.7085 \pm 0.009 \times 10^5 \text{ m}^{-1} \tag{1.19}$$

in good agreement with the manufacturer's specification.

Some care must be exercised when comparing wavelengths, since they do depend on the refractive index, n, of the medium in which they are measured,

$$c' = \frac{c(\text{vacuum})}{n},$$

hence

$$\lambda' = \frac{\lambda(\text{vacuum})}{n}.$$

The wavelengths listed in most tables are given for dry air at a pressure of 760 mm mercury. However, any theoretical calculation, such as in Eq. (1.15) predicts the vacuum wavelengths. The refractive index of air at stp is

$$n(\text{air}) = 1.00029. \tag{1.20}$$

1.5.3. The Balmer Series

Measurements on the first four members of the Balmer series, which lie in the visible region, can be made with the spectrometer described above. The data obtained by a student and their reduction are given in Table 1.3.

We observe that the obtained values for the wavelengths of the Balmer series are in agreement with the accepted values at the level of 1/1000. We can now test Eq. (1.15) and obtain the Rydberg wave number. We note that

$$\frac{1}{\lambda} = R_{\text{H}} \left[\frac{1}{4} - \frac{1}{n^2} \right].$$

So that from a least-squares fit

$$R_{\text{H}} = \frac{\sum \rho_i^2}{\sum \lambda_i \rho_i},$$

[13] In reaching this result we have constrained $\theta_i = 19°12'$.

TABLE 1.3 Data on the Balmer Series of Hydrogen as Obtained with a Grating Spectrometer

Color	θ_n	$\sin\theta_n - \sin\theta_i$	Order	Calculated λ	Accepted λ	Balmer series identification
Violet	33°12′	0.22199	2	410.75 ± 6	410.17 H_δ	$n_i = 6$
	41°15′	0.33378	3			
Blue	26°16′	0.11698	1			
	34°06′	0.23483	2	433.82 ± 8	434.05 H_γ	$n_i = 5$
	42°42′	0.35259	3			
Green	27°10′	0.13001	1			
	36°04′	0.26316	2	485.75 ± 10	486.13 H_β	$n_i = 4$
	46°09′	0.39559	3			
Red	30°11′	0.17720	1			
	42°57′	0.35579	2	657.94 ± 14	656.28 H_α	$n_i = 3$
	59°29′	0.53532	3			

Note. All wavelengths are in nm. These measurements used $d = 3692.1 \pm 30$ nm as determined by the previous measurements on the sodium standard lines, and $\sin\theta_i = 0.32557$.

where

$$\rho_i = \frac{4n_i^2}{n_i^2 - 4},$$

giving

$$R_H = (1.09601 \pm 0.003) \times 10^7 \text{ m}^{-1}$$

in good agreement with the accepted value[14]

$$R_H = \frac{M}{M+m} R_\infty = 1.096776 \times 10^7 \text{ m}^{-1}.$$

Here M is the mass of the proton and m the mass of the electron.

1.5.4. The Prism Spectrograph

Long before gratings became widely available, prisms were used as the dispersive element in spectrographs. Prism spectrographs are handy for viewing a large span of the spectrum and come in various ingenious optical

[14] The difference between R_H and R_∞ is due to the motion of the electron about the center of mass rather than about the proton.

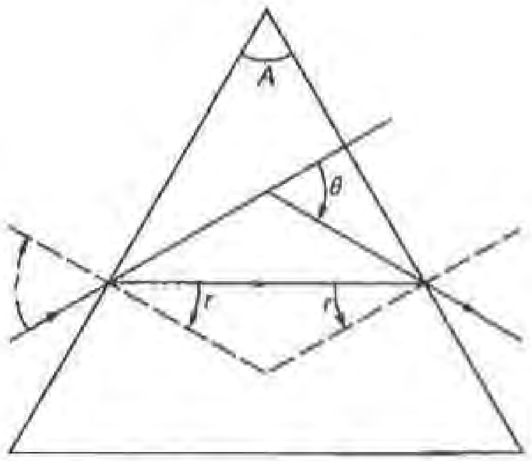


FIGURE 1.16 Diffraction of a ray at minimum deviation through a prism of apex angle A.

arrangements. The dispersion of a prism is a function of the refractive index; thus it cannot be used for absolute measurements without careful calibration.

In the case of a simple prism at minimum deviation (see Fig. 1.16) the diffraction angle θ is given by

$$\frac{\sin\theta_i}{\sin\theta_r} = n \qquad 2\theta_r = A \qquad \theta_i = \theta_r + \frac{1}{2}\,\theta;$$

thus

$$\sin\left(\frac{A+\theta}{2}\right) = n\sin\frac{A}{2}, \tag{1.21}$$

where θ_i and θ_r are the angles of incidence and refraction, respectively, and A is the apex of the prism. In Fig. 1.17 the refractive index of flint glass as a function of wavelength is given. We note that in the determination of wavelength from the diffraction angle the relation is by no means linear and is in general of serious complexity. Further, most modern prism spectrographs do not consist of a single dispersive element, but of some combination of prisms. The instrument used in this laboratory was of the "constant-deviation" type, and Fig. 1.18 gives the optical paths for an incident ray. It may be seen that the angle of incidence and the angle of exit can remain fixed for all wavelengths by an appropriate rotation of the prism; this has obvious advantages for positioning and alignment of source and detector.

The rotation of the prism is calibrated to give rough wavelength indications, but measurements are made on the exposed photographic plate

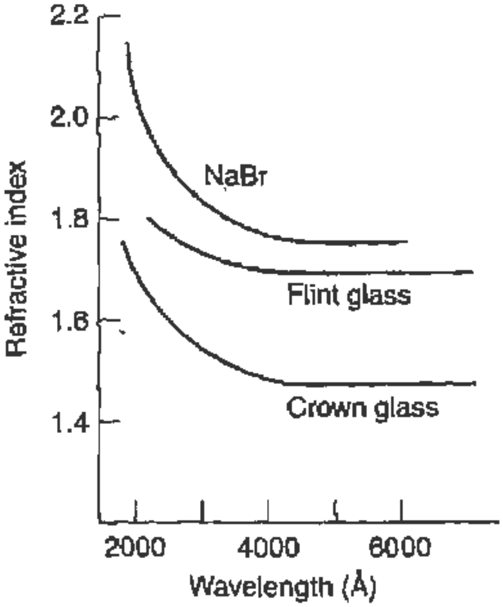


FIGURE 1.17 Refractive index of various materials as a function of wavelength.

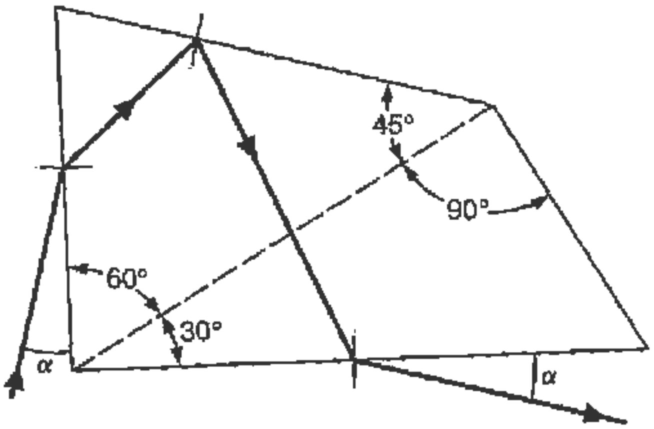


FIGURE 1.18 A constant-deviation prism and the diffraction of a ray passing through it.

or film. A known spectrum is superimposed on the spectrum that is to be investigated, and an interpolation between the known lines is used.

The general arrangement of the spectrograph is shown in Fig. 1.19. Source, lens, and slit should be aligned and the source focused on the slit. By viewing through the eyepiece and varying the prism position, one can get a feeling for the dispersion and the range of the instrument. To obtain photographs of a spectrum, the telescope is replaced by the camera assembly. Several exposures can be had on the same plate; to distinguish different spectra superimposed at the same location on the plate, the "fishtail," which controls the length of the slit, can be used.

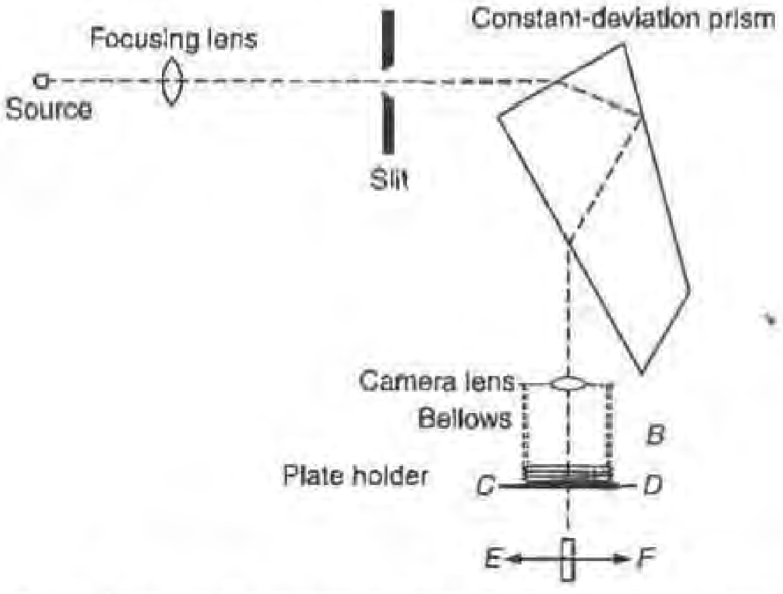


FIGURE 1.19 Schematic arrangement of the constant-deviation spectrograph.

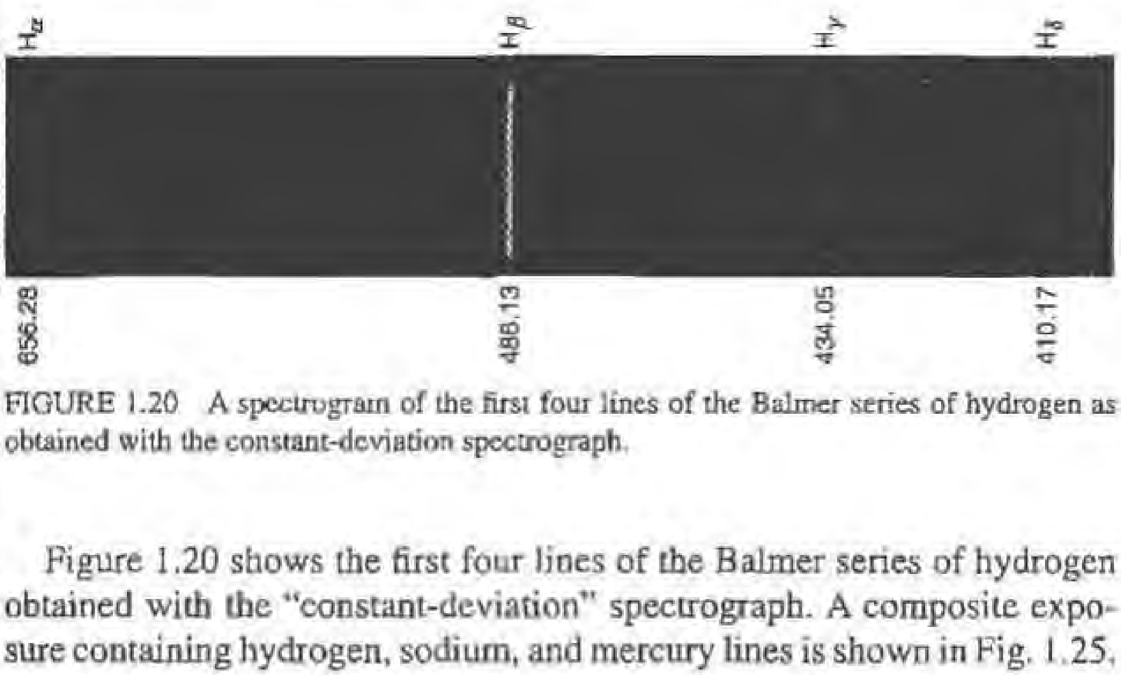


FIGURE 1.20 A spectrogram of the first four lines of the Balmer series of hydrogen as obtained with the constant-deviation spectrograph.

Figure 1.20 shows the first four lines of the Balmer series of hydrogen obtained with the "constant-deviation" spectrograph. A composite exposure containing hydrogen, sodium, and mercury lines is shown in Fig. 1.25.

1.6. THE SPECTRA OF SODIUM AND MERCURY

1.6.1. General

Mention has been made in the previous section of the spectrum of sodium (Na) and mercury (Hg); a brief analysis will be given here, since both elements have been investigated in detail and are representative of the one-electron spectrum (Na) and two-electron spectrum (Hg) correspondingly. Sodium has 11 electrons, so that the $n = 1$ and $n = 2$ shells are completely filled and one electron ($n = 3$) is found outside closed shells. In this respect the sodium spectrum should be equivalent to that of hydrogen except for the central charge that the free electron sees. Indeed, since the nucleus with $Z = 11$ is "screened" by 10 negative charges (the $n = 1$ and $n = 2$ electrons) the free electron sees a potential $-e/r$ when far from the nucleus and a potential $(-Ze)/r + C$ when close to it, where C is the potential generated at the nucleus by the other electrons. However, whereas in hydrogen only one energy level was found for each value of n, a more complex situation arises in sodium, with several levels corresponding to the same n. This splitting is to be attributed to the fact that the time-independent Schrödinger equation for the hydrogen-like atom,

$$\nabla^2\psi + \frac{2m}{\hbar^2}(E - V)\psi = 0,$$

admits solutions with a principal quantum number n, and angular momentum quantum number l, such that $n \geq l + 1$; when the potential that the electron sees is exactly of the Coulomb type as in the case of hydrogen, where $V = (-Ze^2)/r$ the energy eigenvalues

$$E_n = -\left[\frac{mZ^2e^4}{2(4\pi\varepsilon_0)^2\hbar^2}\right]\frac{1}{n^2} \tag{1.22}$$

are independent[15] of l, and agree with the Bohr theory. However, the screened potential that the free electron sees is no longer of the simple Coulomb type, and the energy of the level depends on both n and l. Orbits with smaller values of l are expected to come closer to the nucleus and thus be bound with greater strength; as a consequence their energy will be lower (more negative).

The energy level diagram of sodium is shown in Fig. 1.21, where the levels have been grouped according to their l value. The customary notation is used, namely, $l = 0 \rightarrow S$ state, $l = 1 \rightarrow P$ state, $l = 2 \rightarrow D$ state, $l = 3 \rightarrow F$ state, and so on, alphabetically. The last column in Fig. 1.21 gives the position of the levels of a hydrogen-like atom.

[15] This is the so-called Coulomb degeneracy: a peculiar coincidence for the Coulomb potential when used in the Schrödinger equation.

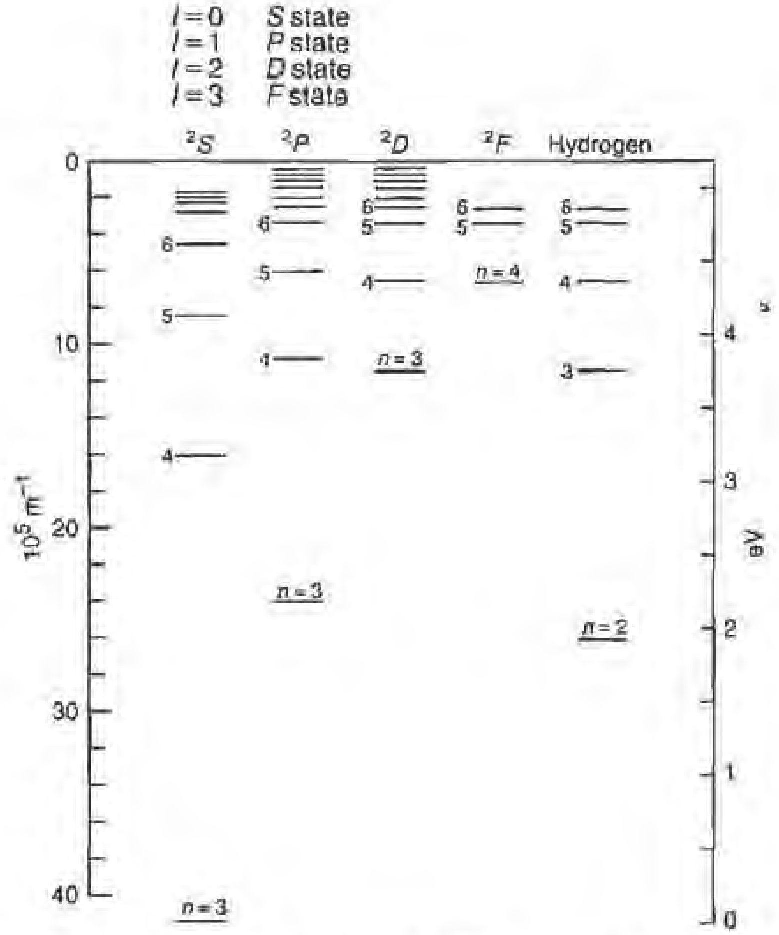


FIGURE 1.21 The energy-level diagram of sodium, grouped according to the orbital angular momentum. The last column gives the corresponding position of the levels of hydrogen. The left-hand scale is in 10^5 m^{-1}, referred to 0 for the singly ionized sodium atom; the right-hand scale is in electron volts referred to 0 at the ground state of the sodium atom.

We note that the higher the value of l, the smaller the departures from the hydrogen-like levels (as suggested qualitatively previously), and that for *given* l the energy levels for different n's follow the same ordering as the hydrogen-like atom, but with an effective charge Z^*, which for sodium is as follows: S states $Z^* \sim 11/9.6$; P states $Z^* \sim 11/10.1$; D states $Z^* \sim 1$; F states $Z^* \sim 1$.

1.6.2. Selection Rules

The spectral lines that we observe are due to transitions from one energy state to a lower one; however, in analyzing the spectrum of sodium, it becomes immediately evident that not *all possible* transitions occur. Thus certain "selection rules" for atomic transitions must be operative, and it is found that for all spectral lines[16]

$$\Delta l = \pm 1. \tag{1.23}$$

This selection rule is readily explained by the quantum-mechanical theory of radiation; it then means that only "electric dipole" transitions occur. Indeed, the transition probability for electric dipole is larger by a factor of $(c/v)^2$ (c, velocity of light) from the next order, while under no conditions do transitions occur in which the angular momentum does not change at all ($\Delta l = 0$). By applying the selection rule of Eq. (1.23) to the energy-level diagram of Fig. 1.21, we obtain Fig. 1.22, which gives the principal lines of the sodium spectrum; since l must change by one unit, transitions will always occur between adjacent columns and never within the same one.

Figure 1.23 is a reproduction of the visible part of the above spectrum obtained by a student with the constant-deviation spectrograph. Beginning from the top (long wavelengths) we recognize the following lines (where the wavelength is given in nanometers)

(a)	Red	615.43–616.07 nm	
(b)	Yellow	589.00–589.59	(famous Na D lines)
(c)	Green	568.27–568.82	
(d)		514.91–515.36	
(e)		497.86–498.29	
(f)	Blue	474.80–475.19	
(g)		466.49–466.86	
(h)	Blue–Violet	449.43–449.77	

1.6.3. Fine Structure

The data in Table 1.4 on the red, yellow, and green lines of sodium, viewed with the grating, were obtained by a student simultaneously with the data used for the determination of the grating spacing d of Eq. (1.19). In the above data two wavelengths were given for each sodium line. Indeed, by viewing through the constant deviation or the grating spectrometer it is easy to resolve into a doublet each of the lines that appear in Fig. 1.23; the spacing is on the order of several tenths of a nanometer.

[16] Exceptions (such as quadrupole transitions) are found in steller spectra.

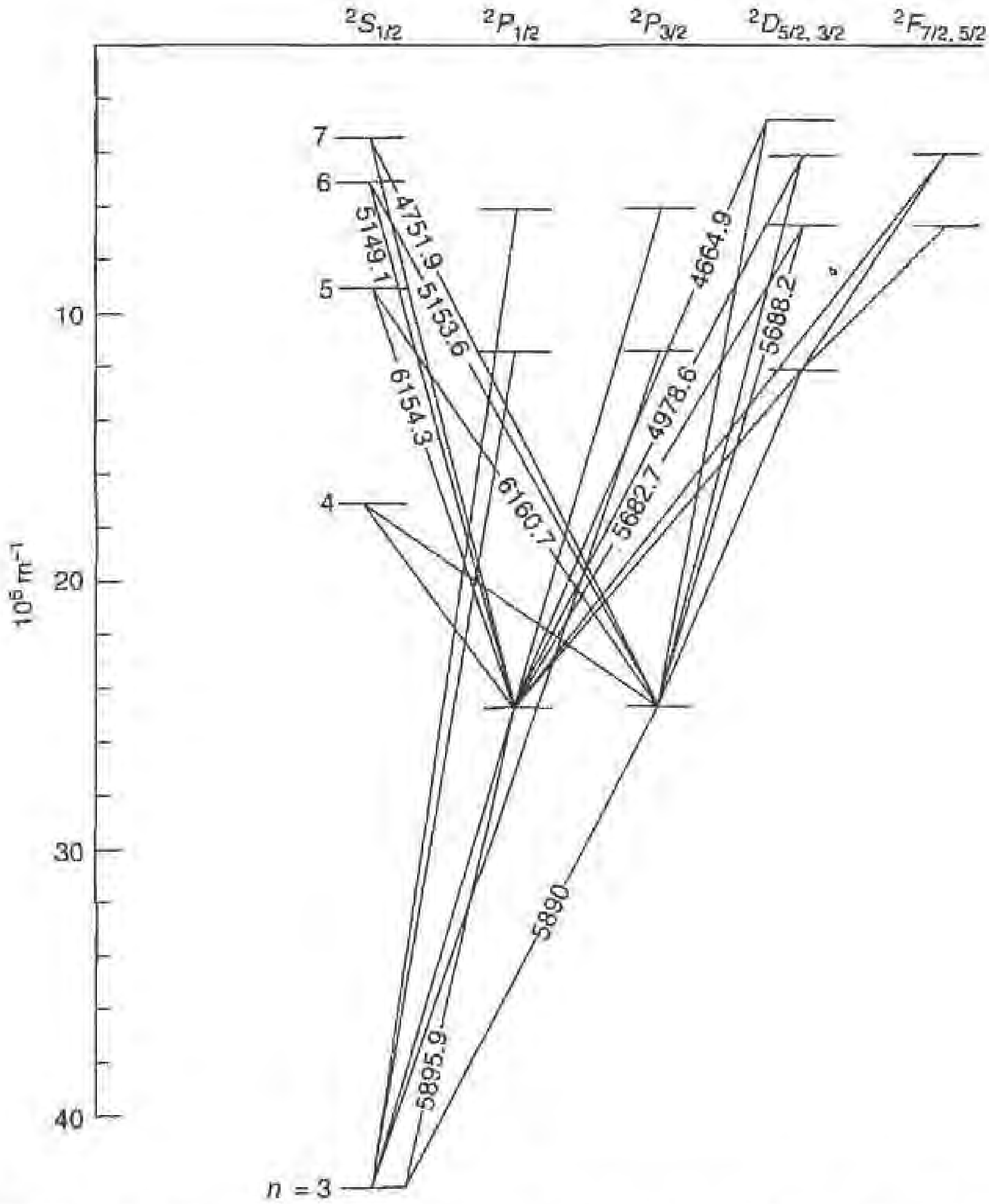


FIGURE 1.22 The "allowed" transitions between the energy levels of sodium. The wavelengths in angstroms (10 Å = 1 nm) of some of the principal lines are indicated. Note that the P states have now been shown in two columns, one referred to as $P_{1/2}$ the other as $P_{3/2}$; the small difference between their energy levels is the "fine structure."

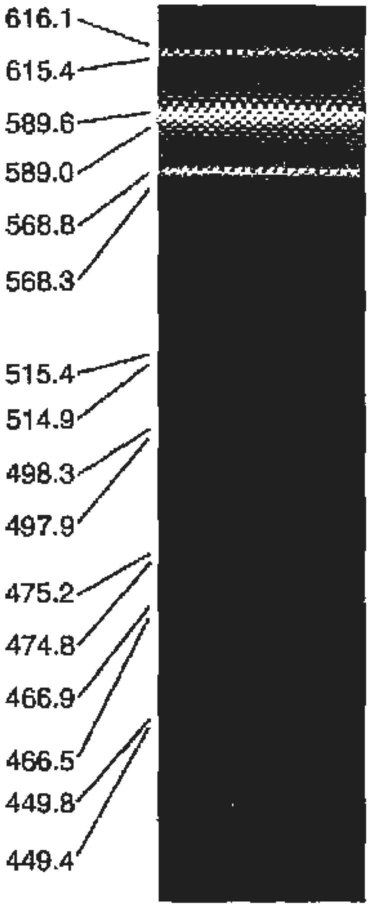


FIGURE 1.23 Photograph of the visible spectrum (in nm) of sodium as obtained with a constant-deviation spectrograph.

TABLE 1.4 Data on the Fine Structure of Sodium as Obtained with a Grating Spectrometer

Line	Order	θ_1	θ_2	$\Delta\theta$ (radians)
Red	2	41°27′	41°29′	5.8×10^{-4}
	3	55°58′	56°00′	5.8
Yellow	2	40°21′	40°23′	5.8
	3	53°49′	53°52′	8.7
	4	75°15′	75°23′	23.2
Green	2	39°32′	39°33′	2.9
	4	70°48′	70°56′	23.2

To reduce the data we note that

$$n_k\lambda = d(\sin\theta_k - \sin\theta_i),$$

where θ_i is the angle of incidence. Also

$$\theta_2 = \theta_1 + \Delta\theta$$

By letting $\sin\Delta\theta_k \approx \Delta\theta_k$, $\cos\Delta\theta_k \approx 1$,

$$n_k\Delta\lambda = d\cos\theta_k\Delta\theta_k. \tag{1.24}$$

Using $d = 3{,}692.1$ nm and averaging over orders within each line, namely writing

$$\Delta\lambda = d\frac{\sum\cos\theta_k\Delta\theta_k}{\sum n_k}, \tag{1.25}$$

we obtain for $\Delta\lambda$:

Line	$\Delta\lambda$ *(Experiment,* **nm)**	$\Delta\lambda$ *(Exact value,* **nm)**
Red	0.57	0.651
Yellow	0.63	0.597
Green	0.59	0.555

The experimental data are thus in ~10% agreement with the exact values.

This splitting of spectral lines was named "fine structure" and must reflect a splitting of the energy levels of sodium; if we express the wavelengths of the sodium lines in wave numbers ($\bar{\nu} = 1/\lambda = \nu/c$, i.e., in a scale proportional to energy since $\Delta E = hc\Delta\bar{\nu}$), it becomes evident that the spacing in all doublets is exactly the same and equal to $\Delta\bar{\nu} = 1.73 \times 10^3$ m^{-1}. Indeed, the doublet structure of all the above lines is due to the splitting of only the $3P$ ($n = 3$, $l = 1$) level as can be seen by referring back to Fig. 1.22. The splitting of the $3P$ level is due to the effect of the electron "spin" and its coupling to the orbital angular momentum (designated by l). According to the Dirac theory, the electron possesses an additional degree of freedom, called "spin," which has the properties of angular momentum of magnitude $s = \hbar/2$ (and therefore two possible orientations with respect to any axis, $m_s = +\frac{1}{2}$ or $m_s = -\frac{1}{2}$). The spin $\mathbf{s}$ can then be coupled to $\mathbf{l}$ according to the quantum-mechanical rules of addition for angular momenta; this will result in a total angular momentum of magnitude $j = l + \frac{1}{2}$ or $j = l - \frac{1}{2}$, and the energy of the state will depend on j. In the case of sodium, the $3P$ level splits into two levels, with $j = \frac{1}{2}$ and $j = \frac{3}{2}$ designated as $3P_{1/2}$ and $3P_{3/2}$ separated by $\Delta\bar{\nu} = 1.73\times 10^3$ m^{-1}.

1.6.4. Electron–Electron Coupling; the Mercury Spectrum

The mercury atom ($Z = 80$) has 80 electrons. These fill the shells $n = 1$, $n = 2$, $n = 3$, and $n = 4$ completely (60 electrons), and in addition, from the $n = 5$ shell, the $l = 0, 1, 2$ subshells account for another 18 electrons. The remaining two electrons instead of occupying the $l = 3$ and $l = 4$ subshells are in the $n = 6$ shell with $l = 0$, giving rise to a configuration equivalent to that of the helium atom.

We thus have an atom with two electrons outside closed shells in contrast to the one-electron systems of the hydrogen and sodium type. In the two-electron system, we can hardly speak of the n number of the atom, since each electron may be in a different shell; however we can still assign a total angular momentum $\mathbf{J}$ to the system, which will be the resultant of the values of each of the two electrons, and (as we saw in the previous section) of their additional degree of freedom, their spin. The addition of these four angular momenta, $\mathbf{l}_1$, $\mathbf{l}_2$, $\mathbf{s}_1$, $\mathbf{s}_2$, to obtain the resultant $\mathbf{J}$ can be done in several ways. For the helium or mercury atom, the Russell–Saunders coupling scheme holds, in which $\mathbf{l}_1$ and $\mathbf{l}_2$ are coupled into a resultant orbital angular momentum $\mathbf{L}$ and $\mathbf{s}_1$ and $\mathbf{s}_2$ into a resultant spin $\mathbf{S}$; finally $\mathbf{L}$ and $\mathbf{S}$ are coupled[17] to give the total angular momentum of the system $\mathbf{J}$. Since $\mathbf{s}_1$ and $\mathbf{s}_2$ have necessarily magnitude $\frac{1}{2}$, the resultant $\mathbf{S}$ has magnitude $S = 0$ or $S = 1$. It is customary to call the states with $S = 0$ singlets, those with $S = 1$ triplets, since when $S = 0$ for any value of L, only a single state can result, with $J = L + S = L$; when $S = 1$, however, three states can result with $J = L + S, L, L - S$, namely $J = L + 1, L, L - 1$ (provided $L \neq 0$). In systems where energy states have total angular momentum J, the selection rules for optical transitions are different, namely

$$\begin{aligned} &\Delta L = \pm 1 \\ &\Delta J = 0, \pm 1 \qquad \text{but not} \qquad J = 0 \to J = 0, \end{aligned} \tag{1.26}$$

and in principle no transitions between triplet and singlet states occur.

[17] In the ensuing discussion the quantum-mechanical rules of addition of angular momentum are used. Even if the reader is not familiar with them, he can infer them from following the development of the argument.

With these remarks in mind we consider the energy-level diagram of mercury. Since there are two electrons outside a closed shell, in the ground state they will both be in the $n = 6, l = 0$ orbit, and hence (due to the Pauli principle) must have opposite orientations of their spin, leading to $S = 0$; the spectroscopic notation is $^1S_0$. For the excited states one should expect both a family of singlet states and a family of triplet states; the singlets, $S = 0$, will be

$$\begin{array}{ll} ^1S_0 & \text{for } L = 0, \text{ and necessarily, } J = 0 \\ ^1P_1 & \text{for } L = 1, \text{ and necessarily, } J = 1 \\ ^1D_2 & \text{for } L = 2, \text{ and necessarily, } J = 2 \text{ etc.} \end{array}$$

Note the spectroscopic notation, where the upper left index is $2S + 1$, indicating the total spin of the state; the capital letter indicates the total L of the atom (according to the convention); and the lower right index stands for J. For the triplets, $S = 1$, and the states are

$$\begin{array}{ll} ^3S_0 & \text{for } L = 0,\ J = 1 \\ ^3P_{0,1,2} & \text{for } L = 1,\ J = 0, 1, 2 \\ ^3D_{1,2,3} & \text{for } L = 2,\ J = 1, 2, 3 \text{ etc.} \end{array}$$

The energy levels for mercury are shown in Fig. 1.24 with some of the strongest lines of the spectrum. It is seen that the selection rules on ΔL and ΔJ always hold, but that transitions with $\Delta S \neq 0$ do occur. It is also to be noted that the fine structure, that is, the splitting of the $6s6p$ ^3P level, is of considerable magnitude: $\Delta\bar{\nu}(^3P_0 - {}^3P_1) = 1.9 \times 10^4\ \text{m}^{-1}$; $\Delta\bar{\nu}(^3P_1 - {}^3P_2) = 4.6 \times 10^4\ \text{m}^{-1}$. Figure 1.25 is a reproduction of the superimposed spectra of hydrogen (longest lines), mercury (medium length), and sodium (shortest lines) obtained by a student with the prism spectrograph. Beginning with long wavelengths (from the left) one identifies the following lines of mercury:

(a) Red	690.75 nm
(b) Yellow doublet	578.97–576.96
(c) Green	546.07
(d) Blue triplet	435.84
(e) Violet	404.66.

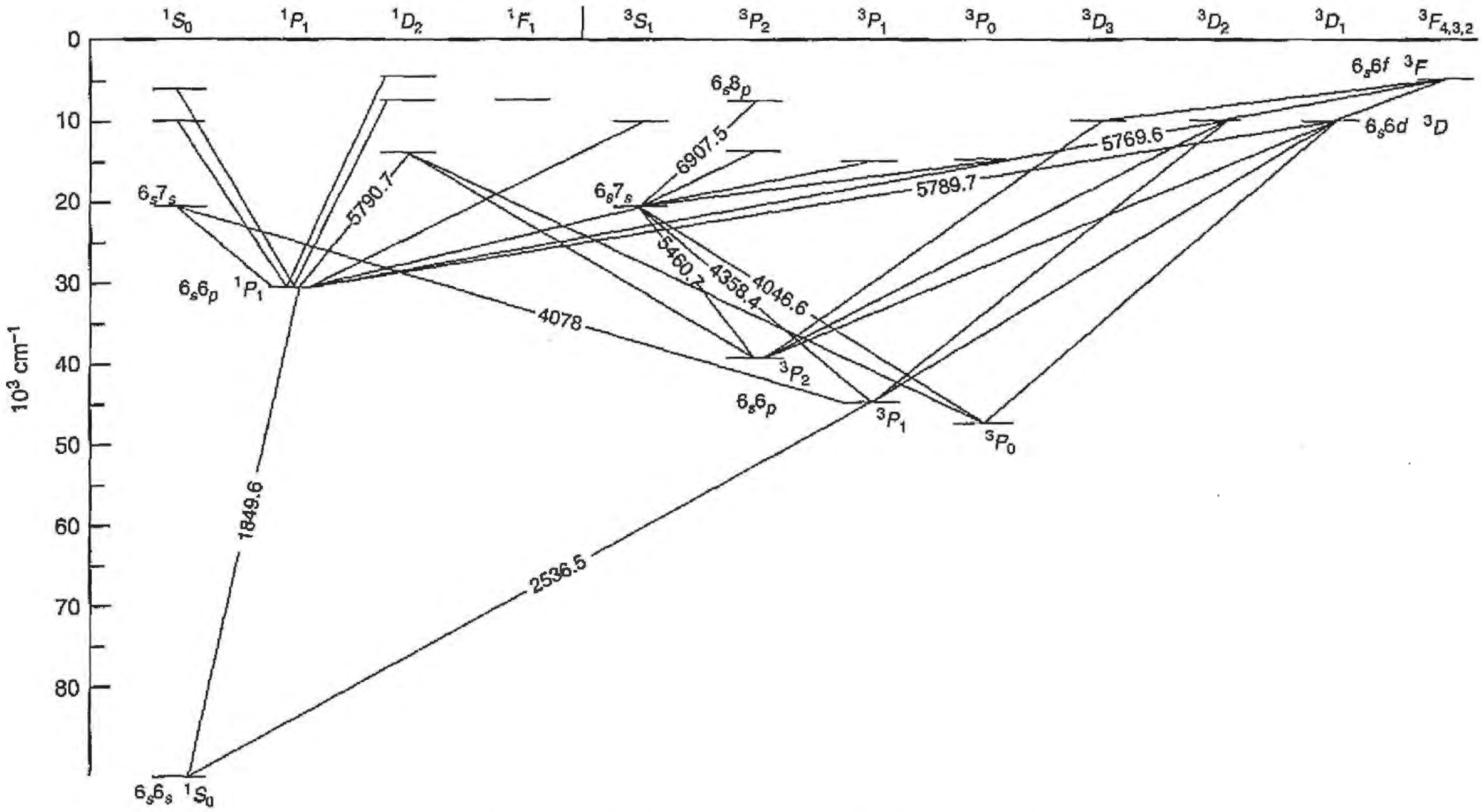


FIGURE 1.24 Energy-level diagram and the principal lines (in Å) in the spectrum of the mercury atom.

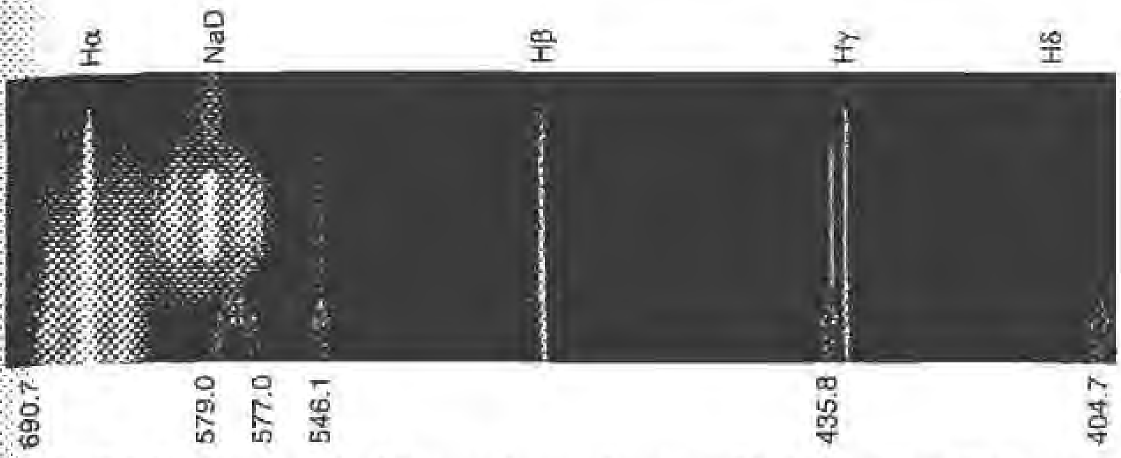


FIGURE 1.25 Photograph of the superimposed spectra of hydrogen (long slit), mercury (medium slit length), and sodium (short slit).

This concludes our discussion of the spectra and energy levels of the sodium and mercury atoms. The same treatment applies to all other one- or two-electron atoms, as well as to those with a one- or two-electron deficiency (hole) from a closed shell. Atoms with more electrons outside closed shells are treated on analogous lines, but the coupling schemes become more complicated, giving rise, as in the case of the rare earths, to extremely complex spectra.

CHAPTER 2

Electrons in Solids

2.1. SOLID MATERIALS AND BAND STRUCTURE

Most matter, as it can be perceived with our senses, consists of systems with very large numbers of interacting particles. In matter in the gaseous state, the distance between molecules is great, and therefore the forces are weak. In solids, however, the forces are much stronger. Understanding of the thermodynamic properties of "bulk" matter, based on the microscopic behavior of the constituent molecules or atoms, was first achieved through the statistical mechanics developed by Boltzmann. Because of the immense number of interacting bodies, the statistical approach is quite valid and has proved highly successful. Classical statistical mechanics, however, was unable to explain several phenomena until quantum-mechanical principles were incorporated. As we know, particles with half-integral spin—such as the electrons—obey "Fermi–Dirac" statistics, while particles with integral spin—such as photons and helium atoms—obey "Bose–Einstein" statistics. The fundamental distinction is that the former type of particles must have a completely antisymmetric wave function, whereas the latter ones must have a symmetric wave function. This leads to a different distribution function for the probability that a particle will occupy a certain cell in phase space.

The experiments in this chapter are primarily concerned with the electronic properties of solids. Since these properties are determined by the behavior of their electrons, it is Fermi statistics that are relevant. Most solid-state materials have a crystalline structure; that is, the atoms form a periodic lattice. Advantage can be taken of this periodicity so that the macroscopic behavior of the crystal is predicted from the general parameters of the lattice and the atoms that form it. It is found that the free electrons, instead of occupying distinct energy levels—as they do in atoms and molecules—are contained in certain energy bands. Knowledge of the "band structure" is necessary in most considerations of the solid state and specifically in the understanding of the behavior of semiconductors. The motion of the free electrons or holes (contained in the valence band) through the lattice can be studied in terms of a single-particle approach. Such phenomena as scattering and the absorption or emission of vibrational quanta (phonons) are invoked and are useful in explaining further details in the macroscopic behavior of the sample.

2.1.1. The Fermi–Dirac Distribution

Let us consider a large ensemble of free Fermi particles (such as electrons); the assumption is made that in phase space[1] there exist many states that these electrons can occupy. Each "cell" has a phase-space volume of h^3 (where h is again Planck's constant), so that the number of available cells for a differential volume of phase space is

$$dn = (h^3)^{-1} dp_x dp_y dp_z \, dx \, dy \, dz. \tag{2.1}$$

According to the exclusion principle, however, each cell can be occupied by two electrons (one with spin up and one with spin down), so that the number of available electron states is $2n$. If we integrate over the space

[1] Phase space is a space spanned by the momentum and position vectors of a particle. Thus, a particle moving in ordinary three-dimensional space will have six components in phase space.

coordinates and divide by the volume, we obtain the number of states n' per unit volume per differential element in momentum space:

$$n' = \frac{2}{V_0}\int_x\int_y\int_z dn = \left(\frac{2}{h^3}\right) dp_x dp_y dp_z.$$

Further, we can obtain the number of states per unit volume per unit energy interval dw_i

$$n_i = \frac{n'}{dw_i} = \frac{2}{h^3} 4\pi p_i^2 dp_i \frac{1}{dw_i}$$

and since for nonrelativistic velocities

$$w_i = \frac{p_i^2}{2m} \qquad dw_i = \frac{2p_i dp_i}{2m}$$

$$\frac{dN(w_i)}{dw_i} \equiv n_i = \frac{8\pi}{h^3}\sqrt{2m^3 w_i}. \tag{2.2}$$

Equation (2.2), which was obtained from very simple considerations, represents the number of states per unit volume per unit energy interval (at a given energy) and is called the "energy density of states." We note that for a simple ensemble of free Fermi particles (a) all energies are permissible (since $dN(w)/dw$ is a continuous and not singular function), namely, the energy is *not quantized*; and (b) the number of states increases with increasing energy.

Proceeding further to specify our system, we would like to know which of these infinitely many states are occupied, or in a statistical fashion, what is the probability that a state i of given energy w_i will be occupied. This is the Fermi–Dirac distribution and is given by

$$\frac{N_i}{2n} = \left[\exp\left(\frac{w_i - w_F}{kT}\right) + 1\right]^{-1}, \tag{2.3}$$

where k is the Boltzmann constant, T is the temperature of the system, and w_F is a characteristic energy, called the Fermi energy or Fermi-level energy.

It is interesting to note the properties of this function, graphed in Fig. 2.1:

(a) It is properly bounded, so that it can represent a probability

$$0 < N_i/2n < 1.$$

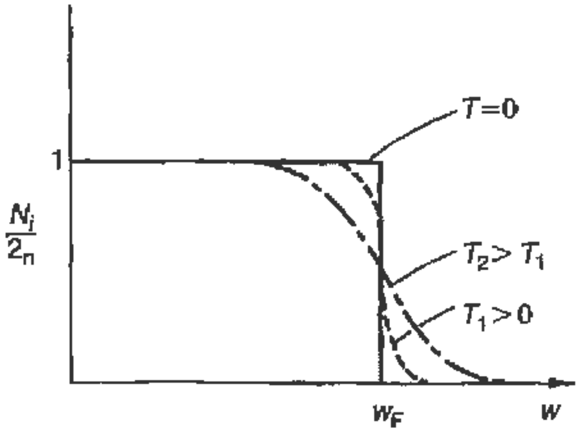


FIGURE 2.1 Probability of occupancy of a state of energy w_i as derived from Fermi–Dirac statistics.

(b) For large values of w_i it assumes the form of the Boltzmann distribution

$$\text{Const} \times \exp(-w_i/kT).$$

(c) For $T = 0$ it is a step function, with

$$N_i/2n = 1 \qquad w_i < w_F$$
$$N_i/2n = 0 \qquad w_i > w_F.$$

(d) For $T \neq 0$, w_F has the property that $N(w_F) = \frac{1}{2}$, and as many states above w_F are occupied, that many states below w_F are empty.

(e) In solids and for average $T \neq 0$, the distribution function is only slightly modified from its shape at $T = 0$ (for solids w_F is on the order of a few electron volts, while $1/kT = 40 \text{ eV}^{-1}$ at $T = 300$ K).

Combining the Fermi–Dirac distribution (Eq. (2.3)) with the energy density of states (Eq. (2.2)) it is possible to obtain any desired distribution. For example, the number of electrons per unit volume (density) at an energy w in the interval dw is given by

$$N(w)\,dw = \frac{8\pi}{h^3}\sqrt{2m^3 w}\left\{\exp\left(\frac{w - w_F}{kT}\right) + 1\right\}^{-1} dw. \qquad (2.4)$$

If we express Eq. (2.4) in terms of the Cartesian coordinates of the velocity, v_x, v_y, and v_z, and integrate over v_x and v_y, we obtain the number of electrons per unit volume with a given velocity in the z direction, v_z (in the

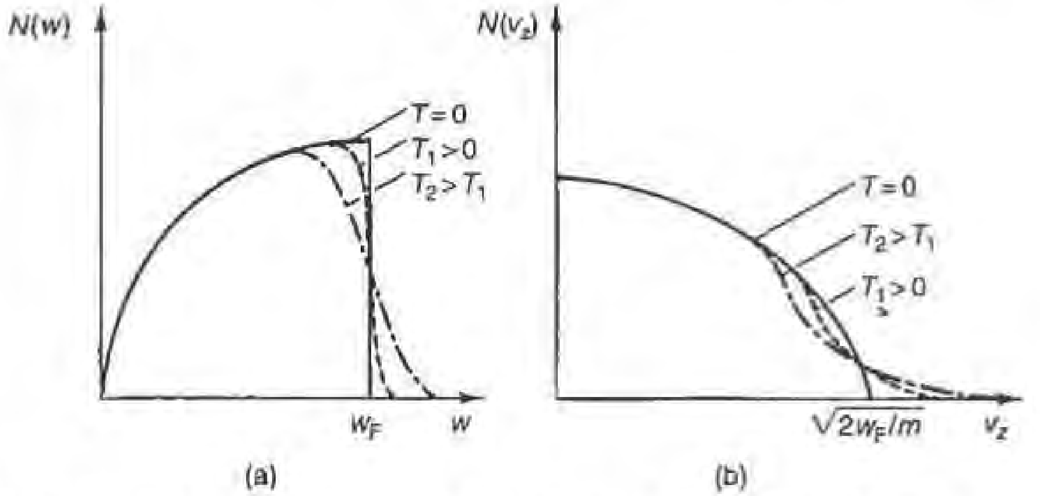


FIGURE 2.2 (a) Number of electrons with an energy w in the interval dw. (b) Number of electrons with z component of velocity v_z in the interval dv_z.

interval dv_z). The result of this integration is[2]

$$N(v_z)\,dv_z = \frac{8\pi}{h^3}\frac{m^2kT}{2}\ln\left\{1+\exp\left(\frac{w_F - mv_z^2/2}{kT}\right)\right\}dv_z. \qquad (2.5)$$

The two distributions given by Eqs. (2.4) and (2.5) are shown in Fig. 2.2.

Even though the majority of the electrons in a solid are not free (as we originally assumed), Fermi–Dirac statistics are applicable, especially to metals. In metals at least one electron per atom has several states available (is in the conduction band), so that it can be considered free; since there will be 6×10^{23} free electrons per gram mole, statistical methods are well justified.

2.1.2. Elements from the Band Theory of Solids

Up to now, no account has been taken of the interatomic or intramolecular forces that might act on the free electrons. Indeed, we expect (from previous experience) that the consideration of some potential in the region where the electrons move will result in the appearance of energy levels; however, because of the periodic structure of this potential, instead of energy levels, *energy bands* appear, and only the states contained in these bands can be

[2] A. Sommerfeld, *Thermodynamics and Statistical Mechanics*, p. 285, Academic Press, New York, 1956.

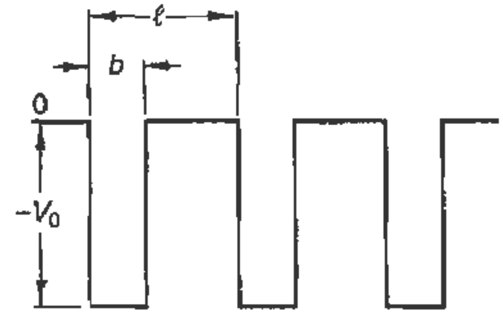


FIGURE 2.3 A periodic potential that may be considered as an idealization to the actual potential of a crystal lattice.

occupied (with any significant probability). In the following paragraphs we will sketch two approaches toward the understanding of the physical origin of the energy bands.

Consider first the one-dimensional problem[3] of an electron moving in a potential consisting of an infinite sequence of "square" wells of depth V_0 and width b and spaced at a distance l from one another (Fig. 2.3). The solution of the Schrödinger equation for such a potential gives for the electron wave function

$$\Psi_k = u_k(x)e^{ikx} \tag{2.6}$$

with $k = 2\pi/\lambda = p/\hbar$ the wave vector of the electron. This wave function consists of the plane wave part e^{ikx}, and $u_k(x)$, which must have the periodicity of the lattice, namely, $u_k(x \pm l) = u_k(x)$. If there are N lattice sites, the length of the crystal is Nl and we impose the periodic boundary condition $\Psi_k(x + Nl) = \Psi_k(x)$. This leads to $e^{ikNl} = 1$, or

$$\begin{aligned} kNl &= n2\pi \\ k &= n2\pi/Nl \qquad n = 0, \pm 1, \pm 2, \ldots. \end{aligned} \tag{2.7}$$

Equation (2.7) determines the allowed values of k, which form almost a continuum because of the very large integer value of N. Note that for $N = 1$ one obtains the familiar "particle in a box" energy levels, with

$$E = \frac{p^2}{2m} = \frac{k^2\hbar^2}{2m} = \frac{n^2\hbar^2}{2ml^2}.$$

[3] E. Merzbacher, *Quantum Mechanics*, third ed., Wiley, New York, 1998.

Having determined the wave function, it is possible to solve the Schrödinger equation for the energy eigenvalues

$$H|\Psi_k\rangle = E(k)|\Psi_k\rangle \qquad \text{or} \qquad \langle\Psi_k^*|H|\Psi_k\rangle = E(k), \tag{2.8}$$

where H is the one-dimensional Hamiltonian operator

$$H = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V(x)$$

and $V(x)$ is now the potential of Fig. 2.3.

The solution of Eq. (2.8) is given in graphical form in Fig. 2.4. We note the following:

(a) Even though all values of k are allowed, discontinuities arise at $k = n\pi/l$ (note that for this particular electron wavelength, Bragg reflection from the lattice will occur with a half-angle $\theta = 90°$; $n\lambda = 2l\sin\theta$, hence $\lambda = 2l/n$, and since $\lambda = 2\pi/k$, it follows that $k = n\pi/l$).

(b) Not all values of the energy are allowed, but only certain "bands"; other bands of energy are forbidden.

(c) The relation between E and p (or k) is no longer the familiar parabolic

$$E = \frac{p^2}{2m} = \frac{k^2\hbar^2}{2m}. \tag{2.9}$$

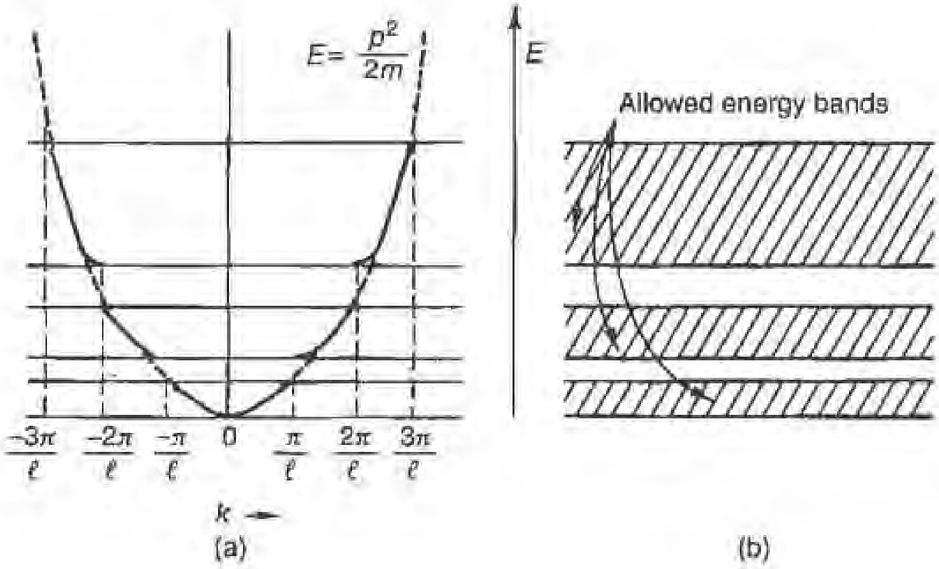


FIGURE 2.4 Results of the solution of the simplified one-dimensional lattice problem. (a) Plot of energy E versus wave number $k = p/\hbar$ for an electron in a crystal lattice. (b) The allowed and forbidden energy bands.

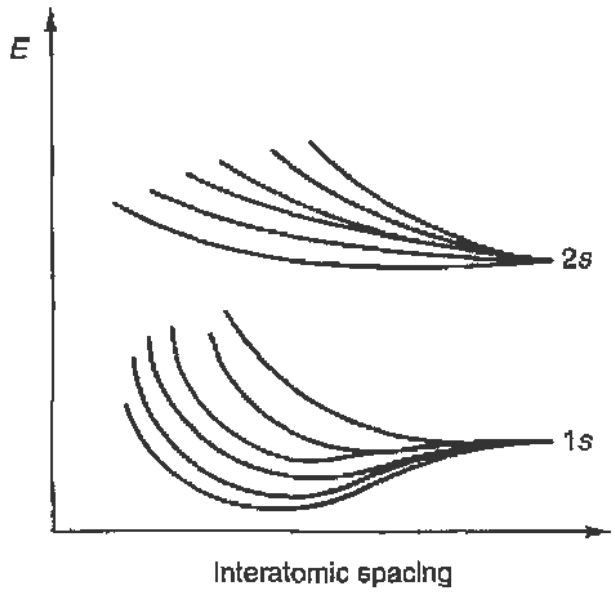


FIGURE 2.5 Energy levels of a system of six similar atoms placed in a linear array.

We can, however, retain this relation if the mass m is assumed variable and a function of k, namely,

$$m^*(k) = \frac{\hbar^2}{(d^2E/dk^2)}. \tag{2.10}$$

The same formalism is carried over into three dimensions, but now the bands are replaced by allowed (Brillouin) surfaces and the axes of symmetry of the crystal must be taken into account.

A different approach is to start with a molecular wave function and study its behavior as the number of identical atoms is increased. In Fig. 2.5 are plotted the energy levels against interatomic distance for the 1s and 2s states of a linear array of six atoms (after Shockley). If, then, in the limit the (almost infinite) array of the crystal is considered, the energy levels coalesce into bands. This is shown in the left-hand side of Figs. 2.6 and 2.7, where the energy bands plotted against interatomic spacing are given for diamond which is an insulator (after Kimball), and for sodium (after Slater), which is a conductor. If the lattice spacing for the particular crystal is known (from experiment), it is possible to read off from the graphs the limits of the energy bands. This is done diagrammatically on the right-hand side of Figs. 2.6 and 2.7; also indicated is the position (in electron volts) of the Fermi level (as it can be calculated, for example, from Eq. (2.4) and the electron density within each band).

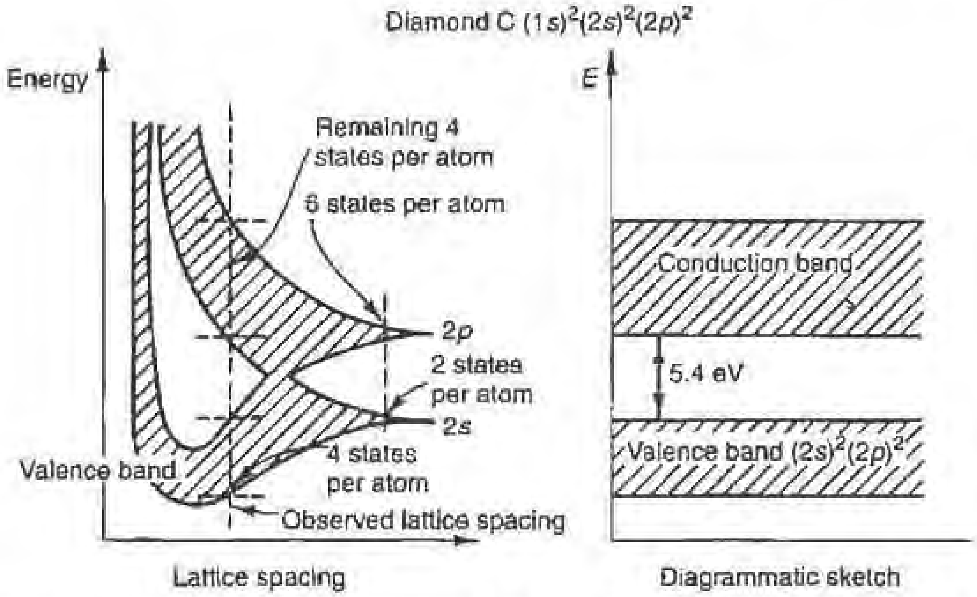


FIGURE 2.6 The energy band structure of diamond (insulator) as a function of lattice spacing. The observed lattice spacing is also indicated.

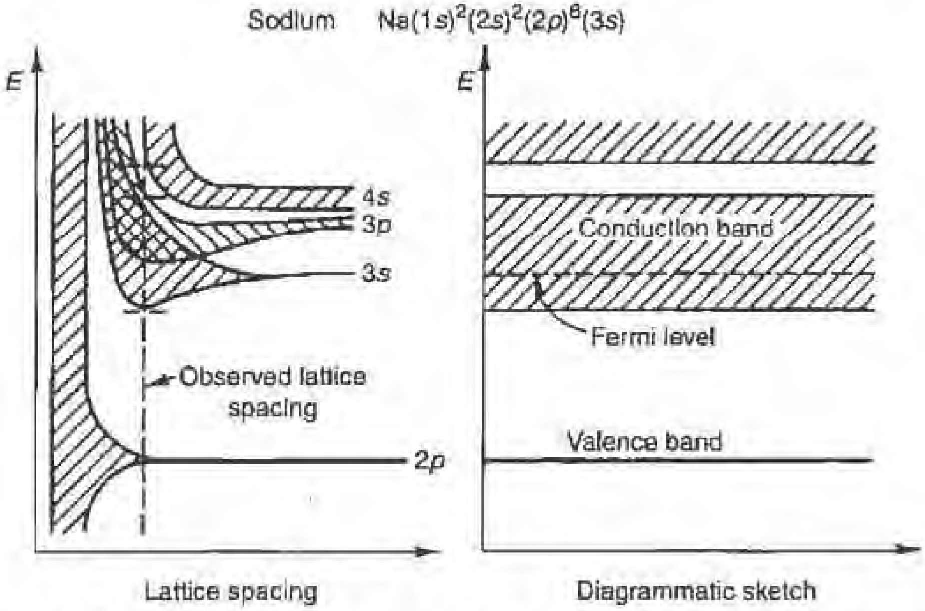


FIGURE 2.7 The energy band structure of sodium (conductor) as a function of lattice spacing. The observed lattice spacing and position of the Fermi level are also indicated.

From these considerations it is possible to understand the difference between conductors, insulators, and semiconductors. For diamond, for example, the valence band is completely filled (this fact follows also from the atomic structure of carbon and the deformation of the energy levels).

The next available states are approximately 5.4 eV higher and hence cannot be reached by the electrons, with a consequent inhibition of their mobility; diamond therefore behaves as an insulator. For sodium, in contrast, the Fermi level lies in the middle of an energy band, so that many states are available for the ($3s$) electron, which can move in the crystal freely; sodium behaves as a conductor. Pure semiconductors, such as germanium, have a configuration such that the valence band is completely filled, but the conduction band lies fairly closely to it (0.80 eV). At high enough temperatures (that is, on the order of a few thousands of degrees), the electrons in the valence band acquire enough energy to cross the gap and occupy a state in the conduction band; when this happens the material that was previously an insulator becomes intrinsically conducting.

Both the electric and thermal conductivity of a solid depend on the density and mobility of the free electrons. Completely analogous to the motion of electrons is the motion of "holes"; holes can be thought of either as "vacancies" in an almost-filled band, or as electrons with negative effective mass.[4] Due to their thermal energy, the carriers have a random motion characterized by $(3/2)kT = E = m^*v^2/2$. When an electric field is applied, a drift velocity is superimposed on the random motion of the carriers, resulting in a steady-state current flow.

2.2. EXPERIMENT ON THE RESISTIVITY OF METALS

In this experiment we will explore the physics behind electrical resistance in metals. What's more, we will do it with a novel technique that measures the *resistivity* of the metal, a property only of the type of material and independent of the size or shape of the conductor. This technique, in fact, can make measurements of the sample without actually touching it, and has found a lot of use in modern applications. It is based on the paper C. P. Bean, R. W. DeBlois, and L. B. Nesbitt, Eddy current method for measuring the resistivity of metals, *J. Appl. Phys.* **30**, 1976 (1959).

First, we make the connection between resistance and resistivity. We assume that Ohm's law is valid, that is, $V = IR$, where R is independent

[4] This can be seen from Eq. (2.10) and the negative curvature of some parts of the $E(k)$ curve of Fig. 2.4a.

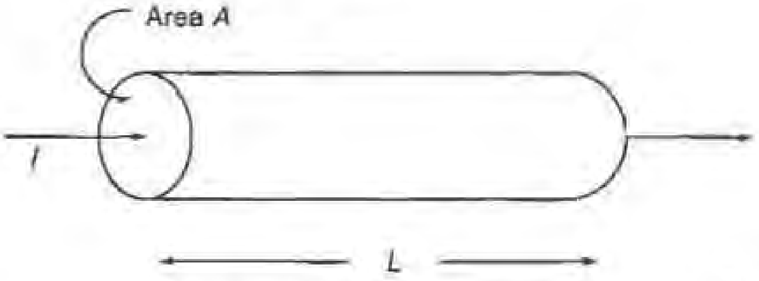


FIGURE 2.8 An idealized resistor.

of voltage or current. Consider the idealized resistor pictured in Fig. 2.8. The resistor has a length L and a cross-sectional area A. A voltage V is applied across the ends of the resistor. A current I of electrons flows from one end to the other, against a resistance R, which is due to the electrons interacting somehow with the atoms of the material.

Consider Ohm's law on a microscopic level. The magnitude of the electric field setup across the ends of the resistor is just $E = V/L$. The electrons that carry the current will be spread out over the area A, so at any point within the resistor the *current density* is (magnitude) $j = I/A$. Therefore Ohm's law becomes

$$E = j\rho, \tag{2.11}$$

where

$$R = \rho\frac{L}{A}$$

and ρ is the "resistivity," a property of the material that is independent of the dimensions of the resistor. Equation (2.11) can be derived from the theory of electrons in metals. The resistivity arises from collisions between the electrons and the atoms of the material. In a metal, the electrons are essentially free, so without any collisions they would continually accelerate under the applied field with an acceleration $a = eE/m$, where e and m are the electron charge and mass. However, the collisions cause the electrons to stop and then start up again, until the next collision. If the time between collisions is called τ, then the "drift" velocity v_{d} is just

$$v_{\mathrm{d}} = a\tau = \frac{eE\tau}{m}. \tag{2.12}$$

Now if there are n electrons per unit volume in the resistor, then a total charge $q = (nAL)e$ passes through the resistor in a time $t = L/v_{\mathrm{d}}$.

TABLE 2.1 Electrical and Thermal Properties of Metals

Name	Z	A	Electrical resistivity ($\mu\Omega \cdot \text{cm}$)	Temperature coefficient (10^{-3}/K)	Thermal conductivity $\left(\frac{\text{cal}}{\text{cm} \cdot \text{K} \cdot \text{s}}\right)$	Θ_D (K)
Al	13	26.98	2.65	4.29	0.53	395
Fe	26	55.85	9.71	6.51	0.18	420
Cu	29	63.55	1.67	6.80	0.94	333
Zn	30	65.38	5.92	4.19	0.27	300
Sn	50	118.69	11.50	4.70	0.16	260
Pb	82	207.19	20.65	3.36	0.083	86
Bi	83	208.98	106.80	—	0.020	118

Therefore, the current density is

$$j = \frac{I}{A} = \frac{1}{A}\frac{q}{t} = \frac{1}{A}\frac{nALe}{L/v_\text{d}} = nev_\text{d}, \tag{2.13}$$

and therefore,

$$\rho = \frac{m}{ne^2}\frac{1}{\tau}. \tag{2.14}$$

Often the "conductivity" $\sigma \equiv 1/\rho$ is used instead of the resistivity.

Electrical resistivities are listed[5] for various metals at room temperature in Table 2.1. Also included are some thermal properties, which are closely related to the resistivity through the underlying physics.[6] One of these is the temperature coefficient of resistivity, defined as $(1/\rho)d\rho/dT$. This quantity is in fact temperature dependent as we shall see, and the quoted numbers should be valid near room temperature.

Clearly, the fundamental physics of resistivity lies in the values for the collision time τ. The interaction of the quantum-mechanical electron waves and the quantized lattice of the metal crystal accounts for the collision time

[5] Values for Z, A, resistivity, and thermal conductivity are taken from L. Montanet *et al.* Review of particle properties, *Phys. Rev. D* **50**, 1241–1242 (1994). The temperature coefficient of resistivity, and all data for Zn and Bi, is from D. R. Lide, *CRC Handbook of Chemistry and Physics*, 56th ed., p. F-166, CRC Press, Boca Raton, FL, 1975. The Debye temperature is from E. U. Condon and H. Odishaw (Eds.), *Handbook of Physics*, 2nd ed., Part 4, Tables 6.1 and 6.3, McGraw–Hill, New York, 1967.

[6] An interesting exercise is to plot the electrical conductivity $1/\rho$ against the thermal conductivity (see Exercise 30 in Appendix G).

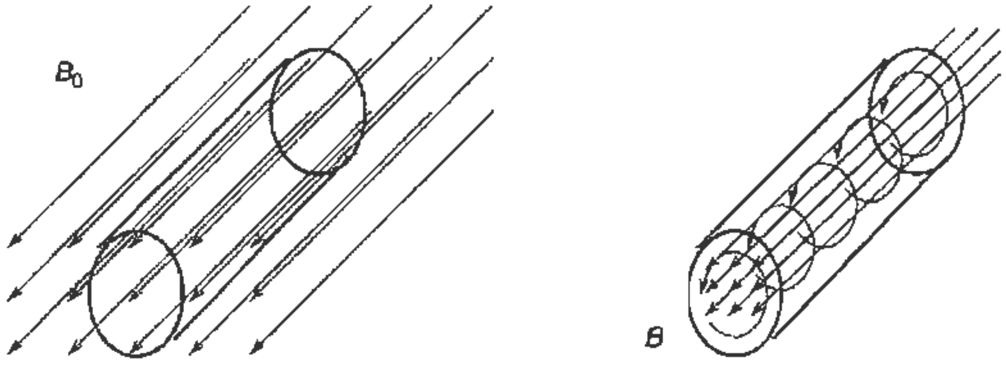
in a *pure* metal crystal. If there are impurities, then the scattering will contain an additional contribution. We can write

$$\frac{1}{\tau} = \frac{1}{\tau_{\text{CRYSTAL}}} + \frac{1}{\tau_{\text{IMPURITY}}}.$$

The scattering from the crystal depends crucially on the vibrational energy stored in the crystal lattice, and therefore on temperature. The impurity scattering is essentially independent of temperature.

The technique we use measures resistivity directly. The idea is based on Faraday's law, which gives the EMF (i.e., voltage) induced in a coil that surrounds a magnetic field that changes with time. That is, we measure a signal $V(t)$ that is proportional to some dB/dt. The magnetic field B is generated by the "eddy currents" left in a metallic sample when the sample is immersed in a constant magnetic field that is rapidly switched off. Figure 2.9 shows how this is done. In Fig. 2.9a, a cylindrical metallic bar is placed in a constant magnetic field whose direction is along the axis of the cylinder. We assume the bar is not ferromagnetic, so the magnetic field inside is essentially the same as it is outside. Remember that the bar is filled with electrons that are essentially free to move within the metal.

Now we shut the field off abruptly. By Faraday's law, the electrons in the metal will move and generate a current that tries to oppose the change in the external magnetic field. These so-called eddy currents are loops in the plane perpendicular to the axis of the sample, and they generate a magnetic



(a) Field on

(b) Field shut off

FIGURE 2.9 The eddy current technique for measuring resistivity. (a) A magnetic field B_0 permeates a cylindrical metal sample. (b) Eddy currents set up when the field is shut off generate a field B of their own. The eddy currents, and therefore B, decrease with time at a rate that depends on the resistivity.

field of their own. See Fig. 2.9b. However, as soon as the external field is gone, there is nothing left to drive these eddy currents, and they start to decay away because of the finite resistivity of the metal. The time it takes for the currents to decay away is directly related to the resistivity, as we shall see.

We again use Faraday's law to detect the decaying eddy currents. The magnetic field set up by the eddy currents also decays away with the same time dependence as the currents. Therefore, if we wrap a coil around the sample, Faraday's law says that an induced EMF shows up as a voltage drop across this coil. This voltage drop is the signal, and the rate at which it decays to zero is a measure of the resisitivity of the metal sample.

In order to determine the voltage signal as a function of time, one needs to solve Maxwell's equations in the presence of the metal. The derivation is complicated, but outlined in Bean *et al.* (1959), where a series solution is obtained by expanding in exponentials. For a cylindrical rod, this series takes the form

$$V(t) \propto \sum_{i=1}^{\infty} \exp\left(-\lambda_i^2 \alpha t\right),$$

where α is proportional to ρ and the λ are roots of the zero-order Bessel function, i.e., $\lambda_1 = 2.405$, $\lambda_2 = 5.520$, $\lambda_3 = 8.654$, and so on. Since the λ increase with each term, for long enough times, only the first term is significant because all the rest die away much faster. That is, the falloff of $V(t)$ with time will look like a single exponential if one waits long enough, but will be more complicated at shorter times.

For a cylindrical metal sample where the external magnetic field points along the axis of the cylinder, the result is

$$V(t) = V_0 e^{-t/t_E}, \tag{2.15}$$

where

$$t_E = 2.17 \times 10^{-9} \left[\frac{\Omega \cdot \mathrm{s}}{\mathrm{cm}}\right] \frac{r^2}{\rho}, \tag{2.16}$$

$$V_0 = 10 N \rho B_0, \tag{2.17}$$

and $t = 0$ is the time when the external field is switched off. In this equation, r is the radius of the cylinder, expressed in centimeters, and ρ is the resistivity of the metal, expressed in ohms-centimeters. Also, N is the number of turns in the detector or "pickup" coil and $B_0 = \mu_0 in$ (in SI units) gives the magnetic field B_0 set up by a solenoid carrying a current i through n turns. *This equation is only valid for times t on the order of t_E or larger.* At earlier times, there are transient terms left over that cause $V(t)$ to fall off more rapidly than given by Eq. (2.15).

2.2.1. Measurements

The lifetime t_E given by Eq. (2.16) is on the order of tenths of milliseconds. Therefore, the magnetic field must be switched off considerably more rapidly than that. This is hard to do mechanically, so we will resort to an electrical switch, using a transistor.[7] The circuit that produces the switching magnetic field is shown in Fig. 2.10.[8] A garden variety 6-V/2-A power supply puts current through the solenoid, creating the magnetic field B_0. However, after passing through the solenoid, the current encounters a transistor (321/TIP 122) instead of passing directly back to ground. The lead out of the solenoid is connected to the collector of the transistor, and the emitter is connected to ground. The base is connected through a 1-kΩ resistor to the 600-Ω output of the HP 3311A waveform generator. The waveform generator is set to produce a square wave, oscillating between around -10 V and $+10$ V with a period of a few milliseconds.

Consider the current through the solenoid. First, the DC power supply is connected so that the solenoid is always positive with respect to ground, thus the collector voltage is always above the emitter voltage. Second, the base-emitter acts like a conducting diode, so there will be a voltage drop across it of around 0.6 V when it conducts. Also, if there is no current through the base, then the base-collector junction is reversed biased and no current flows through the transistor, or therefore through the solenoid. That is, the switch is off. Now when the waveform generator is at $+10$ V, the current through the base is $i_B \approx 10\text{ V}/1\text{ k}\Omega = 10\text{ mA}$. This turns the switch on and lets the current flow through the solenoid pretty much as if the transistor wasn't there, so long as $I_C \ll \beta I_B = 10$ A. You might want to measure the resistance in the solenoid coil to make sure it does not

[7] This transistor is actually a "Darlington pair," which effectively gives a single transistor with a gain parameter $h_{FE} = \beta = 1000$ or so. $V_{CE} = 6$ V does not exceed the specifications.

[8] For students with minimal experience in laboratory electronics, Sections 3.1, 3.2, and 3.3 should be consulted.

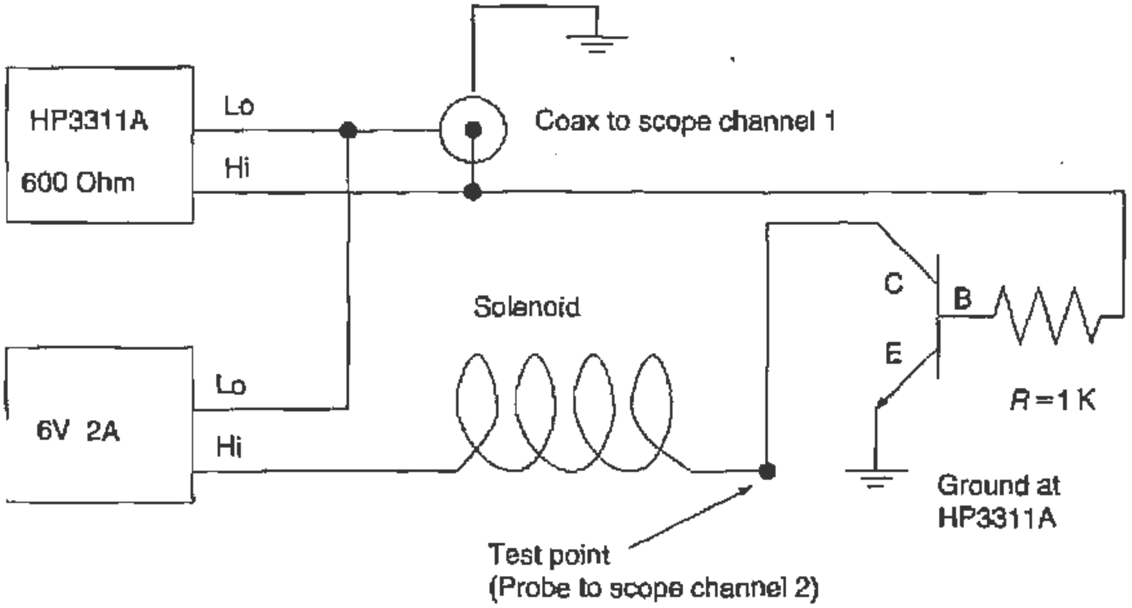


FIGURE 2.10 Switching circuit for turning the magnetic field on and off. It is a good idea to check the current through the solenoid by measuring the voltage at the testpoint, timed against the HP3311A square wave generator.

draw a lot of current, but since you are using a 2-A power supply, it is a good bet that you are in the clear. So, when the square wave generator is at +10 V, the solenoid conducts. However, when the generator switches to −10 V (or presumably anything less than around 0.6 V), the solenoid and the magnetic field shut off. This is, $t = 0$ in Eq. (2.15).

The pickup coil is wound on a separate tube, which can be inserted inside the solenoid. One can then introduce and remove different metal samples from inside the pickup coil. By connecting the terminals of the pickup coil to a digital oscilloscope, we record values of $V(t)$ corresponding to Eq. (2.15). There is one complication. The magnetic field shuts off so fast that the instantaneous induced voltage in the pickup coil is very large. That is, Δt is so small that $dB/dt \approx \Delta B/\Delta t$ and therefore also V are very large. An oscilloscope would typically have circuitry that protects it, but one should take some care to avoid damaging the equipment. To fix this problem, the simple circuit shown in Fig. 2.11 is used to connect the pickup coil terminals to the oscilloscope input. The two diodes are arranged so that any current is taken to ground, so long as the voltage is bigger than +0.6 V or smaller than −0.6 V, for diodes with $V_F = 0.6$ V. That is, the circuit "clamps" the input to the oscilloscope so that it never gets more negative, but still big enough to make the measurement.

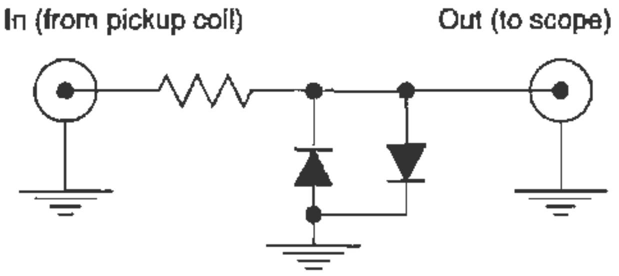


FIGURE 2.11 Clamping circuit for the oscilloscope input.

Sometimes we see the signal "ring" just as the switch shuts off. That is, we see the decaying exponential but a rapid oscillation[9] is superimposed on it, and this gets in the way of measuring the decay time. If the ringing goes away while the signal is still decaying exponentially, just use the data past the point where the ringing is gone. Otherwise, a resistor should be attached in parallel with the scope input. It is best if you can get a variable resistor, and play with the values so that the exponential decay is unaffected but the ringing is thoroughly damped out.

Before measuring the resistivity, one should know what the solenoid circuit is doing. Connect a probe to the junction between the solenoid and the transistor collector. View this on the other channel of the oscilloscope, and confirm that you see what you expect. That is, when the square wave is high, the solenoid is conducting and the voltage at this point should be around $+1.2$ V, i.e., the sum of the two forward voltage drops for the CB and BE diode equivalents for the transistor. On the other hand, when the square wave is low, the solenoid should not be conducting and there is no voltage drop across it, so the voltage at this junction should be around $+6$ V, i.e., the voltage of the DC power supply. This probe should now be removed since the oscilloscope channel is needed to make the resistivity measurements. Next, connect the pickup coil to the clamping circuit and plug it into the second channel of the scope. Do not put any metal sample in just yet. You should see a voltage spike, alternatively positive and negative, when the magnetic field switches on and off, clipped by the diode clamping circuit.

Now insert a sample into the pickup coil. Watch the pickup coil signal on the scope as you do this. The effect of the decaying eddy currents

[9] The circuit has lots of "loops," each of which is essentially an inductor. Any capacitance somewhere will cause oscillations, but the exact source can be hard to pin down. One should take care to wind the pickup coil in a way that minimizes the inherent capacitance. A good way to do this is to crisscross the windings of each layer.

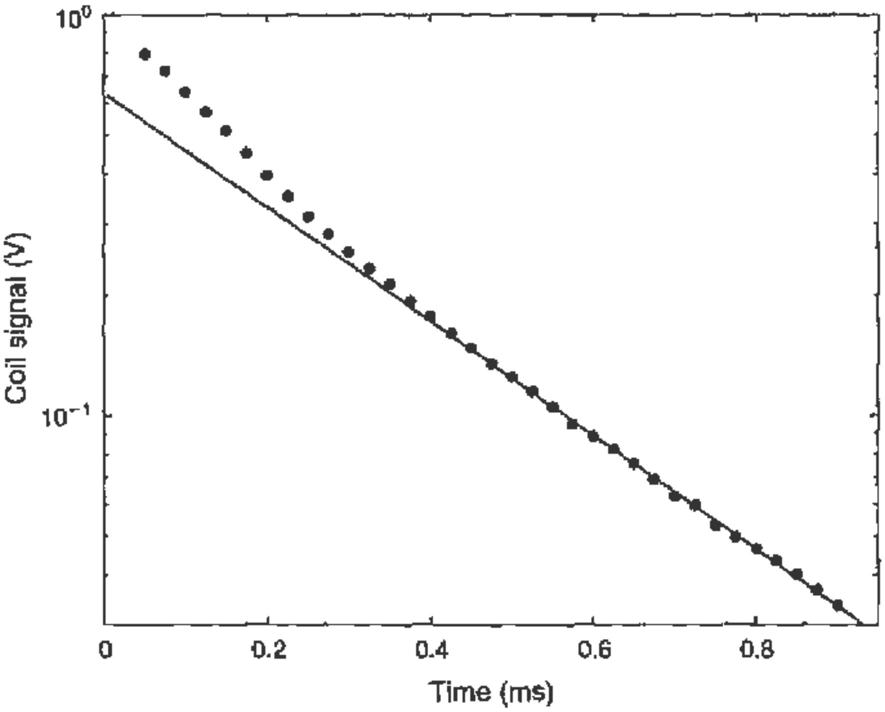


FIGURE 2.12 Resistivity data taken with a high purity aluminum rod as the sample. The decay is clearly not described by a single exponential at the earlier times.

should be clear. You may see some transient oscillations of the signal right after the field shuts off, but there should be plenty of time left after these oscillations die away for you to get a smooth curve. Figure 2.12 shows data acquired with a $\frac{1}{2}$-in. diameter high-purity aluminum rod[10] at room temperature as a sample. The data points are the output of a digital oscilloscope displayed using MATLAB. Note that at the earliest times, there are higher order contributions to the signal (as described by Bean *et al.*), and one must choose a suitable range over which the data are indeed described by a single exponential.

The fit shown in Fig. 2.12 yields a decay time $t_E = 3.051 \times 10^{-4}$ s. Then, from Eq. (2.16) we find for the resistivity

$$\rho = \frac{2.17 \times 10^{-9}}{t_E\ (\mathrm{s})} \times r^2\ (\mathrm{cm}^2) = 2.87 \times 10^{-6}\ \Omega \cdot \mathrm{cm},$$

where we used the fitted value of t_E and $r = 0.635$ cm. This compares well with the value listed in Table 2.1.

[10] From the Alfa Aesar company, http://www.alfa.com/.

The main source of systematic uncertainty is likely to come from the times over which the decaying voltage signal is fitted. At short times, the decay is not a pure exponential because the transient terms have not all died away, so we want to exclude these times when we fit. At long times, there may be some left over voltage level that is a constant added to the exponential, and again, a pure exponential fit will be wrong. Varying the upper and lower fit limits until we get a set that gives the same answer as a set that is a little bit larger on both ends is one approach. One should be convinced that the results are consistent. For example, use aluminum alloy rods of the same composition but different radii, and check to make sure that the decay lifetimes t_E scale like r^2. This should certainly be the case to within the estimated experimental uncertainty.

Having learned how to take and analyze data on resistivity, we can now investigate the temperature dependence. It is best to start simply by comparing the two samples of $\frac{1}{2}$-in. diameter aluminum rods, one an alloy and the other a (relatively) pure metal. Vary the temperature by immersing the samples in baths of ice water, dry ice and alcohol, and liquid nitrogen. Boiling water or hot oil can also be used. These measurements are tricky. One must remove the sample from the bath and measure the eddy current decay before the temperature changes very much. Probably the best way to do this is to take a single trace right after inserting the sample, stop the oscilloscope, and store the trace. Then one analyzes the trace offline to get the decay constant. One might also try to estimate how fast the bar warms up by making additional measurements after waiting several seconds, e.g., after saving the trace. This would best be done with a sample whose resistivity, and therefore t_E, can be expected to change a lot with temperature. Pure aluminum is a good choice. Remember that the temperature dependence will be much different for the pure metal than for the alloy. Try to estimate the contribution to the mean free path of the electrons due to the impurities.

2.3. EXPERIMENT ON THE HALL EFFECT

In Section 2.2 we saw how collisions of electrons with the crystal lattice lead to an electrical resistance, when those electrons are forced to move under an electric field. If one also applies a magnetic field, in a direction perpendicular to the electric field, then the electrons (and other current carriers) will be deflected sideways. As a result an electric field appears in this direction, and therefore also a potential difference. This phenomenon is called the Hall effect, and has important applications both in identifying the current carriers in a material and for practical use as a technique for measuring magnetic fields.

Let us rewrite the microscopic formula for Ohm's law, but this time taking care to indicate current density and electric fields as vectors, and to also note the negative sign of the charge on the electron. Following Eqs. (2.12) and (2.13) we write

$$\mathbf{j} = -ne\mathbf{v}_\mathrm{d} = ne^2\tau\mathbf{E}/m \tag{2.18}$$

or

$$\frac{m\mathbf{v}_\mathrm{d}}{\tau} = -e\mathbf{E}. \tag{2.19}$$

It is clear that in Eq. (2.19) we have made an approximation, replacing the time rate of change of momentum, i.e., $d\mathbf{p}/dt = md\mathbf{v}/dt$, with an expression that uses the average acceleration $\mathbf{v}_\mathrm{d}/\tau$. This is how we have taken into account collisions with the crystal lattice.

It is straightforward to modify Eq. (2.19) to take into account the effect of a magnetic field $\mathbf{B}$. We have

$$\frac{m\mathbf{v}_\mathrm{d}}{\tau} = -e(\mathbf{E} + \mathbf{v}_\mathrm{d} \times \mathbf{B}).$$

If we assume that the magnetic field lies in the z direction, and define the cyclotron frequency $\omega_\mathrm{c} \equiv eB/m$, then we can rewrite this equation as

$$\begin{aligned} v_{\mathrm{d}_x} &= -\frac{e\tau}{m}E_x - \omega_\mathrm{c}\tau v_{d_y} \\ v_{\mathrm{d}_y} &= -\frac{e\tau}{m}E_y + \omega_\mathrm{c}\tau v_{d_x} \\ v_{\mathrm{d}_z} &= -\frac{e\tau}{m}E_z. \end{aligned} \tag{2.20}$$

Consider now a long rectangular section of a conductor, as shown in Fig. 2.13. A longitudinal electric field E_x is applied, leading to a current density flowing in the x direction. As this electric field is initially turned on, the magnetic field deflects electrons along the y direction. This leads to a buildup of charge on the faces parallel to the xz plane, and therefore an electric field E_y within the conductor. In the steady state, this electric field cancels the force due to the magnetic field, and the current density is strictly

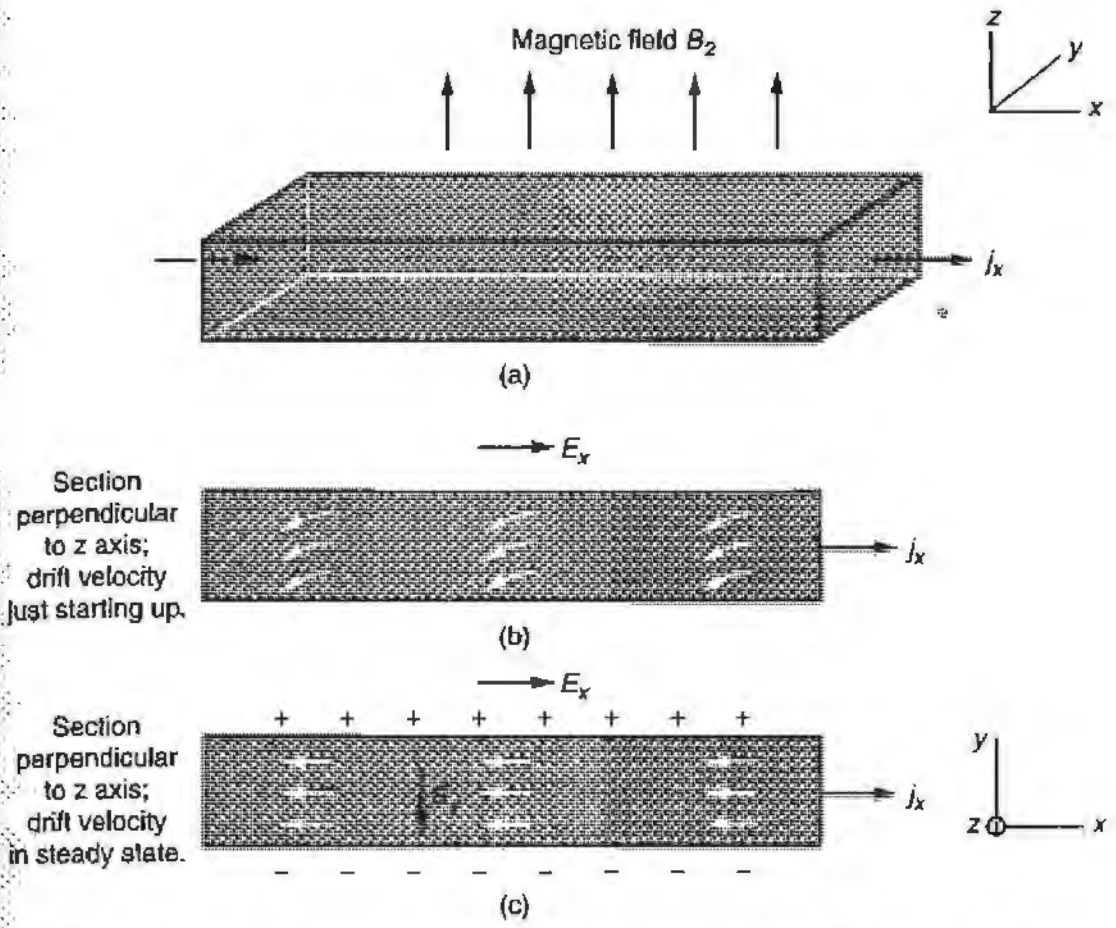

Magnetic field B_2
z
y
x
j_x
(a)
E_x
Section perpendicular to z axis; drift velocity just starting up.
j_x
(b)
E_x
+ + + + + + + +
Section perpendicular to z axis; drift velocity in steady state.
E_y
j_x
y
z
x
− − − − − − − −
(c)

FIGURE 2.13 The standard geometry for discussing the Hall effect (after Kittel).

in the x direction, hence $v_{d_y} = 0$. From Eqs. (2.20) we therefore have

$$E_y = \frac{m\omega_c}{e} v_{d_x} = \frac{m\omega_c}{e}\left(-\frac{e\tau}{m}E_x\right) = -\omega_c \tau E_x = -\frac{eB\tau}{m}E_x.$$

The appearance of the electric field E_y is the Hall effect.

A convenient experimental quantity is the Hall coefficient R_H, defined as

$$R_H \equiv \frac{E_y}{j_x B} \tag{2.21}$$

The quantities E_y, j_x, and B are all straightforward to measure, and in our simple approximation for electrons in conductors we have (from Eq. (2.18)) $j_x = ne^2 \tau E_x / m$; therefore,

$$R_H = \frac{eB\tau E_x/m}{(ne^2\tau E_x/m)B} = \frac{1}{ne}. \tag{2.22}$$

That is, the Hall coefficient is the inverse of the carrier charge density. In fact, the Hall effect is a useful way to measure the concentration of charge carriers in a conductor. It is also convenient to define the Hall resistivity as the ratio of the transverse electric field to the longitudinal current density, that is,

$$\rho_{\mathrm{H}} \equiv E_y/j_x = BR_{\mathrm{H}}, \tag{2.23}$$

which depends (in our approximation) only on the material and the applied magnetic field.

2.3.1. Measurements

In order to measure the Hall effect, one needs a sample of a conductor, but not an especially good conductor. This is because one also needs a relatively low carrier density ne in order to get a sizable effect; this of course leads to a relatively high resistivity. As seen in Table 2.1, bismuth is a good candidate metal, and we describe such an experiment here.[11]

The setup uses a bismuth sample with rectangular cross section, mounted on a probe with attached leads for measuring current and voltage. A thermocouple is also attached to the sample so that temperature measurements can be carried out. The magnetic field is provided by an electromagnet capable of delivering a field up to ~5 kG over a volume roughly 1 cm^3. The bismuth sample probe is shown in Fig. 2.14. The width of the bismuth sample is $w = 6.5$ mm and its thickness, measured with a micrometer, is $t = 1.65 \times 10^{-4}$ m. The effective length of the sample is the distance between the leads used to measure the current ("white" and "brown," as shown in Fig. 2.14). In our case, this distance is $\ell = 7$ mm. Current is supplied by a DC power supply, connected to the sample through the "red" and "black" leads. The Hall voltage is measured with a digital multimeter, using the "green" lead and the output of a potentiometer used to balance the voltage on the "white" and "brown" leads. A separate bundle of wires are connected to leads that carry current to the heating resistor, and to a thermocouple that measures the temperature of the bismuth sample.

Begin by determining the Hall coefficient at room temperature and for a relatively high magnetic field. Turn on the electromagnet power supply to

[11] Semiconductors also make good candidates, with a very low carrier density compared to a metal. For a description of such a setup, see A. Melissinos, *Experiments in Modern Physics*, First ed., Academic Press, New York, 1966.

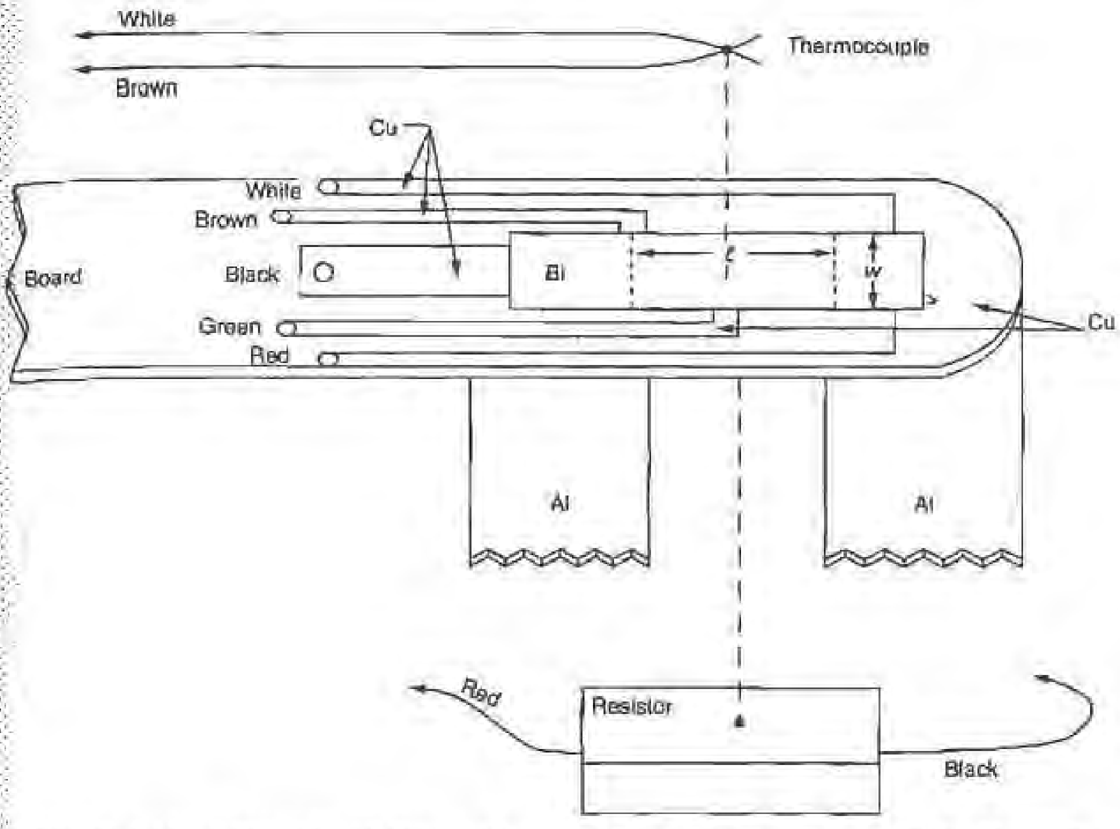


FIGURE 2.14 Schematic of the probe used to make measurements of the Hall effect in bismuth. Electrical connections are made to the bismuth sample using copper leads. A thermocouple, as well as a resistor which acts as a heat source, is also attached to the sample. Two separate bundles of wires emerge from the probe, one of which is used exclusively for heating the sample and for measuring its temperature.

around 4 kG. It will likely need an hour or so to stabilize. In the meantime, with the sample probe removed from the magnetic field, run about 3 A through the bismuth sample, and adjust the potentiometer so that the Hall voltage is zero. Return the current through the sample to zero. *The sample can get quite hot while it is conducting so much current. Be careful not to touch it, or to touch it to anything else.*

When the electromagnet is stabilized, measure and record the magnetic field using a gaussmeter, or by some other technique. Now, place the sample probe in the center of the magnetic field. Quickly raise the current I through the sample to 3.0 A, and record the Hall voltage V_H. Then, quickly, reduce the current by 0.25 A, and record the Hall voltage again. You should carry this series of measurements out rather rapidly to avoid leaving the bismuth sample at high temperature for any extended period of time. When you have reduced the current to near zero, and recorded the final value of the

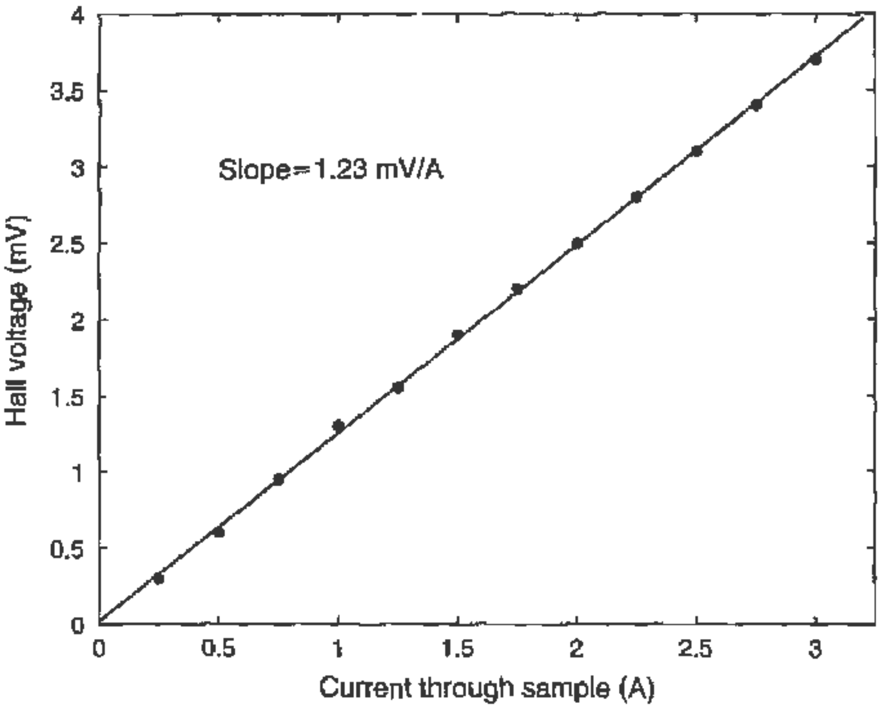


FIGURE 2.15 Sample of Hall effect data, taken at room temperature and with a magnetic field $B = 4.42$ kG.

Hall voltage, remove the probe and recheck the value of the magnetic field.

A sample of data taken in this way, at room temperature and with $B =$ 4.42 kG, is shown in Fig. 2.15. A free linear straight line fit gives a slope of 1.23 mV/A, with an intercept very close to zero. In terms of quantities related to our measurement, the Hall coefficient (Eq. (2.21)) is expressed by

$$R_H \equiv \frac{E_y}{j_x B} = \frac{V_H/w}{I/(w \times t)B} = \frac{V_H t}{IB} = \frac{dV_H}{dI}\frac{t}{B},$$

where we note that our data yields a very good direct proportional relationship between V_H and I. Using SI units, this yields

$$R_H = \left(1.23 \times 10^{-3}\frac{\text{V}}{\text{A}}\right)\left(\frac{1.65 \times 10^{-4}\ \text{m}}{0.442\ \text{T}}\right) = 4.59 \times 10^{-7}\ \text{m}^3/\text{C}$$

This is quite close to an accepted room temperature value of $R_H = 5.4 \times 10^{-7}$ m^3/C for pure bismuth metal. The uncertainties in measuring the dimensions of the sample can easily account for the discrepancy.

Of course, this sample and this setup can be used to determine the resistivity of bismuth. Outside of the magnetic field, measure the voltage

TABLE 2.2 Sample data, taken by a student, for the resistivity ρ of bismuth as a function of temperature, using the Hall effect apparatus

T (°C)	T (K)	ρ (μΩ-cm)
−80	193	70
−60	213	85
−40	233	96
−20	253	110
0	273	121
20	293	134
40	313	150
60	333	163

drop along the length ℓ of the bismuth sample, as a function of the applied current, and determine the resistivity ρ from the ratio

$$\rho = \frac{E_x}{j_x} = \frac{dV_x}{dI}\frac{wt}{\ell}.$$

The temperature dependence of each of these quantities can be determined by heating (and cooling) the probe, and recording values as a function of temperature using readings from the thermocouple.

Table 2.2 lists some results for the resistivity ρ in (μΩ-cm) as a function of temperature. To examine the temperature dependence it is best to make a log–log plot of the data vs T since we expect a power law dependence. This is shown in Fig. 2.16 and when fitted gives

$$\rho \propto T^{1.52}.$$

Note that at room temperature ($T = 25°\text{C}$)

$$\rho = 1.4 \times 10^{-4}\ \Omega\text{-cm}$$

in reasonable agreement with the data of Table 2.1.

Indeed, one expects a $T^{3/2}$ dependence of the resistivity on the temperature because of the following argument. From Eq. (2.14) the resistivity is inversely proportional to the mean time between collisions, as long as the carrier density remains constant. Now the mean time between collisions is given by

$$\tau = \lambda/v,$$

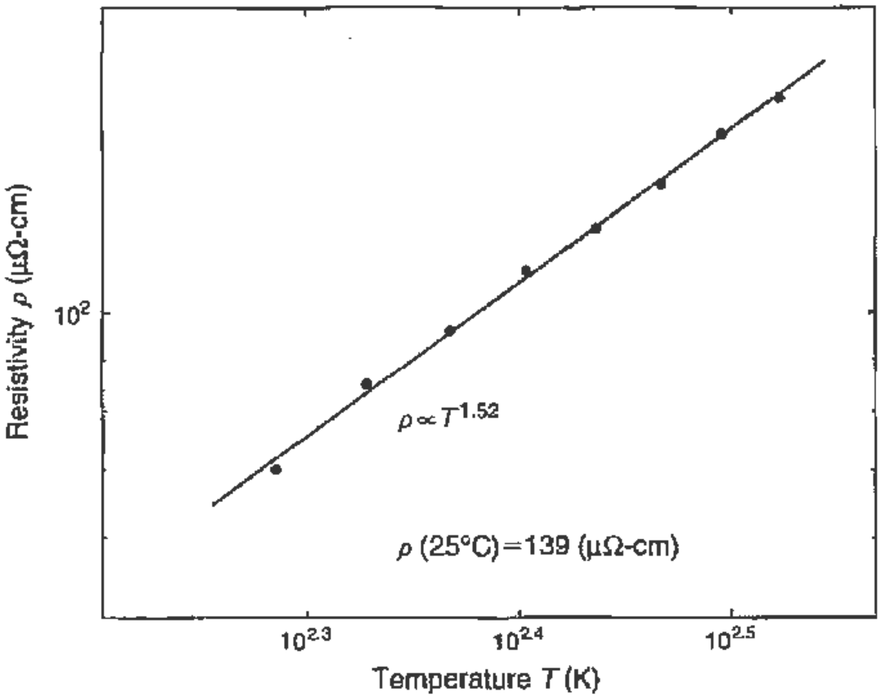


FIGURE 2.16 The resistivity of bismuth as a function of temperature, taken with the Hall effect apparatus (data from Table 2.2.) The data are fitted to a power law form.

where λ is the mean free path for scattering, and v the thermal velocity of the electrons. For v we can use

$$\frac{1}{2}mv^2 = \frac{3}{2}kT \qquad \text{or} \qquad v = \sqrt{3kT/m}.$$

The mean free path, λ, decreases as the collision cross section increases, namely as the lattice vibrations increase with temperature. It is found that λ is inversely proportional to the temperature, and therefore

$$\tau \propto 1/T^{3/2}$$

or using Eq. (2.14),

$$\rho \propto T^{3/2}.$$

We can also examine the temperature dependence of the Hall coefficient. In this case it is best to plot R_{H} on a semi-log plot vs $1/T$. The reason is that the Hall coefficient (see Eq. (2.22)) is directly inversely proportional to the carrier density, and we expect the carrier density to depend on the temperature by an exponential factor, such as for instance shown in Eq. (2.28). The data are plotted in this way in Fig. 2.17, and we recognize two distinct slopes. As expected, R_{H} falls with increasing temperature

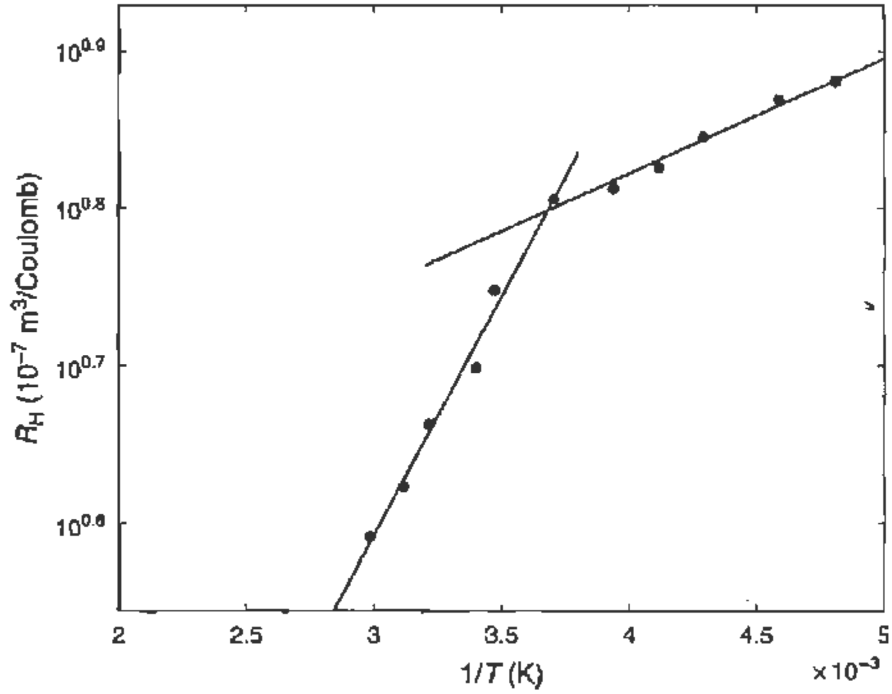


FIGURE 2.17 Measurements of the Hall coefficient as a function of temperature.

because the carrier density increases. By fitting the data to the form

$$n \propto \exp(-E/2kT),$$

we find for the two regions

$$\begin{array}{ll} \text{low } T, & E = 0.029\text{ eV} \\ \text{high } T, & E = 0.120\text{ eV}. \end{array}$$

Such energy differences are typical of the excitation of impurities. It is also relevant to note that the carrier density at room temperature is

$$n = 1/eR_H = 1.35 \times 10^{19}\,\text{cm}^{-3}.$$

This is quite high and typical of a conductor.

2.4. SEMICONDUCTORS

2.4.1. General Properties of Semiconductors

We have seen in the first section how a free-electron gas behaves, and what can be expected for the band structure of a crystalline solid. In the second section we applied the model of a free-electron gas to the behavior of the resistivity of metals. In the present section we will study some properties of semiconductors that can be verified easily in the laboratory, where we will make use both of the free electron gas model and of the band structure of the material. As mentioned before, a semiconductor is a crystalline solid in which the conduction band lies close to the valence band, but is not populated at low temperatures; semiconductors are unlike most metals in that both *electrons* and *holes* are responsible for the properties of the semiconductor. If the semiconductor is a pure crystal, the number of holes (positive carriers, p) is equal to the number of free electrons (negative carriers, n), since for each electron raised to the conduction band, a hole is created in the valence band: these are called the *intrinsic* carriers. All practically important semiconductor materials, however, have in them a certain amount of impurities that are capable either of donating electrons to the conduction band (making an n-type crystal) or of accepting electrons from the valence band, thus creating holes in it (making a p-type crystal). These impurities are called *extrinsic* carriers and in such crystals $n \neq p$.

Let us then first look at the energy-band picture of a semiconductor as it is shown in Fig. 2.18; the impurities are all concentrated at a single energy level usually lying close to, but below, the conduction band. The density of states must be different from that of a free-electron gas (Eq. (2.4) and Fig. 2.2a) since, for example, in the forbidden gaps it must be 0; close to the ends of the allowed bands it varies as $E^{1/2}$ and reduces to 0 on the edge.

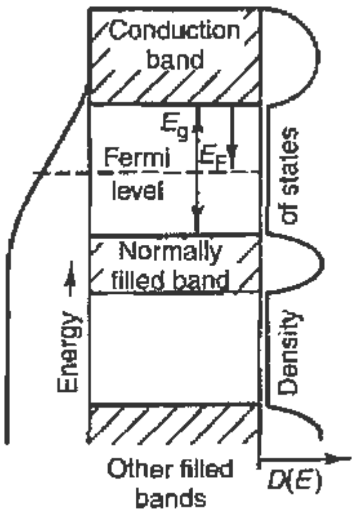


FIGURE 2.18 Energy band structure of a semiconductor without impurities. On the left-hand side the Fermi distribution for a free-electron gas is shown; on the right-hand side the actual density of states $D(E)$ is shown.

On the other hand, the Fermi distribution function, Eq. (2.3), remains the same. The only parameter in this function is the Fermi energy, which can be found by integrating the number of *occupied* states (Fermi function times density of states) and setting it equal to the electron density. It is clear, however, that if we are to have as many empty states in the valence band as occupied ones in the conduction band, the Fermi level must lie exactly in the middle of the forbidden gap[12] (because of the symmetry of the trailing edge of the distribution). In Fig. 2.18, the density of states is shown to the right and the Fermi distribution function to the left. We measure the position of the Fermi level *from the conduction band* and define it by E_{F}; the exact value of E_{F} is

$$E_{\mathrm{F}} = -\frac{E_{\mathrm{g}}}{2} + kT \ln \left(\frac{m_{\mathrm{h}}^*}{m_{\mathrm{e}}^*} \right)^{3/4} . \tag{2.24}$$

Since the Fermi level lies below the conduction band, E_{F} is a *negative* quantity, E_{g} is the energy gap always taken to be positive, and m_h^* and m_e^* are the effective masses of holes and electrons, respectively. If w_{C} and w_{F} stand for the actual position of the conduction band and Fermi level above the zero point energy, then

$$w_{\mathrm{F}} = w_{\mathrm{C}} + E_{\mathrm{F}}.$$

To find the density of electrons in the conduction band (or holes in the valence band) we simply substitute Eq. (2.24) for w_{F} into Eq. (2.4), multiply by the density of states, and integrate over w from $w = w_{\mathrm{C}}$ to $+\infty$. When, however, the exponent

$$-(w_{\mathrm{F}} - w) \approx \frac{E_{\mathrm{g}}}{2} + E \gg kT, \tag{2.25}$$

the Fermi distribution degenerates to a Boltzmann distribution. (Here E is the energy of the electrons as measured from the top of the conduction band; obviously it can take either positive or negative values.) With this *assumption* the integration is easy, yielding

$$n = \left(\frac{2\pi m_e kT}{h^2} \right)^{3/2} e^{E_{\mathrm{F}}/kT} \approx \left(\frac{2\pi m_e kT}{h^2} \right)^{3/2} e^{-E_{\mathrm{g}}/2kT}; \tag{2.26}$$

[12] If the effective masses of p- and n-type carriers are the same.

similarly,

$$p = \left(\frac{2\pi m_h kT}{h^2}\right)^{3/2} e^{-(E_g+E_F)/kT} \approx \left(\frac{2\pi m_h kT}{h^2}\right)^{3/2} e^{-E_g/2kT}. \quad (2.27)$$

It is interesting that the product np is independent of the position of the Fermi level[13]—especially if we take $m_e = m_h$

$$n_i^2 = np = 2.31 \times 10^{31} T^3 e^{-E_g/kT}.$$

From the analysis we expect that as the temperature is raised, the density of the intrinsic carriers in a semiconductor will increase at an exponential rate characterized by $E_g/2kT$. This temperature is usually very high since $E_g \approx 0.7$ V (see Eqs. (2.29)).

We have already mentioned that impurities determine the properties of a semiconductor, especially at low temperatures where very few intrinsic carriers are populating the conduction band. These impurities, when in their ground state, are usually concentrated in a single energy level lying very close to the conduction band (if they are donor impurities) or very close to the valence band (if they are acceptors). As for the intrinsic carriers, the Fermi level for the impurity carries lies halfway between the conduction (valence) band and the impurity level; this situation is shown in Figs. 2.19a and 2.19b. If we make again the low temperature approximation of Eq. (2.25), the electron density in the conduction band is given by

$$n = N_d \left(\frac{2\pi mkT}{h^2}\right)^{3/2} e^{-E_d/2kT}, \quad (2.28)$$

where N_d is the donor density and E_d the separation of the donor energy level from the conduction band. In writing Eq. (2.28), however, care must be exercised because the conditions of Eq. (2.25) are valid only for very low temperatures. Note, for example, that for germanium

$$E_g = 0.7 \text{ eV}, \quad \text{and for } kT = 0.7 \text{ eV}, \quad T = 8000 \text{ K}$$

whereas

$$E_d = 0.01 \text{ eV}, \quad \text{and for } kT = 0.01 \text{ eV}, \quad T = 120 \text{ K}. \quad (2.29)$$

Thus at temperatures $T \approx 120$ K most of the donor impurities will be in the conduction band and instead of Eq. (2.28) we will have $n \approx N_d$; namely,

[13] This result is very general and holds even without the approximation that led to Eqs. (2.26) and (2.27).

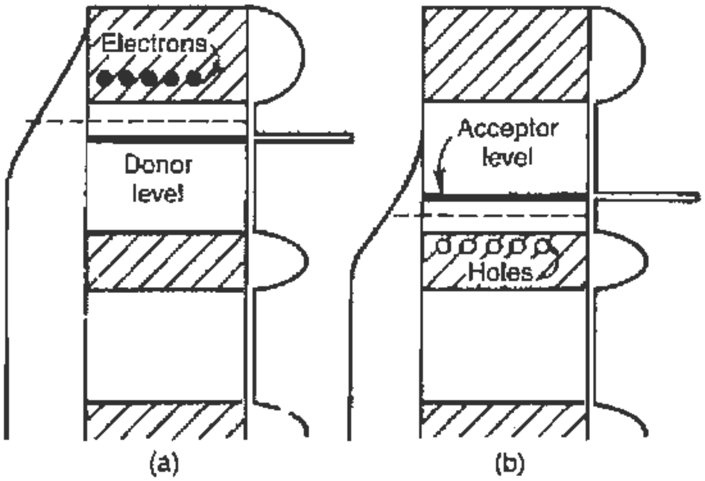


FIGURE 2.19 Same as described in the legend to Fig. 2.18 but with the addition of impurities. (a) The impurities are of the donor type and lie at an energy slightly below the conduction band. (b) The impurities are of the acceptor type and lie slightly above the valence band. Note the shift of the Fermi level as indicated by the dotted line.

the density of impurity carriers becomes saturated. Once saturation has been reached the impurity carriers in the conduction band behave like the free electrons of a metal.

2.4.2. Sketch of *p–n* Semiconductor Junction Theory

Semiconductor materials with high impurity concentration, when properly combined, form a transistor. Junction transistors consist of two junctions of dissimilar-type semiconductors, one *p* type and one *n* type; the intermediate region, the base, is usually made very thin. We will briefly sketch the behavior of such a *p–n* junction and then see how the combination of two junctions can provide power amplification; for this we will use our knowledge of the band structure of semiconductors and the position of the Fermi level, as developed previously (Figs. 2.18 and 2.19). When two materials with dissimilar band structure are joined, it is important to know at what relative energy level one band diagram lies with respect to the other: the answer is that *the Fermi levels of both materials must be at the same energy position* when no external fields are applied; this is shown in Fig. 2.20.

From the energy diagram of Fig. 2.20, it follows that only electrons with $E_e > \Delta W_e$ will be able to cross the junction from the *n* material into the *p* region and only holes with $E_h > \Delta W_h$ from the *p* region into the *n* region.

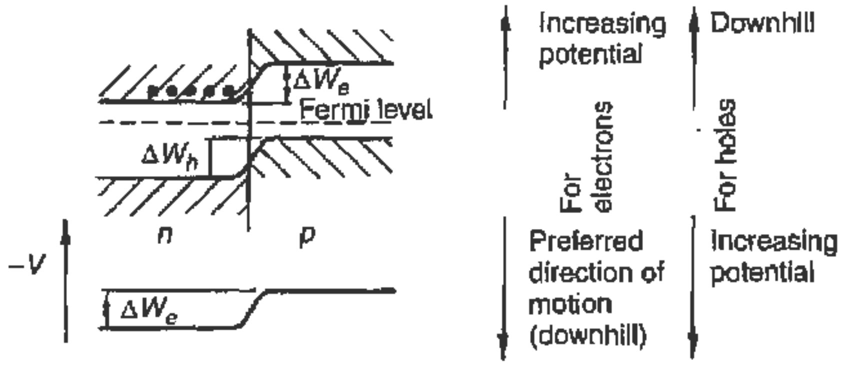


FIGURE 2.20 Structure of the energy bands at the junction of an n-type and a p-type semiconductor.

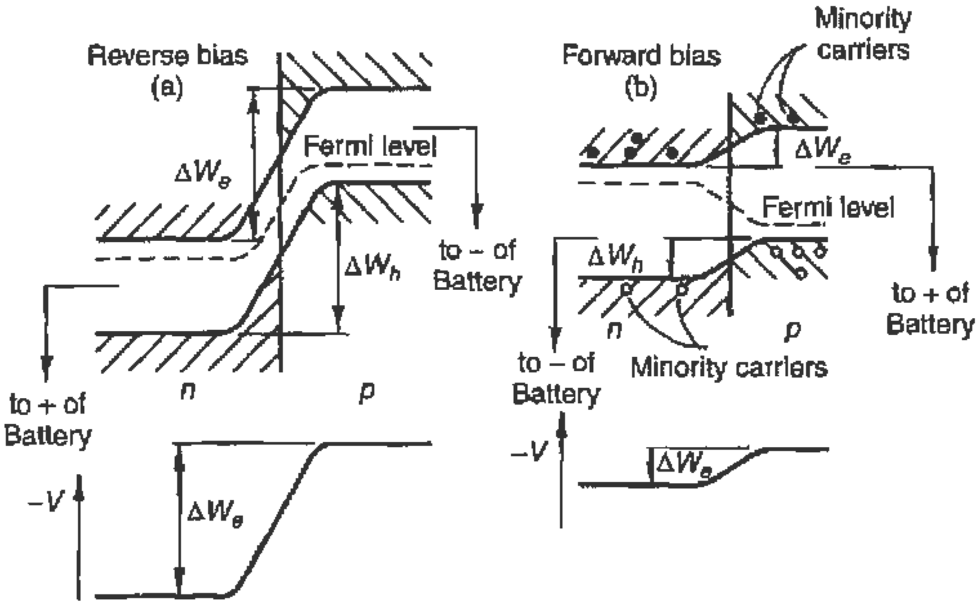


FIGURE 2.21 Structure of the energy bands at a biased n–p junction: (a) reverse bias and (b) forward bias. The solid dots represent electrons, whereas the open circles holes.

Holes in the n region or electrons in the p region are called "minority carriers." Indeed, there will be diffusion of some minority carriers across the junction, but since no electric field is present these carriers will remain in the vicinity of the junction.[14]

If now a *reverse bias* is applied—that is, one that opposes the further motion of the minority carriers—the Fermi levels will become displaced by the amount of the bias, as shown in Fig. 2.21a. We see that the barriers

[14] The result of such diffusion is the buildup of a local charge density, which prevents further diffusion. Throughout the present analysis, however, we will neglect the local effects at the junction.

ΔW_e and ΔW_h are increased by almost the full voltage, making any motion of minority carriers across the junction very improbable. Figure 2.21b, on the other hand, shows the situation when *forward bias* is applied (favoring the motion of minority carriers). The Fermi levels are now displaced in the opposite direction so that the barriers are lowered. However, the full bias voltage does not appear as a difference between the Fermi levels because dynamic equilibrium prevails. There is a continuous flow of minority carriers in the direction of the electric field (holes obviously moving in the opposite direction from electrons) and as a result a potential gradient exists along the material; thus the entire bias voltage does not necessarily appear at the junction itself.

We will now consider two junctions put together; in Fig. 2.22a, p-type, n-type, and again p-type material are joined. When no bias is applied, we expect the Fermi levels to be at the same position, with the resulting configuration shown in the diagram; in agreement with our previous conclusions from the consideration of a simple junction, we see that barriers exist for the motion of holes from the p regions into the n region, and also for the motion of electrons from the n region into either of the p regions.

Figure 2.22b shows the double junction under operating biases; note that one junction is biased *forward*, the other is biased in the *reverse* direction. The n-type material common to both junctions is called the *base*,

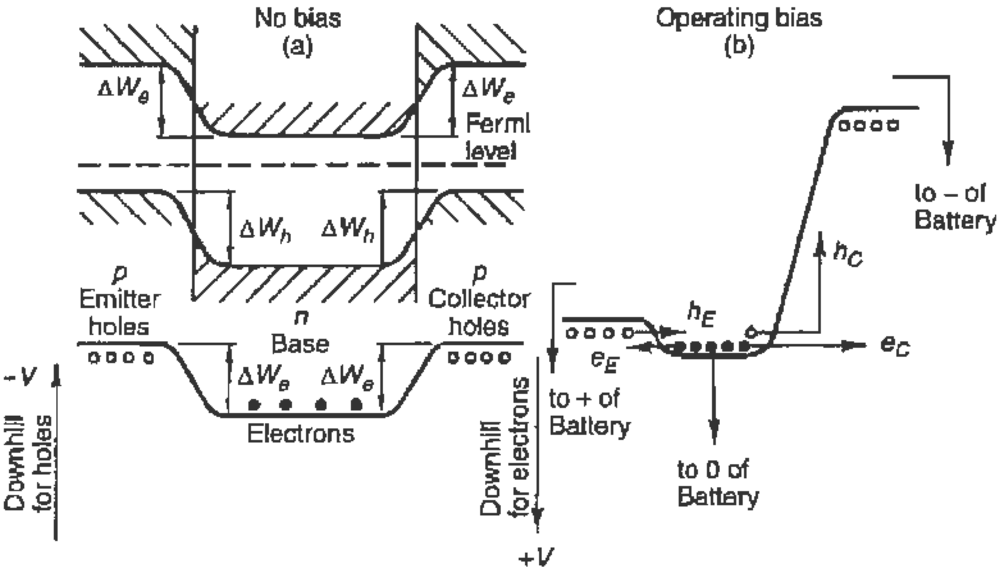


FIGURE 2.22 Structure of the energy bands for a p–n–p junction transistor: (a) with no bias applied, and (b) with operating biases. Note that the emitter is forward-biased, whereas the collector is reverse-biased.

while the p type of the forward-biased junction is the *emitter*; the p-type material of the reverse junction is the *collector*. A completely symmetric device consisting of n–p–n materials will perform similarly when the biases are reversed. From the energy diagram of Fig. 2.22b we can see that by varying the emitter junction bias we can control the injection of minority carriers into the base region; if the base region is made thin, it is possible for these holes to reach the collector junction, at which point they will immediately cross it, since it represents a gain in potential energy. If h_E is the minority carrier current injected into the base over a potential barrier $\Delta W_h(EB)$, the power required for injection is $P_{\text{in}} = h_E \Delta W_h(EB)$; similarly if h_C is the hole current into the collector down a potential drop $\Delta W_h(BC)$, the power gained is $P_{\text{out}} = h_C \Delta W_h(EC)$. Thus if $P_{\text{out}} > P_{\text{in}}$, the device is a power amplifier; since usually $\Delta W(CB) \gg \Delta W(EB)$, it suffices for $h_C \sim h_E$ to give power gain.

2.4.3. Measurements of the I–V curve of a pn Junction

A simple experiment that demonstrates the properties of a pn junction is discussed below. One simply measures the current as a function of (positive and negative) voltage across a diode. Additional properties can be demonstrated by varying the temperature of the diode, which changes the number of carriers in the conduction band. That is, the carriers (be they electrons or holes) will lead to a current density of the form $J_{e,h} = (J_{e,h})_0 \exp(eV_{\text{B}}/kT)$, where V_{B} is the bias voltage across the diode. The minority carriers will cancel this current exactly when there is no bias voltage applied, so the net current through an ideal diode has the form

$$I = I_0\left(e^{eV_{\text{B}}/kT} - 1\right). \tag{2.30}$$

A photograph of the experimental setup is shown in Fig. 2.23. A silicon pn junction diode is attached to one side of a copper plate with conductive epoxy. A power resistor is attached to the other side of the plate, to be used as a heat source. A thermocouple is also attached to record the temperature. A Keithley Model 617 Programmable Electrometer is used to vary the voltage across the diode, and to record the current. The result of a V–I scan, and the temperature as determined by the thermocouple are recorded using a Universal Laboratory Interface. (See Section 3.9.) Measurements

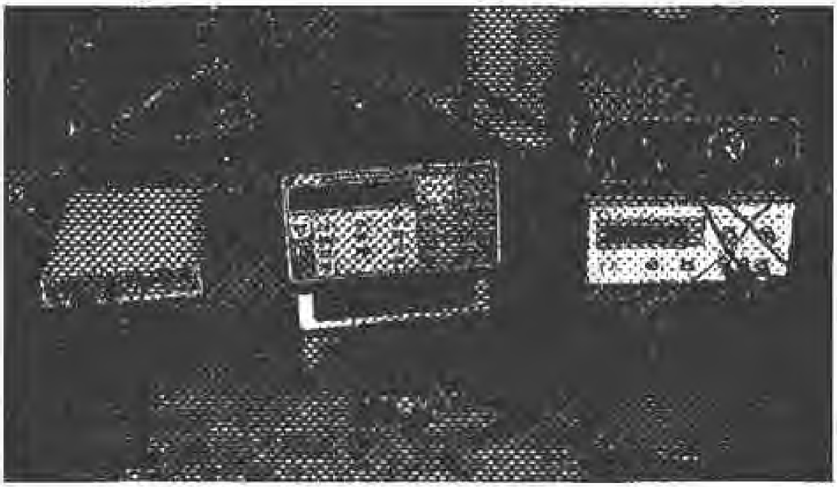


FIGURE 2.23 Photograph of the setup used to measure the properties of a diode.

are repeated for different currents through the power resistor, giving the setup time to come to thermal equilibrium.

To analyze the data we must appreciate that the diode does not obey the ideal diode equation (2.30) but operates in the recombination regime,[15] where

$$I = I_0(e^{eV_B/2kT} - 1) \simeq I_0 e^{eV_B/2kT}, \tag{2.31}$$

and the last approximation is justified because the term $\exp(eV_B/2kT) \gg 1$. Therefore, we present the V–I curves for $V_B > 0$, on a semi-log plot in Fig. 2.24a. From the fit we find the slopes

$$\begin{aligned} T = 24°\mathrm{C} = 297\,\mathrm{K} \qquad & e/2kT = 21.3\ \mathrm{V}^{-1} \\ 319\,\mathrm{K} \qquad & = 19.6\ \mathrm{V}^{-1} \\ 342\,\mathrm{K} \qquad & = 18.9\ \mathrm{V}^{-1}. \end{aligned}$$

Note the onset of saturation for biases $V_B \gtrsim 0.5\ V$ and also the different intercepts at $V = 0$.

We observe that the measured slopes do indeed scale with temperature as expected and if we average the three results we obtain

$$e/2k = (6.28 \pm 0.19) \times 10^3\ \mathrm{K/V}.$$

[15] G. W. Neudeck, *The p-n Junction Diode*, Addison-Wesley, Reading, MA, 1983.

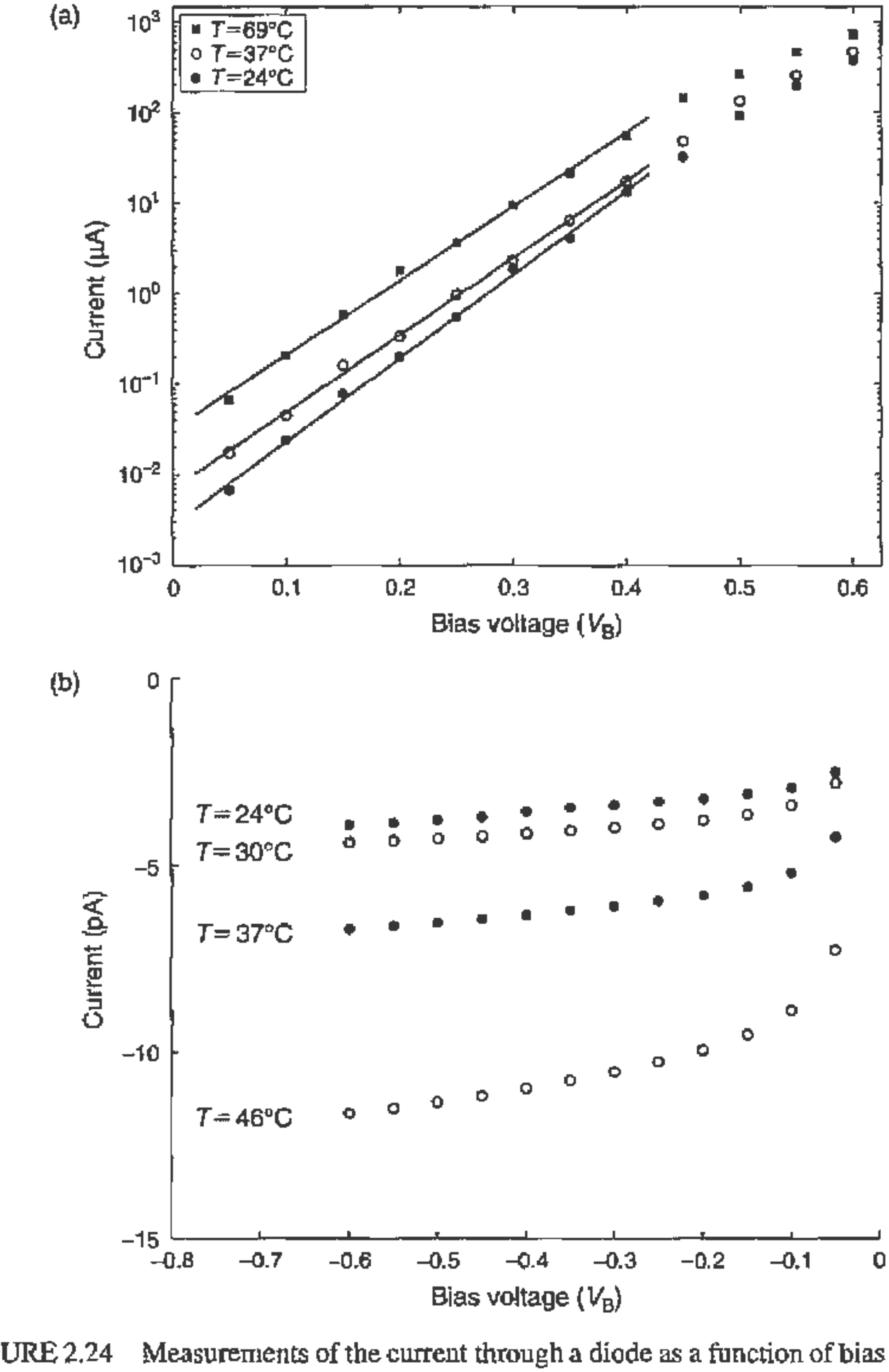


FIGURE 2.24 Measurements of the current through a diode as a function of bias voltage, for different temperatures. (a) is for positive bias, plotted on a semilogarithmic scale. Exponential fits are indicated. (b) is for negative bias voltage, plotted on a linear scale.

Thus, using the value of the Boltzman constant $k = 1.38 \times 10^{-23}$ J/K we find that

$$e = (1.73 \pm 0.05) \times 10^{-19}\ \mathrm{C}$$

in good agreement with the value of the electron charge.

The different intercepts are an indication of the variation of I_0 with temperature. (Of course at $V_B = 0$, $I = 0$ but this point cannot be reached on the logarithmic plot.) A better way of determining I_0 is by applying negative bias. From the negative bias data (Fig. 2.24b) we find that

$T = 297$ K	$I_0 = 3.9$ pA
303 K	$= 4.4$ pA
310 K	$= 6.7$ pA
319 K	$= 11.6$ pA.

The reverse current is proportional to the minority carrier density. As the temperature increases, the population density increases as

$$n \propto e^{-E_g/2kT},$$

where E_g is the energy gap between the valence and conduction bands. From the data we find that

$$E_g = 0.84\ \mathrm{eV}.$$

This is in reasonable agreement with the energy gap in silicon (1.1 eV at room temperature). Systematic error can come from a number of sources, including contact potential differences and the extent to which the negative bias data of Fig. 2.24b has reached its asymptotic value.

2.5. HIGH T_c SUPERCONDUCTORS

2.5.1. Introduction

In 1911 it was discovered that certain metals completely lose their electrical resistance when cooled to very low temperatures, typically less than 10 K. The loss of resistivity sets in sharply when the *critical temperature* T_c is crossed. This is analogous to a phase transition between different states of matter, as for instance from ice to water. The phase diagram for the

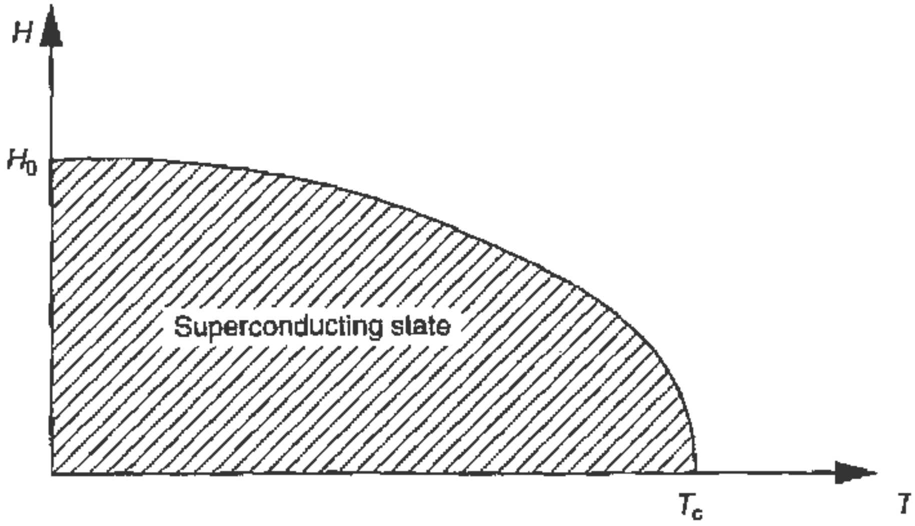


FIGURE 2.25 Typical phase diagram of the superconducting state in the T, H plane.

superconducting state depends on the temperature and the external magnetizing field H as shown in Fig. 2.25. Values of T_c and B_0 for some common metallic superconductors are given below

	T_c	H_0
Niobium (Nb)	9.46 K	0.1944 T
Lead (Pb)	7.81 K	0.0803 T
Mercury (Hg)	4.1 K	0.0411 T

By now many materials have been found to become superconducting at low temperature and such behavior can be explained by the BCS theory.[16] The basic mechanism is that at low temperature electrons bind in pairs with opposite spins. The net spin of the pair is zero so the pairs obey Bose statistics and can move through the lattice without scattering, namely without resistance. This gives rise to a supercurrent, which once started continues to circulate even after the external electromotive force (i.e., the applied voltage) is removed. In fact, all of the pairs occupy the ground state and can be described by a wave function that extends over macroscopic dimensions.

Superconductors not only exhibit zero resistance (when below T_c) but also have the property that no magnetic field can exist inside the superconductor. Inside an idealized "perfect conductor" the magnetic field cannot change, $dB/dt = 0$. This is because any change in the external field induces, by Faraday's law, surface currents that exactly cancel the effect of

[16] J. Bardeen, L. Cooper, and J. Schrieffer received the 1972 Nobel prize for their microscopic theory of superconductivity proposed in 1957.

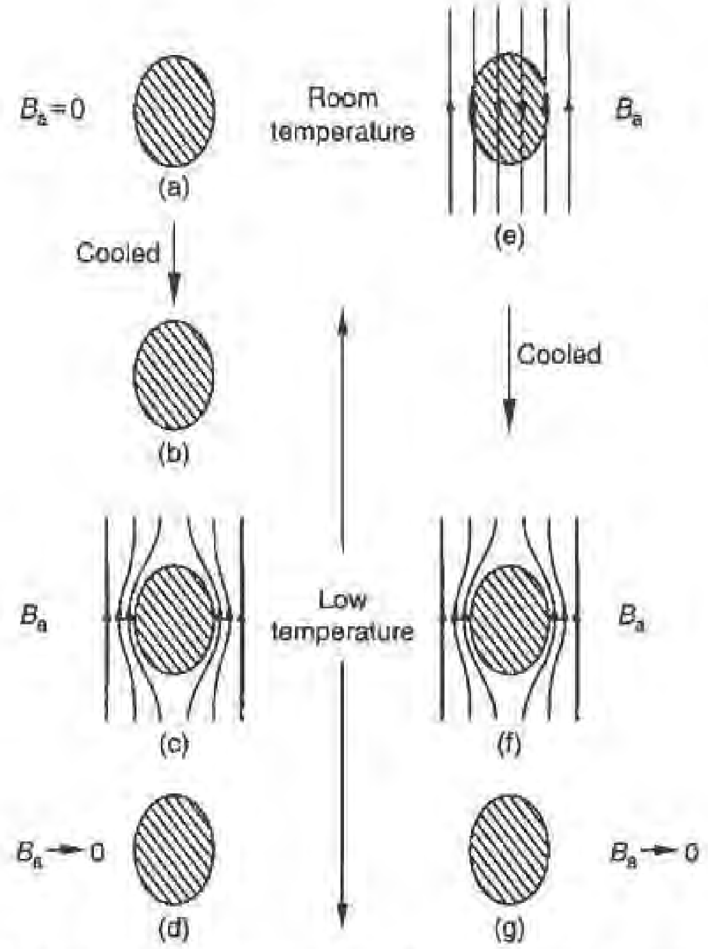


FIGURE 2.26 Behavior of a superconductor when placed in a magnetic field. (a–d) The field is switched on after the sample is cooled below T_c. (e,f) The field is applied before cooling the sample. In either case the flux is expelled from the superconductor and no field is trapped in its interior.

the external field inside the conductor. In a superconductor, however, $B = 0$ in the interior region, irrespective of whether the field is applied before or after the superconductor is cooled below T_c. This is shown in Fig. 2.26.

The exclusion of the magnetic field (flux) from the interior of a superconductor is called the Meissner effect and can be easily demonstrated by levitating a small permanent magnet above the surface of a superconductor. This is shown in Fig. 2.27 where the solid lines are the magnetic field lines of the permanent magnet. Since the superconductor must expel the flux from its interior, the induced surface currents produce the field shown by the dotted lines. They exactly cancel the external field *in the interior* of the superconductor. However, outside the superconductor there now exists a magnetic field opposite to that of the permanent magnet. Thus there is a force pushing the small magnet away from the superconductor. As the

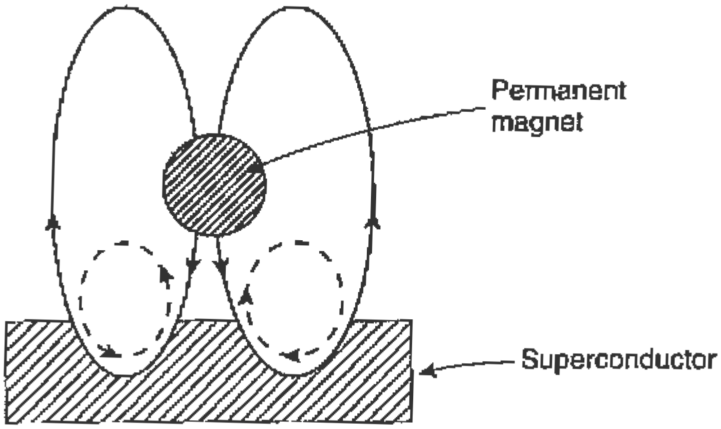


FIGURE 2.27 A permanent magnet is placed above a superconductor. The solid lines are the flux produced by the permanent magnet. The dotted lines are the flux produced by the induced surface currents and completely cancel the external flux in the interior of the superconductor. In the exterior they give rise to a lift force.

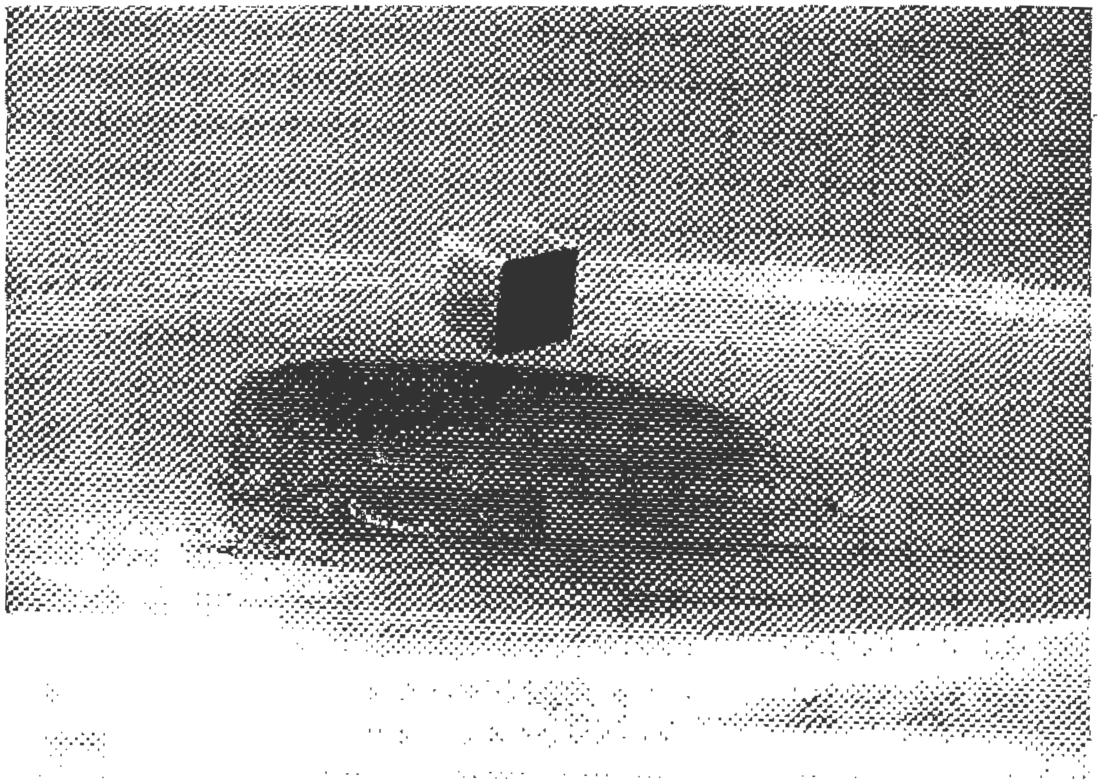


FIGURE 2.28 Levitation of a small permanent magnet above a YBCO pellet cooled by liquid nitrogen. Courtesy Colorado Superconductor, Inc.

distance increases, the magnetic lift force decreases; equilibrium is reached when the lift equals the gravitational force on the magnet. Figure 2.28 is an actual picture of levitation due to the Meissner effect using the high T_c superconductor discussed in the next section.

Superconductors are widely used for the construction of high-field magnets. They are also extensively used in some of the most sensitive scientific instruments; finally, they display fascinating quantum-mechanical effects in a macroscopic system.[17]

2.5.2. Observation of the Superconducting Transition in YBCO

In 1986 Bednorz and Müller reported superconductivity at temperatures in excess of 40 K in certain samples of La–Ba–Cu–O. It was soon discovered that the $YBa_2Cu_3O_7$ ceramic (YBCO) undergoes transition to the superconducting state above 90 K. Pellets of YBCO can be manufactured in the laboratory by mixing the chemicals in powder form and compressing them in a steel die using a hydraulic press (to approximately 15,000 psi). The pellets are then heated in a furnace to about 900°C in an oxygen atmosphere and allowed to cool. However, it is by far more convenient to order 1-in.-diameter disks of YBCO from a commercial supplier. A reliable source is Colorado Superconductor, P.O. Box 8223, Fort Collins, CO 80526.

To measure the resistivity of the sample, a four-point probe, as well as thermocouple leads, are attached to one side of the disk as shown in Fig. 2.29. The probes can be fastened using conductive epoxy.[18] The whole assembly is placed in a flat container that can be filled with liquid Nitrogen. Data can be taken as the sample cools or, as was done for the data presented here, by first cooling the sample for 2 min. and then removing it from the liquid N_2 bath. Temperature and resistivity are recorded as the sample warms through the superconducting transition.

The four connections (see Fig. 2.29) are spaced equidistantly, separated by a distance s, typically ~ 1 mm. A high-impedance source supplies a constant current to the outer terminals, 1 and 4, and the voltage across terminals 2 and 3 is measured. For a flat sample of thickness $t \ll s$, as in the present case, current rings emanate from the outer tips, so that the resistance between terminals 2 and 3 is

$$R = \int_{x_2}^{x_3} \rho \frac{dx}{2\pi x t} = \frac{\rho}{2\pi t} \ln x \bigg|_{s}^{2s} = \frac{\rho}{2\pi t} \ln 2.$$

[17] See R. P. Feynman, *The Feynman Lectures*, Vol. III, Lecture 21.

[18] The commercial pellets can be obtained with all the leads attached.

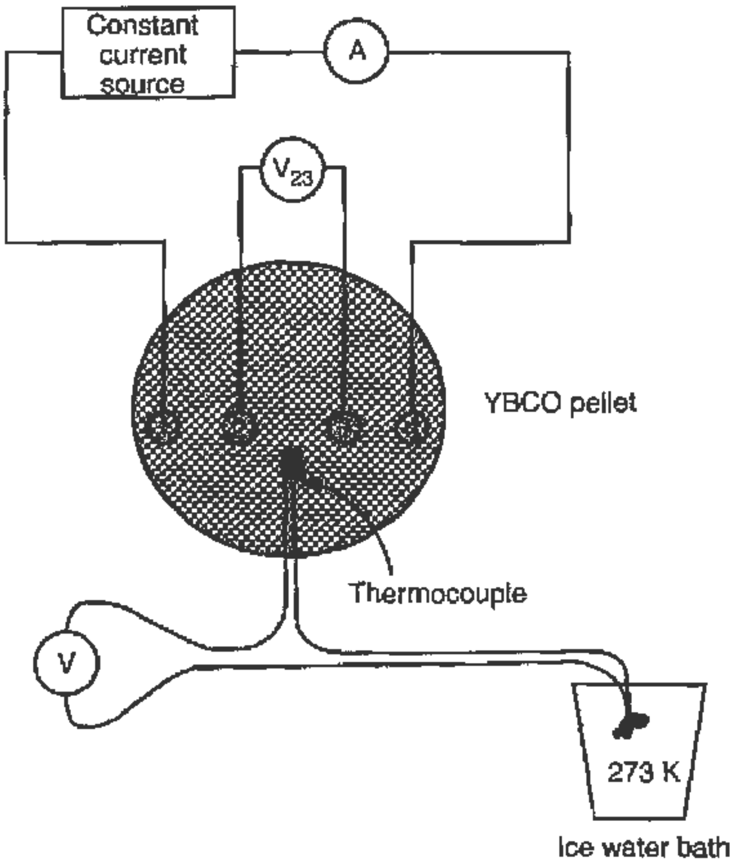


FIGURE 2.29 Schematic of the connections to the four-way probe and of the measuring apparatus.

Furthermore, due to the presence of two outer tips, $V_{23} = 2IR$, so that for a thin sample

$$\rho = \frac{\pi t}{\ln 2}\left(\frac{V}{I}\right). \tag{2.32}$$

Note that the probe spacing s does not enter Eq. (2.32).

In these measurements, the constant current source provided $I \approx$ 500 mA to terminals 1 and 4. Typically, in the normal conducting state $V_{23} \simeq 1$ mV, whereas below the transition, V_{23} is at the noise limit of the HP 34401A meter used for the measurement ($V_{23} \simeq 10\ \mu$V). Since the transition occurs rapidly it is important to use a computer to record the data. In the present case data were recorded every 0.33 s. The HP meter was connected to the computer through an RS232 serial port, the thermocouple voltage and source current through an ADC card.

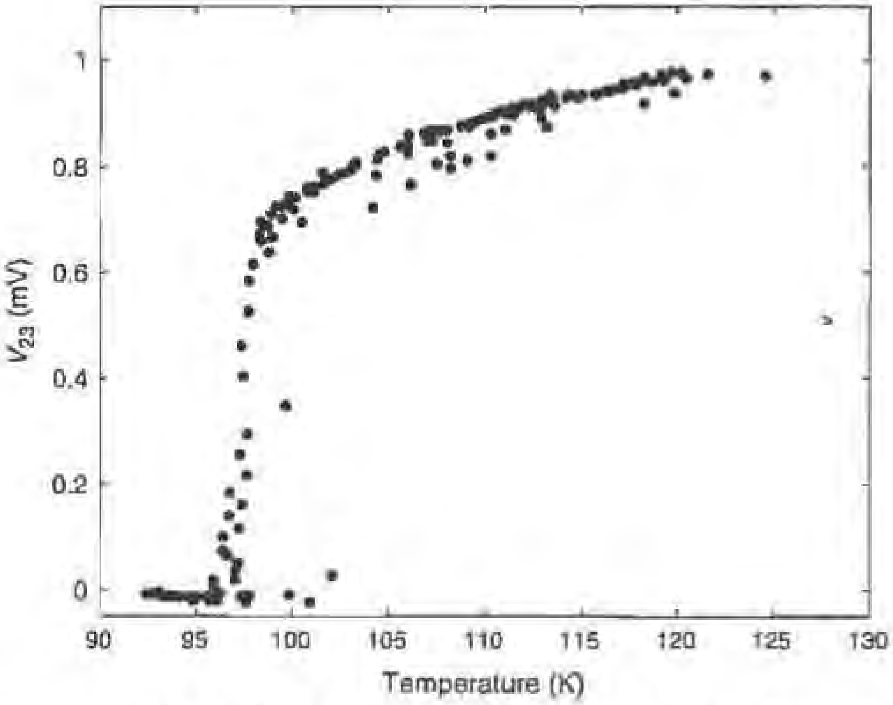


FIGURE 2.30 Plot of V_{23} vs T. Below the transition at $T = 98°\text{K}$ the voltage on the probe terminals is compatible with zero. The transition width $\Delta T < 1°\text{K}$.

Results obtained by a student are shown in Fig. 2.30, which is a plot of V_{23} vs T. It is clear that a phase transition occurs at $T = 98$ K. The width of the transition is $\Delta T < 1$ K. Note that the voltage V_{23} *below* T_c is too small to measure.

For $T > T_c$ the resistance across terminals 2 and 3 of the probe is of order

$$R_{23} = V_{23}/2I \simeq 7 \times 10^{-4}\ \text{V}/1\ \text{A} = 0.7\ \text{m}\Omega.$$

Then, from Eq. (2.32), using $t = 1$ mm we obtain

$$\rho = 3.15 \times 10^{-4}\ \Omega\text{-cm}.$$

This is two orders of magnitude higher than the resistivity of metals (see Table 2.1) but is to be expected for a ceramic material. The gradual increase of the resistivity with temperature for $T > T_c$ is also expected since the "normal" electrons scatter from the lattice thermal vibrations as discussed in Section 2.2.

2.6. REFERENCES

For the material covered on semiconductors, the reader may also consult the following texts:

W. C. Dunlap, Jr., *An Introduction to Semiconductors*, Wiley, New York, 1957. Brief but clear treatment.

R. A. Dunlap, *Experimental Physics: Modern Methods*, Oxford Univ. Press, New York, 1988. Detailed discussion of semiconductors, their physics, and device applications.

C. Kittel, *Introduction to Solid State Physics*, 7th ed., Wiley, New York, 1996. A more general treatment of the solid state.

W. Shockley, *Electrons and Holes*, Van Nostrand, New York, 1950. A thorough presentation of the subject.

The lecture on superconductivity in the Feynman lectures, Vol. III-21 is highly recommended. Also highly recommended is the text

A. C. Rose-Innes and E. H. Rhoderick, *Introduction to Superconductivity*, Pergamon, Elmsford, NY, 1969.

For a practical account including information on high T_c materials one can consult

D. Prochnow, *Superconductivity: Experimenting in New Technology*, Tab Books, Blue Ridge Summit, PA, 1989.

CHAPTER 3

Electronics and Data Acquisition

Up to this point, we have described measurements that require only rudimentary laboratory equipment. Before continuing, however, we will discuss a broader range of topics in electronics and data acquisition.

3.1. ELEMENTS OF CIRCUIT THEORY

Nearly every measurement made in a physics laboratory comes down to determining a voltage, so it is important to have at least a basic understanding of electronic circuits. It is not important to be able to design circuits, or even to completely understand a circuit given to you, but you do need to know enough to get some idea of how the measuring apparatus affects your result. This section introduces the basics of elementary, passive electronic circuits. You should be familiar with the concepts of electric voltage and current before you begin, but something on the level of an introductory physics course should be sufficient. It is helpful to have already learned something about resistors, capacitors, and inductors as well, but we will review them briefly.

3.1.1. Voltage, Resistance, and Current

Figure 3.1a shows a DC current loop. It is just a battery that provides the elcctromotive forcc V, which drives a current I through the resistor R. This is a cumbersome way to write things, however, so we will use the shorthand shown in Fig. 3.1b. All that ever matters is the *relative* voltage between two points, so we specify everything relative to the "common" or "ground." There is no need to connect the circuit loop with a line; it is understood that the current returns from the common point back to the terminals of the battery.

The concept of electric potential is based on the idea of electric potential energy, and energy is conserved. This means that the total change in electric potential going around the loop in Fig. 3.1a must be zero. In terms of Fig. 3.1b, the "voltage drop" across the resistor R must equal V. For ideal resistors, $V = IR$; that is, they obey Ohm's law. The SI unit of resistance is volts/amperes, also known as the ohm (Ω).

Electric current is just the flow of electric charge ($I \equiv dq/dt$, to be precise), and electric charge is conserved. This means that when there is a "junction" in a circuit, like that shown in Fig. 3.2, the sum of the currents flowing into the junction must equal the sum of the currents flowing out. In the case of Fig. 3.2, this rule just implies that $I_1 = I_2 + I_3$. It does not

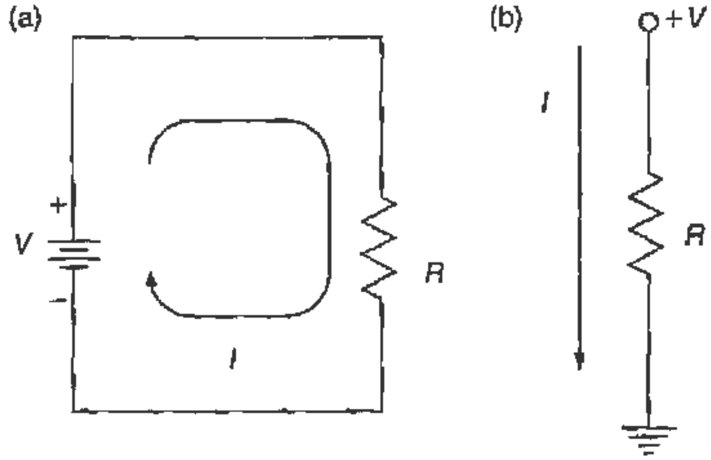


FIGURE 3.1 The simple current loop (a) showing the entire loop, and (b) in shorthand.

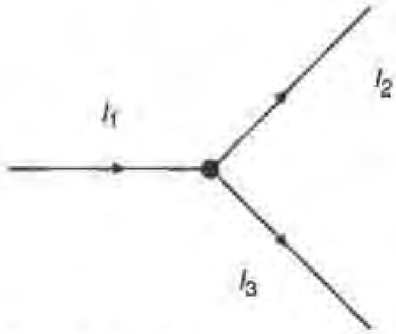


FIGURE 3.2 A simple three-wire circuit junction.

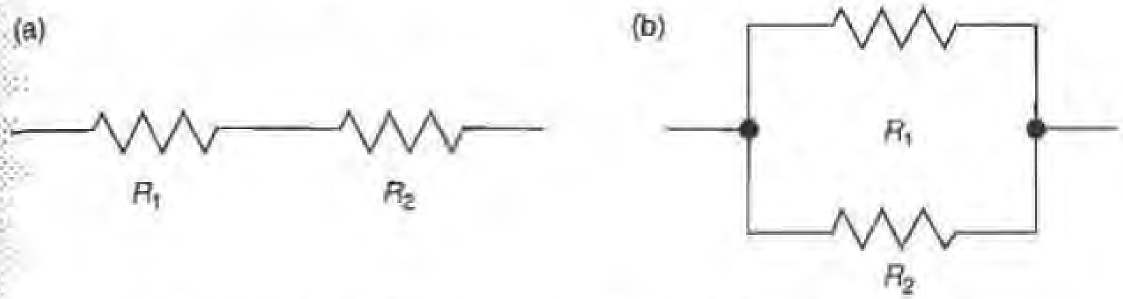


FIGURE 3.3 Resistors connected (a) in series and (b) in parallel.

matter whether you specify the current flowing in or out, so long as you are consistent with this rule. Remember that current can be negative as well as positive.

These rules and definitions allow us to determine the resistance when resistors are connected in series, as in Fig. 3.3a, or in parallel, as in Fig. 3.3b. In either case, the voltage drop across the pair must be IR, where I is the current flowing through them. For two resistors R_1 and R_2 connected in series, the current is the same through both, so the voltage drops across them are IR_1 and IR_2, respectively. Since the voltage drop across the pair must equal the sum of the voltage drops, then $IR = IR_1 + IR_2$, or

$$R = R_1 + R_2 \qquad \text{Resistors in series.}$$

If R_1 and R_2 are connected in parallel, then the voltage drops across each are the same, but the current through them is different. Therefore $IR = I_1R_1 = I_2R_2$. Since $I = I_1 + I_2$, we have

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} \qquad \text{Resistors in parallel.}$$

Remember that whenever a resistor is present in a circuit, it may as well be some combination of resistors that give the right value of resistance.

A very simple, and very useful, configuration of resistors is shown in Fig. 3.4. This is called a "voltage divider" because of the simple relationship between the voltages labeled V_{out} and V_{in}. Clearly $V_{\text{in}} = I(R_1 + R_2)$ and $V_{\text{out}} = I(R_2)$, where I is the current through the resistor string. Therefore

$$V_{\text{out}} = V_{\text{in}} \frac{R_2}{R_1 + R_2}. \tag{3.1}$$

That is, this simple circuit divides the "input" voltage into a fraction determined by the relative resistor values. We will see lots of examples of this sort of thing in the laboratory.

Do not get confused by the way circuits are drawn. It does not matter which directions lines go in. Just remember that a line means that all points along it are at the same potential. For example, it is common to draw a voltage divider as shown in Fig. 3.5. This way of looking at it is in fact an easier way to think about an "input" voltage and an "output" voltage.

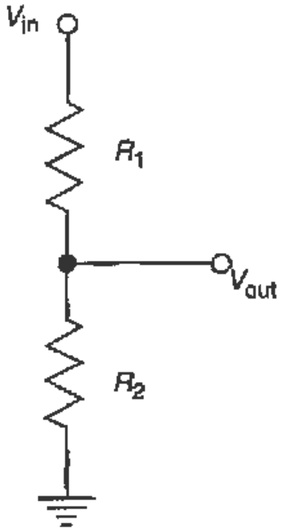


FIGURE 3.4 The basic voltage divider.

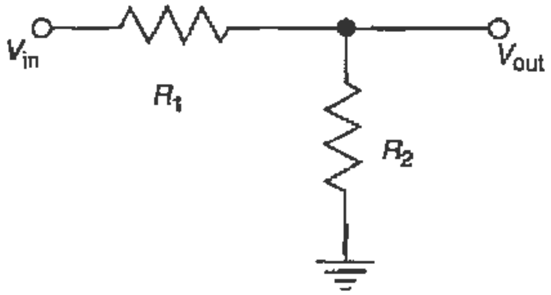


FIGURE 3.5 An alternate way to draw a voltage divider.

3.1.2. Capacitors and AC Circuits

A capacitor stores charge, but does not allow the charge carriers (i.e., electrons) to pass through it. It is simplest to visualize a capacitor as a pair of conducting plates, parallel to each other and separated only by a small amount. If a capacitor has a potential difference V across its leads and has stored a charge q on either side, then we define the *capacitance* $C \equiv q/V$. It is easy to show that for a parallel plate capacitor, C is a constant value independent of the voltage. In general, it is possible, but not easy, to calculate C from the geometry of the conducting surfaces. The SI unit of capacitance is Coulombs/Volts, also known as the Farad (F). As it turns out, one Farad is an enormous capacitance, and laboratory capacitors typically have values between a few microfarads (μF) down to a few hundred micromicrofarads ($\mu\mu$F).[1]

It is pretty easy to figure out what the effective capacitance is if capacitors are connected in series and in parallel, just using the above definitions and the rule about the total voltage drop. The answers are

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} \qquad \text{Capacitors in series}$$

and

$$C = C_1 + C_2 \qquad \text{Capacitors in parallel,}$$

that is, just the opposite from resistors.

Now let's think about what a capacitor does in a circuit. Let's take the resistor R_2 in the voltage divider of Fig. 3.4 and replace it with a capacitor C. This is pictured in Fig. 3.6. The capacitor does not allow any charge carriers to pass through it, so the current $I = 0$. Therefore the voltage drop across the resistor R is zero, and V_{out}, the voltage across the capacitor C, just equals V_{in}. You may wonder, *what good is this?* We might have just as well connected the output terminal to the input! To appreciate the importance of capacitors in circuits, we must consider voltages that change with time.

If the voltage changes with time, we refer to the system as an AC circuit. If the voltage is constant, we call it a DC circuit. Now go back to the voltage divider with a capacitor, pictured in Fig. 3.6, and let the input

[1] 1 $\mu\mu$F = 1 pF (picofarad).

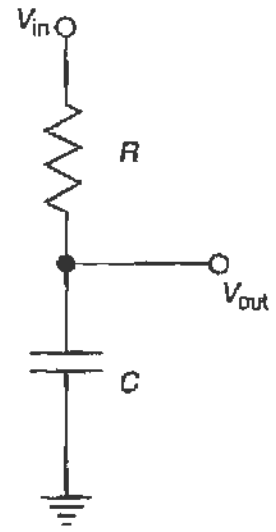


FIGURE 3.6 A voltage divider with a capacitor in it.

voltage change with time in a very simple way. That is, take

$$V_{in}(t) = 0 \qquad \text{for } t \leq 0 \tag{3.2}$$

$$= V \qquad \text{for } t > 0 \tag{3.3}$$

and assume that there is no charge q on the capacitor at $t = 0$. Then for $t > 0$, the charge $q(t)$ produces a voltage drop $V_{out}(t) = q(t)/C$ across the capacitor. The current $I(t) = dq/dt$ through the divider string also gives a voltage drop IR across the resistor, and the sum of the two voltage drops must equal V. In other words

$$V = V_{out} + IR = V_{out} + R\frac{dq}{dt} = V_{out} + RC\frac{dV_{out}}{dt} \tag{3.4}$$

and $V_{out}(0) = 0$. This differential equation has a simple solution. It is

$$V_{out}(t) = V[1 - e^{-t/RC}]. \tag{3.5}$$

Now it should be clear what is going on. As soon as the input voltage is switched on, current flows through the resistor and the charge carriers pile up on the input side of the capacitor. There is induced charge on the output side of the capacitor, and that is what completes the circuit to ground. However, as the capacitor charges up, it gets harder and harder to put more charge on it, and as $t \to \infty$, the current does not flow anymore and $V_{out} \to V$. This is just the DC case, where this circuit is not interesting anymore.

The value RC is called the "capacitive time constant," and it is the only time scale we have in this circuit. That is, statements like "$t \to 0$" and "$t \to \infty$" actually mean "$t \ll RC$" and "$t \gg RC$." The behavior of the circuit will always depend on the time as measured in units of RC. So now we see what is interesting about capacitors. They are sensitive to currents that are changing with time in a way that is quite different from resistors. That is a very useful property that we will study some more, and use in lots of experiments.

The time dependence of any function can always be expressed in terms of sine and cosine functions using a Fourier transform. It is therefore common to work with sinusoidally varying functions for voltage and so forth, just realizing that we can add them up with the right coefficients to get whatever time dependence we want in the end. It is very convenient to use the complex number notation

$$V(t) = V_0 e^{i\omega t} \tag{3.6}$$

for time-varying (i.e., AC) voltages, where it is understood that the voltage we measure in the laboratory is just the real part of this function. The angular frequency $\omega = 2\pi\nu$, where ν is the frequency, that is, the number of oscillations per second. This expression for $V(t)$ is easy to differentiate and integrate when solving equations. It is also a neat way of keeping track of all the phase changes signals undergo when they pass through capacitors and other "reactive" components. You will see and appreciate this better as we go along.

Now is a convenient time to define *impedance*. This is just a generalization of resistance for AC circuits. Impedance, usually denoted by Z, is a (usually) complex quantity and (usually) a function of the angular frequency ω. It is defined as the ratio of voltage drop across a component to the current through it, and just as for resistance, the SI unit is the ohm. For "linear" components (of which resistors and capacitors are common examples), the impedance is not a function of the amplitude of the voltage or current signals. Given this definition of impedance, the rules for the equivalent impedance are the same as those for resistance. That is, for components in series, add the impedances, while if they are in parallel, add their reciprocals.

The impedance of a resistor is trivial. It is just the resistance R. In this case, the voltage drop across the resistor is in phase with the current through it since $Z = R$ is a purely real quantity. The impedance is also independent of frequency in this case. For a capacitor, the voltage drop

$V = V_0 e^{i\omega t} = q/C$ and the current $I = dq/dt = i\omega C \times V_0 e^{i\omega t}$. Therefore, the impedance is

$$Z(\omega) = \frac{V(\omega, t)}{I(\omega, t)} = \frac{1}{i\omega C}. \tag{3.7}$$

Now the behavior of capacitors is clear. At frequencies low compared to $1/RC$, i.e., the "DC limit," the impedance of the capacitor goes to infinity. (Here, the value of R is the equivalent resistance in series with the capacitor.) It does not allow current to pass through it. However, as the frequency gets much larger than $1/RC$, the impedance goes to 0 and the capacitor acts like a short, since current passes through it as if it were not there. You can learn a lot about the behavior of capacitors in circuits just by keeping these limits in mind.

We can easily generalize our concept of the voltage divider to include AC circuits and reactive (i.e., frequency dependent) components like capacitors. We will learn about another reactive component, the inductor, shortly. The generalized voltage divider is shown in Fig. 3.7. In this case, we have

$$V_{\text{out}}(\omega, t) = V_{\text{in}}(\omega, t)\frac{Z_2}{Z_1 + Z_2} = V_{\text{in}}(\omega, t) g e^{i\phi}, \tag{3.8}$$

where we have expressed the impedance ratio $Z_1/(Z_1 + Z_2)$, a complex number, in terms of two real numbers g and ϕ. We refer to $g = |V_{\text{out}}|/|V_{\text{in}}|$

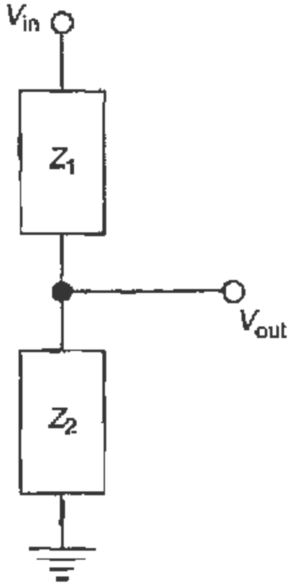


FIGURE 3.7 The generalized voltage divider.

as the "gain" of the circuit, and ϕ is the phase shift of the output signal relative to the input signal. For the simple resistive voltage divider shown in Figs. 3.4 and 3.5, we have $g = R_1/(R_1 + R_2)$ and $\phi = 0$. That is, the output signal is in phase with the input signal, and the amplitude is just reduced by the relative resistor values. This holds at all frequencies, including DC.

The relative phase is an important quantity, so let's take a moment to look at it a little more physically. If we write $V_{\text{in}} = V_0 e^{i\omega t}$, then according to Eq. (3.8) we can write $V_{\text{out}} = gV_0 e^{i\omega t + \phi}$. Since the measured voltage is just the real part of these complex expressions, we have

$$V_{\text{in}} = V_0 \cos(\omega t)$$
$$V_{\text{out}} = gV_0 \cos(\omega t + \phi)$$

These functions are plotted together in Fig. 3.8. The output voltage crests at a time different than the input voltage, and this time is proportional to the phase. To be exact, relative to the time at which V_{in} is a maximum,

$$\text{Time of maximum } V_{\text{out}} = -\frac{\phi}{2\pi} \times T = -\frac{\phi}{\omega},$$

where $T = 2\pi/\omega$ is the period of the driving voltage.

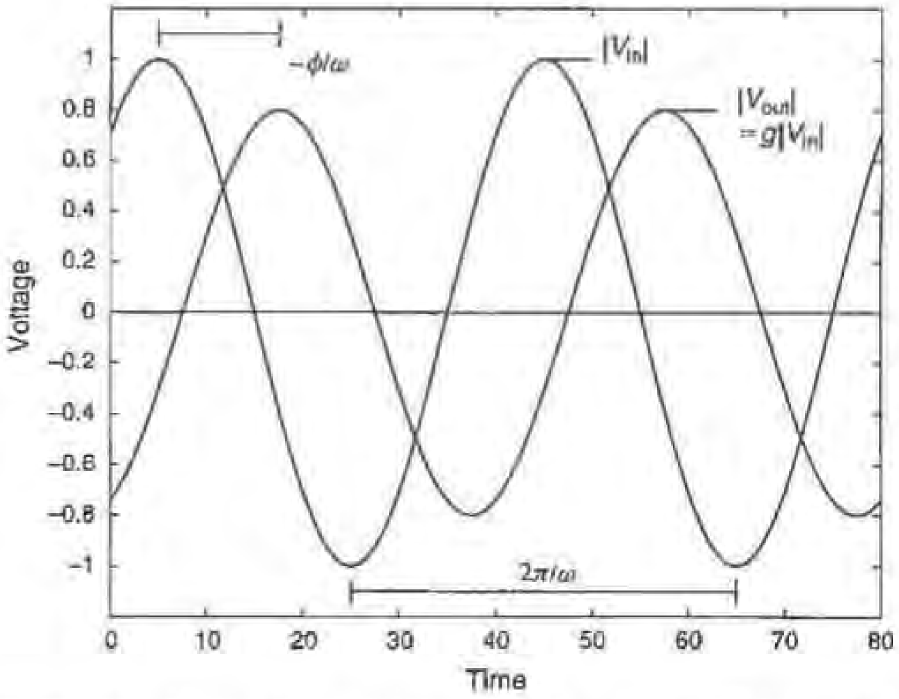


FIGURE 3.8 Input and output voltages for the generalized voltage divider.

Now consider the voltage divider in Fig. 3.6. Using Eq. (3.8) we find

$$V_{\text{out}} = V_{\text{in}} \frac{\dfrac{1}{i\omega C}}{R + \dfrac{1}{i\omega C}} = V_{\text{in}} \frac{1}{1 + i\omega RC}.$$

The gain g of this voltage divider is just $(1 + \omega^2 R^2 C^2)^{-1/2}$ and you can see that for $\omega = 0$ (i.e., DC operation) the gain is unity. For very large frequencies, though, the gain goes to 0. The gain changes from unity to 0 for frequencies in the neighborhood of $1/RC$. We have said all this before, but in a less general language.

However, our new language tells us something new and important about V_{out}, namely the phase relative to V_{in}. Any complex number z can be written as

$$z = |z|e^{i\phi} \quad \text{and} \quad z^* = |z|e^{-i\phi}, \tag{3.9}$$

where

$$\phi = \tan^{-1}\left[\frac{\text{Im}(z)}{\text{Re}(z)}\right] \tag{3.10}$$

is called the "phase" of z. Therefore, we find that

$$\frac{1}{1 + i\omega RC} = \frac{1 - i\omega RC}{1 + \omega^2 R^2 C^2} = \frac{1}{(1 + \omega^2 R^2 C^2)^{1/2}} e^{i\phi}.$$

In other words, the output voltage is phase shifted relative to the input voltage by an amount $\phi = -\tan^{-1}(\omega RC)$. For $\omega = 0$ there is no phase shift, as you should expect, but at very high frequencies the phase is shifted by $-90°$.

3.1.3. Inductors

Just as a capacitor stores energy in an electric field, an inductor stores energy in a magnetic field. An inductor is essentially a wire wound into the shape of a solenoid. The symbol for an inductor is . The key is in the magnetic field that is set up inside the coil, and what happens when the current *changes*. So, just as with a capacitor, inductors are important when the voltage and current change with time, and the response depends on the frequency.

The inductance L of a circuit element is defined to be

$$L = \frac{N\Phi}{I},$$

where N is the number of turns in the solenoid and Φ is the magnetic flux in the solenoid generated by the current I. The SI unit of inductance is the Tesla $\cdot$ m^2/Ampere, or the Henry (H).

Now if the current I through the inductor coil is changing, then the magnetic flux is changing and this sets up a voltage in the coil that opposes the change in the current. The magnitude of this voltage drop is

$$V = \frac{d(N\Phi)}{dt} = L\frac{dI}{dt}.$$

If we write $V = IZ$, where Z is the impedance of the inductor, and $I = I_0 e^{i\omega t}$, then $V = i\omega LI$ or

$$Z = i\omega L. \tag{3.11}$$

We can use this impedance to calculate, for example, V_{out} for the generalized voltage divider of Fig. 3.7 if one or more of the components is an inductor.

You can now see that the inductor is, to a large extent, the opposite of a capacitor. The inductor behaves as a short (that is, just the wire it is) at low frequencies, whereas a capacitor is open in the DC limit. On the other hand, an inductor behaves as if the wire were cut (an open circuit) at high frequencies, but the capacitor is a short in this limit.

One particularly interesting combination is the series LCR circuit, combining one of each in series. The impedance of such a string displays the phenomenon of "resonance." That is, in complete analogy with mechanical resonance, the voltage drop across one of the elements is a maximum for a certain value of ω. Also, as the frequency passes through this value, the relative phase of the output voltages passes through 90°. If the resistance R is very small, then the output voltage can be enormous, in principle.

3.1.4. Diodes and Transistors

Resistors, capacitors, and inductors are "linear" devices. That is, we write $V = IZ$, where Z is some (complex) number, which may be a function of frequency. The point is, though, that if you increase V by some factor, then you increase I by the same factor. Diodes and transistors are examples of "nonlinear" devices. Instead of talking about some impedance Z, we instead consider the relationship between V and I as some nonlinear function. What is more, a transistor is an "active" device, unlike resistors, capacitors, inductors, and diodes, which are "passive." That is, a transistor takes in power from some voltage or current source, and gives an output that combines that input power with the signal input to get a response. It used to be that many of these functions were possible with vacuum tubes of various kinds. These have been almost completely replaced by solid-state devices based on semiconductors. The physics of semiconductors and semiconductor devices was discussed in Sections 2.1 and 2.4.

The symbol for a diode is ▶| where the arrow shows the nominal direction of current flow. An ideal diode conducts in one direction only. That is, its V–I curve would give zero current I for $V < 0$ and infinite I for $V > 0$. (Of course, in practice, the current I is limited by some resistor in series with the diode.) This is shown in Fig. 3.9a. A real diode, however, has a more complicated curve, as shown in Fig. 3.9b. The current I changes approximately exponentially with V, and becomes very large for voltages above some forward voltage drop V_F. For most cases, a good approximation is that the current is zero for $V < V_F$ and unlimited for $V > V_F$. Typical values of V_F are between 0.5 and 0.8 V.

Diodes are *pn* junctions. These are the simplest solid-state devices, made of a semiconductor, usually silicon. The electrons in a semiconductor fill an energy band and normally cannot move through the bulk material, so the semiconductor is really an insulator. If electrons make it into the next

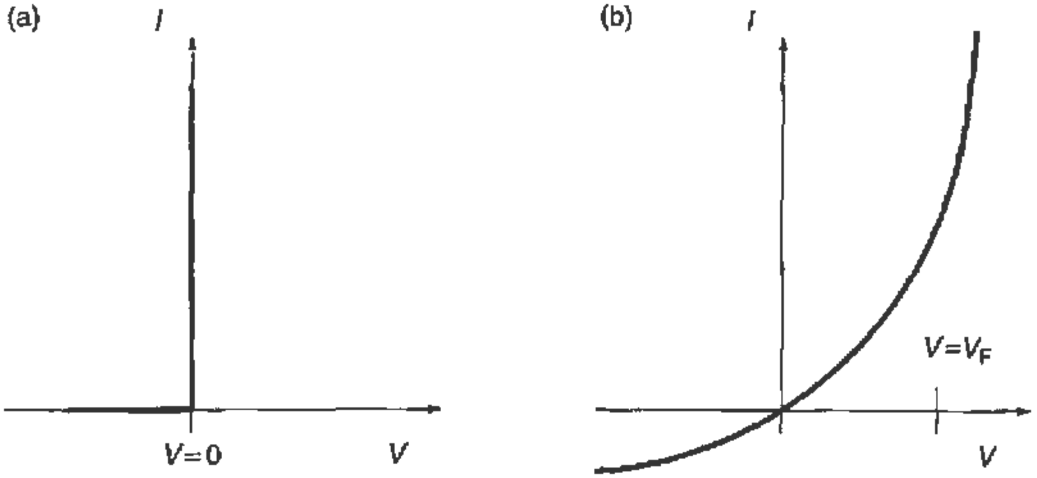


FIGURE 3.9 Current I versus voltage V for (a) the ideal diode and (b) a real diode.

energy band, which is normally empty, then they can conduct electricity. This can happen if, for example, electrons are thermally excited across the energy gap between the bands. For silicon, the band gap is 1.1 eV, but the mean thermal energy of electrons at room temperature is $\sim kT = 1/40$ eV. Therefore, silicon is essentially an insulator under normal conditions, and not particularly useful.

That is where the p and n come in. By adding a small amount (around 10 parts per million) of specific impurities, lots of current carriers can be added to the material. These impurities (called dopants) can precisely control how current is carried in the semiconductor. Some dopants, like arsenic, give electrons as carriers, and the doped semiconductor is called n-type, since the carriers are negative. Other dopants, like boron, bind up extra electrons, and current is carried by "holes" created in the otherwise filled band. These holes act like positive charge carriers, so we call the semiconductor p-type. In either case, the conductivity increases by a factor of $\sim$1000 at room temperature, and this makes some nifty things possible.

So now back to the diode, or pn junction. This is a piece of silicon, doped p-type on one side and n-type on the other. Electrons can only flow from p to n. That is, a current is carried only in one direction. A detailed analysis gives the I–V curve shown in Fig. 3.9b. See Dunlap (1988; full citing in Section 3.10) for more details. If you put voltage across the diode in the direction opposite to the direction of possible current flow, that is called a "reverse bias." A small "leakage" current flows as shown in Fig. 3.9b. If you put too much of a reverse bias on the diode, i.e., $V < -V_R^{\text{Max}}$, it will break down and start to conduct. Typical values of V_R^{Max} are 100 V or less.

Transistors are considerably more complicated than diodes,[2] and we will only scratch the surface here. The following summary closely follows the introduction to transistors in *The Art of Electronics* (full citing in Section 3.10). For details on the underlying theory, see Dunlap (1988). A transistor has three terminals, called the collector, base, and emitter. There are two main types of transistors, namely npn and pnp, and their symbols are shown in Fig. 3.10. The names are based on the dopants used in the semiconductor materials. The properties of a transistor may be summarized in

[2] The invention of the transistor was worth a Nobel Prize in Physics in 1956.

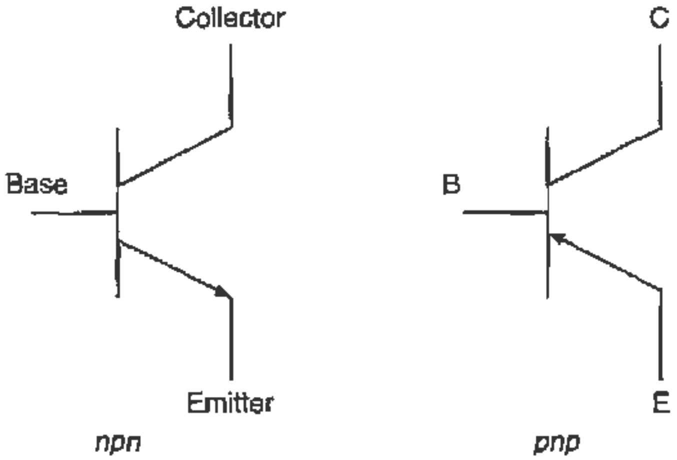


FIGURE 3.10 Symbols for *npn* and *pnp* transistors.

the following simple rules for *npn* transistors. (For *pnp* transistors, just reverse all the polarities.)

1. The collector must be more positive than the emitter.
2. The base–emitter and base–collector circuits behave like diodes. Normally the base–emitter diode is conducting and the base–collector diode is reverse-biased.
3. Any given transistor has maximum values of I_C, I_B, and V_{CE} that cannot be exceeded without ruining the transistor. If you are using a transistor in the design of some circuit, check the specifications to see what these limiting values are.
4. When rules 1–3 are obeyed, I_C is roughly proportional to I_B and can be written as $I_C = h_{FE}I_B$. The parameter h_{FE}, also called β, is typically around 100, but it varies a lot among a sample of nominally identical transistors.

Obviously, rule 4 is what gives a transistor its punch. It means that a transistor can "amplify" some input signal. It can also do a lot of other things, and we will see them in action later on.

3.1.5. Frequency Filters

Simple combinations of passive elements can be used to remove "noise" from a voltage signal. If the noise that is bothering you is in some specific range of frequencies, and you can make your measurement in some other range, then a *frequency filter* can do a lot for you. Frequency filters are usually simple circuits (or perhaps their mechanical analogs) that allow only a specific frequency range to pass from the input to the output. You then make your measurement with the output. Of course, you need to be careful of any noise introduced by the filter itself. The circuit shown in Fig. 3.6 is a "low-pass" filter. It exploits the frequency dependence of the capacitor impedance $Z_C = 1/i\omega C$ to short frequencies much larger than $1/RC$ to ground, and to allow much smaller frequencies to pass. As we showed earlier, the ratio of the output to input voltage as a function of frequency $\nu = \omega/2\pi$ is $(1+\omega^2R^2C^2)^{-1/2}$. You can also use inductors in these simple circuits. Remember that whereas a capacitor is open at low frequencies and a short at high frequencies, an inductor behaves just the opposite. Figure 3.11 shows all permutations of resistors, capacitors, and inductors, and whether they are high- or low-pass filters.

Suppose you only want to deal with frequencies in a specific range. Then, you want a "bandpass" filter, which cuts off at both low and high frequencies, but lets some intermediate bandwidth pass through. Consider the circuit shown in Fig. 3.12. The output voltage tap is connected to ground

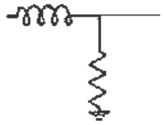
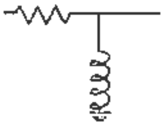
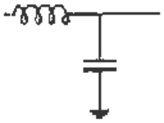
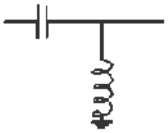
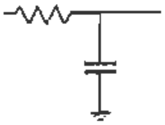
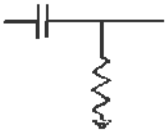
Circuit	Type	Circuit	Type
	Low pass		High pass
	Low pass		High pass
	Low pass		High pass

FIGURE 3.11 Simple passive frequency filters.

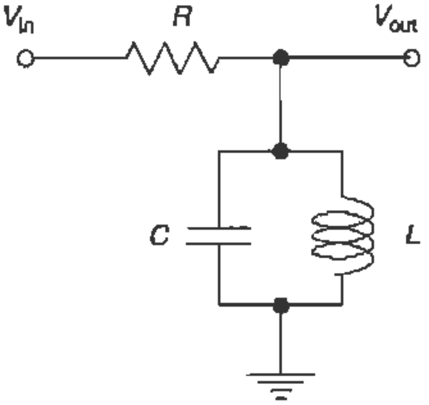


FIGURE 3.12 A simple bandpass filter.

through *either* a capacitor or an inductor. Therefore, the output will be zero at both low and high frequencies. Analyzing this filter circuit is simple

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{Z_{LC}}{Z_R + Z_{LC}},$$

where $Z_R = R$ and $Z_{LC} = \left(Z_L^{-1} + Z_C^{-1}\right)^{-1}$ with $Z_L = 1/i\omega L$ and $Z_C = i\omega C$. (Note that L and C are connected in parallel.) The result is

$$g = \left|\frac{V_{\text{out}}}{V_{\text{in}}}\right| = \frac{1}{\left[1 + \frac{R^2}{\omega^2 L^2}(1 - \omega^2 LC)^2\right]^{1/2}}$$

and as advertised, $g \to 0$ for both $\omega \ll R/L$ and for $\omega \gg 1/RC$. However, frequencies near $\nu = \omega/2\pi = 1/(2\pi\sqrt{LC})$ are passed through with little attenuation. At $\omega = 1/\sqrt{LC}$, $g = 1$ and there is no attenuation at all. Can you see how to build a "notch" filter, or "band reject" filter, that allows all frequencies to pass *except* those in the neighborhood of $\omega = 1/\sqrt{LC}$?

3.2. BASIC ELECTRONIC EQUIPMENT

3.2.1. Wire and Cable

Connections between components are made with wires. We tend to neglect the importance of choosing the right wire for the job, but in some cases it can make a big difference. The simplest wire is just a strand of some conductor, most often a metal such as copper or aluminum. Usually the wire is coated with an insulator so that it will not short out to its surroundings, or to another part of the wire itself. If the wire is supposed to carry some small signal, then it will likely need to be "shielded," that is, covered with another conductor (outside the insulator) so that the external environment does not add noise somehow. One popular type of shielded wire is the "coaxial cable," which is also used to propagate "pulses."

Do not forget about Ohm's law when choosing the proper wire. That is, the voltage drop across a section of wire is still $V = IR$, and you want this voltage drop to be small compared to the "real" voltages involved. The resistance $R = \rho \times L/A$, where L is the length of the wire, A is its cross-sectional area, and ρ is the resistivity of the metal. Therefore, to get the smallest possible R, you keep the length L as short as practical, get a wire with the largest practical A,[3] and choose a conductor with small resistivity. Copper is the usual choice because it has low resistivity ($\rho = 1.69 \times 10^{-8}$ Ω-cm) and is easy to form into wire of various thicknesses and shapes. Other common choices are aluminum ($\rho = 2.75 \times 10^{-8}$ Ω-cm), which can be significantly cheaper in large quantities, or silver ($\rho = 1.62 \times 10^{-8}$ Ω-cm), which is a slightly better conductor, although not usually worth the increased expense.

The resistivity increases with temperature, and this can lead to a particularly insidious failure if the wire must carry a large current. The power dissipated in the wire is $P = I^2R$, and this tends to heat it up. If there is not enough cooling by convection or other means, then R will increase and the wire will get hotter and hotter until it does serious damage. This is most common in wires used to wind magnets, but can show up in other high-power applications. A common solution is to use very-low-gage (i.e., very thick) wire which has a hollow channel in the middle through which water flows. The water acts as a coolant to keep the wire from getting too hot.

A coaxial cable is a shielded wire. The name comes from the fact that the wire sits inside an insulator, another conductor, and another insulator, all in circular cross section sharing the same axis. A cutaway view is shown in Fig. 3.13. Coaxial cable is used in place of simple wire when the signals are very small and are likely to be obscured by some sort of electronic noise in the room. The outside conductor (called the "shield") makes it difficult for external electromagnetic fields to penetrate to the wire, and minimizes the noise. This outside conductor is usually connected to ground.

A second, and very important, use of coaxial cable is for "pulse transmission." The wire and shield, separated by the dielectric insulator, act as a waveguide and allow short pulses of current to be transmitted with little distortion from dispersion. Short pulses can be very common in the laboratory, in such applications as digital signal transmission and in radiation detectors. You must be aware of the "characteristic impedance" of the cable when you use it in this way.

Coaxial cable has a characteristic impedance because it transmits the signal as a train of electric and magnetic fluctuations, and the cable itself has characteristic capacitance and inductance. The capacitance and inductance of a cylindrical geometry like this are typically solved in elementary physics

[3] Wire diameter is usually specified by the "gage number." The smaller the wire gage, the thicker the wire, and the larger the cross-sectional area.

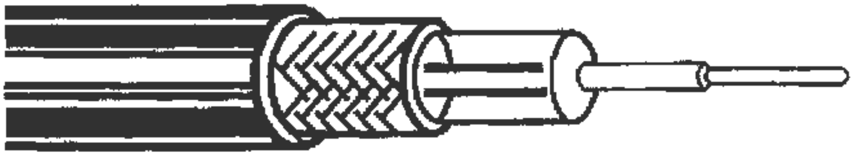


FIGURE 3.13 Cutaway view of coaxial cable.

texts on electricity and magnetism. The solutions are

$$C = \frac{2\pi\epsilon}{\ln(b/a)} \times \ell \qquad \text{and} \qquad L = \frac{\mu}{2\pi} \ln\left(\frac{b}{a}\right) \times \ell,$$

where a and b are the radii of the wire and shield respectively, ϵ and μ are the permittivity and permeability of the dielectric, and ℓ is the length of the cable. It is very interesting to derive and solve the equations that determine pulse propagation in a coaxial cable, but we will not do that here. One thing you learn, however, is that the impedance seen by the pulse (which is dominated by high frequencies) is very nearly real and independent of frequency, and equal to

$$Z_c = \sqrt{\frac{L}{C}} = \frac{1}{2\pi}\sqrt{\frac{\mu}{\epsilon}} \ln\left(\frac{b}{a}\right). \tag{3.12}$$

This "characteristic impedance" is always in a limited range, typically $50 \leq Z_c \leq 200\,\Omega$, owing to natural values of ϵ and μ, and to the slow variation of the logarithm.

You must be careful when making connections with coaxial cable, so that the characteristic impedance Z_c of the cable is "matched" to the load impedance Z_L. The transmission equations are used to show that the "reflection coefficient" Γ, defined as the ratio of the current reflected from the end of the cable to the current incident on the end, is given by

$$\Gamma = \frac{Z_L - Z_c}{Z_L + Z_c}.$$

That is, if a pulse is transmitted along a cable and the end of the cable is not connected to anything ($Z_L = \infty$), then $\Gamma = 1$ and the pulse is immediately reflected back. On the other hand, if the end shorts the conductor to the shield ($Z_L = 0$), then $\Gamma = -1$ and the pulse is inverted and then sent back. *The ideal case is when the load has the same impedance as the cable. In this case, there is no loss at the end of the cable and the full signal is transmitted through.* You should take care in the lab to use cable and electronics that have matched impedances. Common impedance standards are 50 and 90 Ω.

Of course, you will need to connect your wire to the apparatus somehow, and this is done in a wide variety of ways. For permanent connections, especially inside electronic devices, solder is usually the preferred solution. It is harder than you might think to make a good solder joint, and if you are going to do some of this, you should have someone show you who has a decent amount of experience. Another type of permanent connection, called "crimping," squeezes the conductors together using a special tool that ensures a good contact that does not release. This is particularly useful if you cannot apply the type of heat necessary to make a good solder joint.

Less permanent connections can be made using terminal screws or binding posts. These work by taking a piece of wire and inserting it between two surfaces that are then forced together by tightening a screw. You may need to twist the end of the wire into a hook or loop to do this best, or you may use wire with some sort of attachment that has been soldered or crimped on the end. If you keep tightening or untightening screws, especially onto wires with handmade hooks or loops, then the wire is likely to break at some point. Therefore, for temporary connections, it is best to use alligator clips or banana plugs, or something similar. Again, you will usually use wires with this kind of connector previously soldered or crimped on the end.

Coaxial cable connections are made with one of several special types of connectors. Probably most common is the "bayonet N-connector," or BNC, standard, including male cable end connectors, female device connectors, and union and T-connectors for joining cables. In this system, a pin is soldered or crimped to the inner conductor of the cable, and the shield is connected to an outer metal holder. Connections are made by twisting the holder over the mating connector, with the pin inserting itself on the inner part. Another common connector standard, called "safe high voltage" or SHV, works similarly to BNC, but is designed for use with high DC voltages by making it difficult to contact the central pin unless you attach it to the correct mate.

For low-level measurement you must be aware of the thermal electric potential difference between two dissimilar conductors at different temperatures. These "thermoelectric coefficients" are typically around 1 μV/°C, but between copper and copper-oxide (which can easily happen if a wire or terminal is oxidized) it is around 1 mV/°C.

3.2.2. DC Power Supplies

Laboratory equipment needs to be "powered" in one way or another. Unlike the typical 100-V, 60-Hz AC line you get out of the wall socket, though, this equipment usually requires some constant DC level to operate. One way to provide this constant DC level is to use a battery, but if the equipment draws much current the battery will quickly run down. Instead we use DC "power supplies", the power supply in turn gets its power from the wall socket.

Power supplies come in lots of shapes, sizes, and varieties, but there are two general classes. These are "voltage" supplies or "current" supplies, and the difference is based on how the output is regulated. Since the inner workings of the power supply have some effective resistance, when the power supply must give some current, there will be a voltage drop across that internal resistance, which will affect how the power supply works. In a "voltage-regulated" supply, the circuitry is designed to keep the output voltage constant (to within some tolerance), regardless of how much current is drawn. (Typically, there will be some maximum current at which the regulation starts to fail. That is, there is a maximum power that can be supplied.) Most electronic devices and detector systems prefer to have a specific voltage they can count on, so they are usually connected to voltage-regulated supplies.

A "current-regulated" supply is completely analogous, but here the circuitry is designed to give a constant output current in the face of some load on the supply. Such supplies are most often used to power magnets, since the magnetic field only cares about how much current flows through the coils. This is in fact quite important for establishing precise magnetic fields, since the coils tend to get hot and change their resistance. In this case, $V = IR$ and R is changing with time, so the power supply must know to keep I constant by varying V accordingly. In many cases, a simple modification (usually done without opening up the box) can convert a power supply from voltage regulation to current regulation.

The output terminals on most power supplies are "floating." That is, they are not tied to any external potential, in particular not to ground. One output (sometimes colored in red) is positive with respect to the other (black). You will usually connect one of the outputs to some external point at known potential, like a common ground.

You should be aware of some numbers. The size and price of a power supply depends largely on how much power it can supply. If it provides a voltage V while sourcing a current I, then the power output is $P = IV$. A very common supply you will find around the lab will put out several volts and a couple of amperes, so something like 10 W or so. Depending on things like control knobs and settings to computer interfacing, they can cost anywhere from $50 up to a few hundred. So-called "high-voltage" power supplies will give several hundred up to several thousand volts, and can source anywhere from a few microamperes up to 100 mA, and keep the voltage constant to a level of better than 100 mV. Still, the power output of such devices is not enormously high, typically under a few hundred watts. The cost will run into thousands of dollars. Magnet power supplies, though, may be asked to run something like 50 A through a coil that has a resistance of, say, 2 Ω. In this case, the output power is 5 kW.

3.2.3. Waveform Generators

"Waveform generators" produce an output voltage signal $V(t)$ that varies in time. The function $V(t)$ can be anything from a simple sine wave to an arbitrary function you program into the device, but increased flexibility can cost a lot of money. Most waveform generators, though, do have at least sine waves, square waves, or triangle waves, and can vary the frequency over a wide range. Low frequencies are pretty easy to get, but for very high frequencies (above a megahertz or so) things get much harder because of stray capacitance giving effective shorts. You can also vary the amplitude and offset of the output voltage over several volts.

Sometimes instead of a "wave" output, one needs a "pulse"—that is, a signal that is high for some short period of time, with the next signal coming after a much longer time. Most waveform generators can accomodate your wishes either by providing an explicit "pulse" output, or by allowing you to change the symmetry of the waveform so that the "0 to π" portion of the wave is stretched or compressed relative to the "π to 2π" portion.

3.2.4. Meters

Now that you know how to obtain some voltages, including time-varying ones, and how to connect these voltages using wire and cable, you must think about how to measure the voltage. The simplest way to do this is with a meter, particularly if the voltage is DC. (Most meters do provide you with AC capability, but we will not go into the details here.) An excellent reference on the subject of meters is given in the *Low Level Measurements Handbook*, published by Keithley Instruments, Inc. This handbook, as well as other materials, are available from Keithley at http://www.keithley.com/.

At one time, people would use either voltmeters, ammeters, or ohmmeters to measure voltage, current, or resistance, respectively. These days, although you still might want to buy one of these specialized instruments to get down to very low levels, most measurements are done with "digital multimeters," or DMMs for short. (In fact, some DMMs are available now that can effectively take the place of the most sensitive specialized meters.) Voltage and resistance measurements are made by connecting the meter in parallel to the portion of the circuit you are interested in. To measure current, the meter must be in series.

Realize that DMMs work by averaging the voltage measurement over some period of time, and then displaying the result. This means that if the voltage is fluctuating on some time scale, these fluctuations will not be observed if the averaging time is greater than the typical period of the fluctuations. Of course the shorter the averaging time a meter has (the higher the "bandwidth" it has), the fancier it is and the more it costs.

Meters have some effective input impedance, so they will (at some level) change the voltage you are trying to measure. For this reason, voltmeters and ohmmeters are designed to have very large input impedances (many megaohms to as high as several gigaohms), while ammeters "shunt" the current through a very low resistance and turn the job into measuring the (perhaps very low) voltage drop across that resistor.

3.3. OSCILLOSCOPES AND DIGITIZERS

3.3.1. Oscilloscopes

An oscilloscope measures and displays voltage as a function of time. That is, it plots for you the quantity $V(t)$ on a cathode ray tube (CRT) screen as it comes in. This is a very useful thing, and you will use oscilloscopes in nearly all the experiments you do. A good reference is *The XYZ's of Oscilloscopes*, published by Tektronix, Inc. You can download a copy from http://www.tek.com/ under "Application Notes for Oscilloscopes."

The simple block diagram shown in Fig. 3.14 explains how an oscilloscope works. The voltage you want to measure serves two purposes. First, after being amplified, it is applied to the vertical deflection plates of the

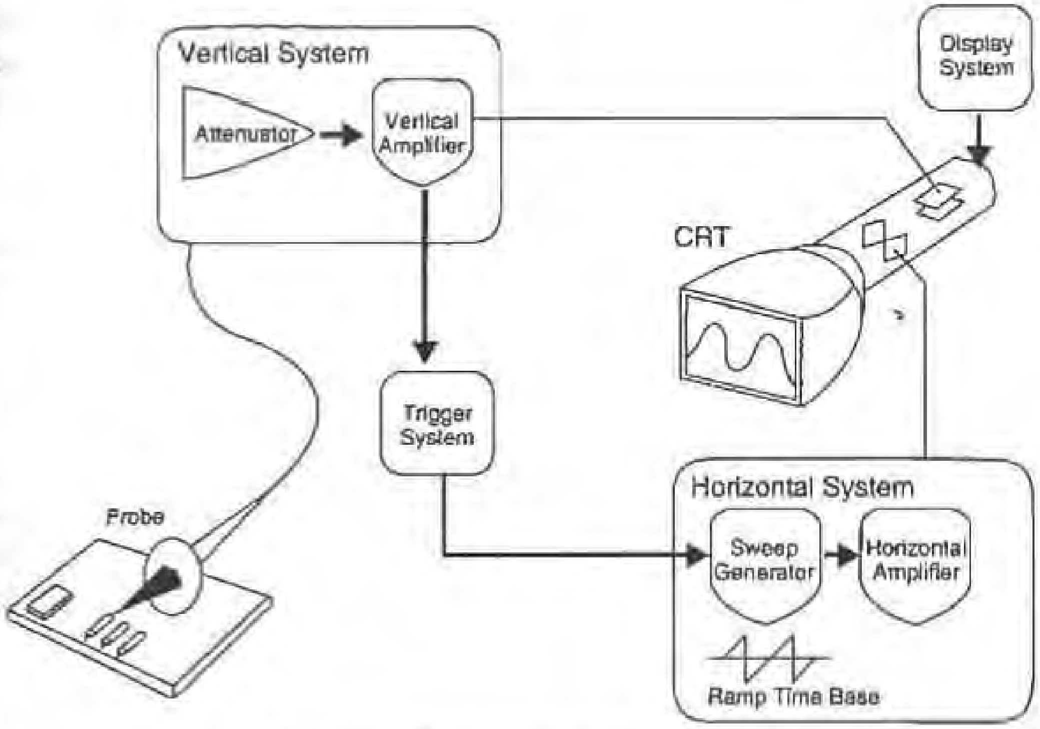


FIGURE 3.14 Block diagram of an oscilloscope.

CRT. This means that the vertical position of the trace on the CRT corresponds linearly to the input voltage, which is just what you want. The vertical scale on the CRT has a grid pattern that lets you know what the input voltage is.

The horizontal position of the trace is controlled by a "sweep generator" whose speed you can control. However, for repetitive signal shapes, you want the signal to "start" at the same time for every sweep, and this is determined by the "trigger" system. The place on the screen where the trace starts is controlled by a "horizontal position" knob on the front panel. One kind of trigger is to just have the scope sweep at the line (i.e., 60 Hz) frequency, but this will not be useful if the signals you are interested in do not come at that frequency. Another kind of simple trigger is to have the trace sweep once whenever the voltage rises or falls past some level, i.e., a "leading edge" trigger. There is usually a light on the front panel that flashes when the scope is triggered.

Oscilloscopes almost always have at least two input channels, and it is possible to trigger on one channel and look at the other. This can be very useful for studying coincident signals or for measuring the relative phase of two waveforms. In any case, the trigger "mode" can either be "normal," in which case there is a sweep only if the trigger condition is met, or "auto" where the scope will trigger itself if the trigger condition is not met in some period of time. Auto mode is particularly useful if you are searching for some weak signal and do not want the trace to keep disappearing on you.

You have several controls on how the input voltage is handled. A "vertical position" knob on the front panel controls where the trace appears on the screen. You will find one of these for each input channel. The input "coupling" can be set to either AC, DC, or ground. In AC mode, there is a capacitor between the input connector and the vertical system circuit. This keeps any constant DC level from entering the scope, and all you see is the time-varying (i.e., AC) part. If you put the scope on DC, then the constant voltage level also shows up. If the input coupling is grounded, then you force the input level to 0, and this shows you where 0 is on the screen. (Make sure that the scope is on "auto" trigger if you ground the input; otherwise, you will not see a trace!)

Sometimes, you also get to choose the input impedance for each channel. Choosing the "high" input impedance (usually 1 MΩ) is best if you want to measure voltage levels and not have the oscilloscope interact with the circuit. However, the oscilloscope will get a lot of use looking at fast pulsed signals transmitted down coaxial cable, and you do not want an "impedance mismatch" to cause the signal to be reflected back. (See Section 3.2.1.) Cables with 50-Ω characteristic impedances are very common in this work, so you may find a 50-Ω input impedance option on the scope. If not, you should use a "tee" connector on the input to put a 50-Ω load in parallel with the input.

By flipping switches on the front, you can look at either input channel's trace separately, or both at the same time. There is obviously a problem, though, with viewing both simultaneously since the vertical trace can only be in one place at a time. There are two ways to get around this. One is to *alternate* the trace from channel one to channel two and back again. This gives complete traces of each, but does not really show them to you at the same time. If the signals are very repetitive and you are not interested in fine detail, this is okay. However, if you really want to see the traces at the same time, select the *chop* option. Here, the trace jumps back and forth between the channels at some high frequency, and you let your eye interpolate between the jumps. If the sweep speed is relatively slow, the interpolation is no problem and you probably cannot tell the difference between *alternate* and *chop*. However, at high sweep speed, the effect of the chopping action will be obvious.

You should realize by now that high-frequency operation gets hard, and the oscilloscope gets more complicated and expensive. Probably the single most important specification for an oscilloscope is its "bandwidth," and you will see that number printed on the front face right near the screen. The number tells you the frequency at which a sine wave would appear only 71% as large as it should be. You cannot trust the scope at frequencies approaching or exceeding the bandwidth. Most of the scopes in the lab have 20- or 60-MHz bandwidths. A "fast" oscilloscope will have a bandwidth of a few hundred megahertz or more. You will find that you can vary the sweep speed over a large range, but never much more than $(\text{bandwidth})^{-1}$. The "vertical sensitivity" can be set independently of the sweep speed, but scopes in general cannot go below around 2 mV/division.

On most oscilloscopes, if you turn the sweep speed down to the lowest value, one more notch puts the scope in the XY display mode. Now, the trace displays channel one (X) on the horizontal axis and channel two (Y) on the vertical. For periodic signals, the trace is a Lissajous pattern from which you can determine the relative phase of the two inputs. Oscilloscopes are also used in this way as displays for various pieces of equipment which have XY output options. Thus, the oscilloscope can be used as a plotting device in some cases.

3.3.2. Digitizers

In order to measure a voltage and deal with the result in a computer, the voltage must be *digitized*. The generic device that does this is the analog-to-digital converter or ADC. ADCs come in approximately an infinite number of varieties and connect to computers in lots of different ways. We will cover the particulars when we discuss the individual experiments, but for now we will review some of the basics.

Probably the most important specification for an ADC is its resolution. We specify the resolution in terms of the number of binary digits ("bits") that the ADC spreads out over its measuring range. The actual measuring range can be varied externally by some circuit, so the number of bits tells you how finely you can chop that range up. Obviously, the larger the number of bits, the closer you can get to knowing exactly what the input voltage was before it was digitized. A "low-resolution" ADC will have 8 bits or less. That is, it divides the input voltage up into 256 pieces and gives the computer a number between 0 and 255, which represents the voltage. A "high-resolution" ADC has 16 bits or more.

High resolution does not come for free. In the first place, it can mean a lot more data to handle. For example, if you want to histogram the voltage being measured with an 8-bit ADC, then you need 256 channels for each histogram. However, if you want to make full use of a 16-bit ADC, every histogram would have to consume 65,536 channels. Resolution also affects the *speed* at which a voltage can be digitized. Generally speaking, it takes much less time to digitize a voltage into a smaller number of bits than it does for a large number of bits.

There are three general classes of ADCs, referred to as *flash*, *peak-voltage sensing*, and *charge integrating* ADCs. A flash ADC, or "waveform recorder," simply reads the voltage level at its input and converts that voltage level into a number. They are typically low resolution, but run very fast. Today you can easily get an 8-bit flash ADC that digitizes at 100 MHz (i.e., one measurement every 10 ns). This is fast enough so that just about any time-varying signal can be converted to numbers so that a true representation of the signal can be stored in a computer.

To get better resolution, you need to decide what it is about the signal you are really interested in. For example, if you only care about the maximum voltage value, you can use a peak-sensing ADC, which digitizes the maximum voltage observed during some specified time. Sometimes, you are interested instead in the area underneath some voltage signal. This is the case, for example, in elementary particle detectors where the net charge delivered is a measure of the particle's energy. For applications like this, you can use an integrating ADC, which digitizes the net charge absorbed over some time period, i.e., $(1/R)\int_{t_1}^{t_2} V(t)\,dt$, where R is the resistance at the input. For either of these types, you can buy commercial ADCs that digitize into 12 or 13 bits in 5 μs or longer, but remember that faster and more bits costs more money.

The opposite of an ADC is a DAC, or digital-to-analog Converter. Here the computer feeds the DAC a number depending on the number of bits, and the DAC puts out an analog voltage proportional to that number. The simplest DAC has just one bit, and its output is either "on" or "off." In this case, we refer to the device as an "output register." These devices are a way of controling external equipment in an essentially computer-independent fashion.

In many cases, you want to digitize a time interval instead of a voltage level. This can be done with a "time-to-analog converter" (TAC), followed by an ADC. However, both of these functions are now available packaged in a single device called a TDC. The rules and ranges are very similar as for ADCs.

Devices known as "latches" or "input registers" will take an external logic level, and digitize the result into a single bit. These are useful for telling whether some device is on or off, or perhaps if something has happened that the computer should know about.

When a device is busy digitizing, it cannot deal with more input. We refer to the cumulative time a device is busy as "dead time." Suppose τ is the time needed to digitize an input pulse, and R_0 is the (presumably random) rate at which pulses are delivered to the digitizer. If R_m is the *measured* rate, then in a time T the number of digitized pulses is R_mT. The dead time incurred in time T is therefore $(R_mT)\tau$, so the number of pulses lost is $[(R_mT)\tau]R_0$. The total number of pulses delivered (R_0T) must equal the number digitized plus the number lost, so

$$R_0T = R_mT + R_mT\tau R_0,$$

and therefore

$$R_m = \frac{R_0}{1 + \tau R_0} \tag{3.13}$$

or

$$R_0 = \frac{R_m}{1 - \tau R_m}. \tag{3.14}$$

The "normal" way to operate a digitizer is so that it can keep up with the rate at which pulses come in. In other words, the rate at which it digitizes $(1/\tau)$ should be much greater than the rate at which pulses are delivered, that is, $\tau R_0 \ll 1$. Equation (3.13) shows that in this case, $R_m \approx R_0$; that is, the measured rate is very close to the true rate, which is just what you want. Futhermore, an accurate correction to the measured rate is given by Eq. (3.14), which can be written as $R_0 = R_m(1 + \tau R_m)$ under normal operation.

On the other hand, if $\tau R_0 \gg 1$, then $R_m \approx 1/\tau$. That is, the digitizier measures a pulse and before it can catch its breath, another pulse comes along. The device is "always dead," and the measured rate is just one per digitizing time unit. Essentially all information on the true rate is lost, because the denominator of Eq. (3.14) is close to 0. You would have to know the value of τ very precisely in order to make a correction that gives you the true rate.

3.3.3. Digital Oscilloscopes

The digital oscilloscope is a wonderful device. Instead of taking the input voltage and feeding it directly onto the deflection plates of a CRT (Fig. 3.14), a digital oscilloscope first *digitizes* the input signal using a flash ADC, stores the waveform in some internal memory, and then has other circuitry to read that memory and display the output on the CRT. We then have the voltage stored as numbers, and the internal computer in the digital oscilloscope can do just about anything with the numbers. Even though it works very differently from analog oscilloscopes, digital scopes have controls that make them look as much like analog scopes as possible. The same terminology is used, and just about any function found on an analog scope will also be found on a digital one.

3.4. SIMPLE MEASUREMENTS

We now outline some simple measurements of elementary circuits. Circuits are most easily put together on a "breadboard." This is a flat, multilayered surface with holes in which you stick the leads of wires, resistors, capacitors, and so on. The holes are connected internally across on the component pads, and downward on the power pads.

Connect two 1-kΩ resistors in series on the breadboard, and then connect the terminals of the power supply to each end of this two-resistor string. Measure the voltage across the output of the terminals. Also, measure the current through the string. Now connect two more 1-kΩ resistors in series with the others. Move the connections from the power supply so that once again it is connected to each end of the string. Repeat your voltage and current measurements. Now measure the voltage drop across each of the four resistors. Compare the result to what you expect based on the voltage divider relation. Use your data and Ohm's law to measure the resistance of each of the resistors. Compare the resistance values you measure with the nomiual value.

Remove the DC power supply and replace it with a waveform generator. Set the waveform to a sine wave. Use an oscilloscope to compare the voltage (as a function of time) across the resistor string from the waveform generator with the voltage across one of the resistors. Put each of these into the two channels of the oscilloscope, and trigger the scope on the channel corresponding to the waveform generator output. Look at both traces simultaneously (on either *chop* or *alternate*) and compare the relative amplitudes of the "input" sine wave across the string, and the "output" sine wave across the single resistor.

Now connect a resistor and capacitor in series. Choose a resistance R and capacitance C so that the inverse time constant $1/RC$ is well within the frequency range of the waveform generator and the oscilloscope. Just as you did for the resistor string, measure the amplitude of the voltage across either the resistor or capacitor, relative to the waveform generator signal applied across the front and back of the pair. (You should take care to set the DC offset of the waveform generator to 0 using the oscilloscope to measure the offset relative to ground.) Do this as a function of frequency, spanning well on either side of $1/RC$. Also measure the phase of the output sine wave, relative to the input sine wave. Figure 3.15 shows how to make these measurements on the oscilloscope CRT, using the circuit shown. Refer to Fig. 3.8 for interpreting the input and output waveforms in terms of gain and phase. It would be a good idea to select your frequency values logarithmically instead of linearly. That is, use ν_0, $2\nu_0$, $4\nu_0$, ..., $\nu_{\max}$ where ν_0 is your starting low frequency. Make a clear table of your measurements and plot the gain (i.e., the relative amplitudes) and the relative phase as a function of frequency. Do not forget that you measure frequency ν, but most

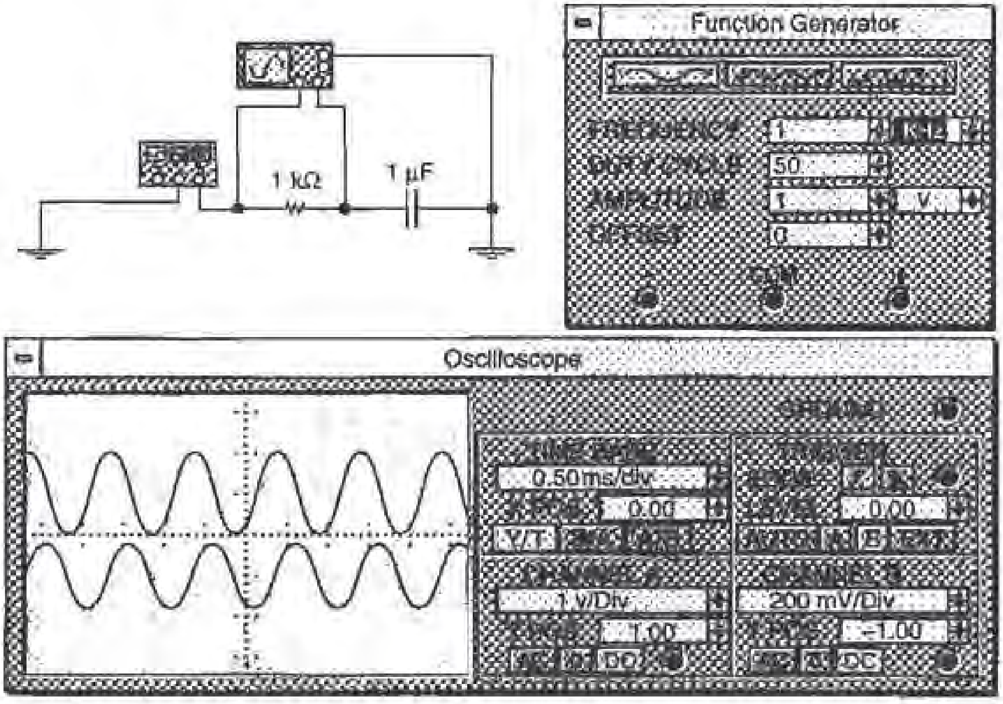


FIGURE 3.15 Measuring gain and relative phase on an oscilloscope.

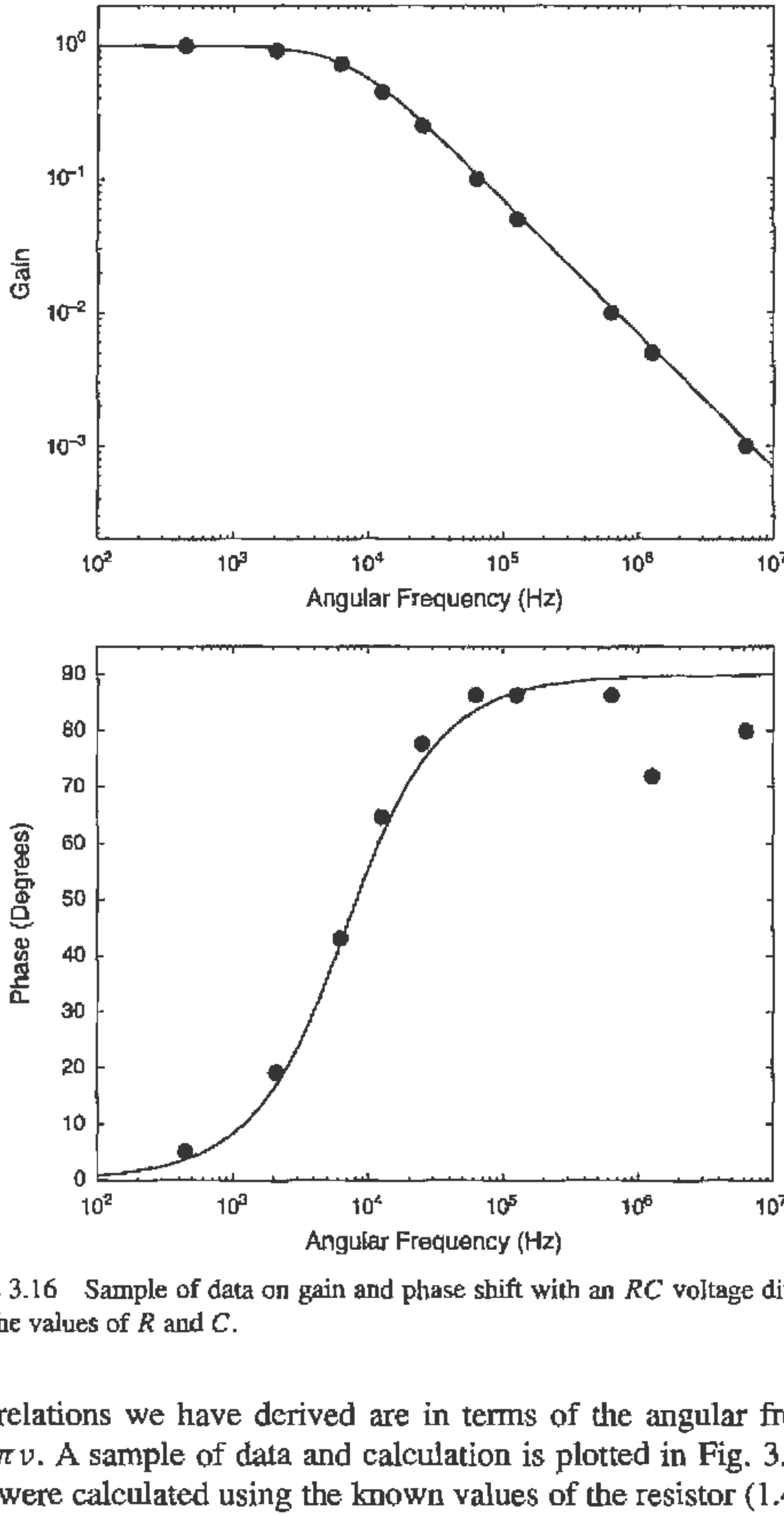


FIGURE 3.16 Sample of data on gain and phase shift with an *RC* voltage divider. See text for the values of *R* and *C*.

of the relations we have derived are in terms of the angular frequency $\omega = 2\pi\nu$. A sample of data and calculation is plotted in Fig. 3.16. The curves were calculated using the known values of the resistor (1.453 kΩ) and capacitor (0.1 μF).

Finally, use the waveform generator as a pulse generator and study the output using your RC voltage divider circuit. Compare the input and output pulse shapes as a function of the width Δt of the pulse. What happens if $\Delta t \gg RC$? What about $\Delta t \ll RC$?

3.5. OPERATIONAL AMPLIFIERS

Noise can get in the way of your measurements by causing things to change when you do not want it. These changes can happen as a function of time, frequency, temperature, etc. To fight this, you want your apparatus to be stable against time, frequency, temperature, etc. The most common way to achieve this is using *negative feedback*. The idea behind negative feedback is that you take a part of the "output" and *subtract* it away from the "input," causing it to "feed back" to the output and discourage it from changing.

Consider a generic amplifier, like that shown in Fig. 3.17, which amplifies the difference voltage between its inputs to give an output voltage. Let the gain of the amplifier be α. That is, for the circuit in Fig. 3.17 we have $V_{\rm out} = \alpha V_{\rm in}$. We apply negative feedback by taking some of the output voltage and subtracting it from the input. This is shown in Fig. 3.18. A resistor voltage divider is used to take a fraction $\beta = R_2/(R_1 + R_2)$ of the output voltage $V_{\rm out}$ and subtract it from the input. The amplifier now does not amplify $V_{\rm in}$ directly, but instead amplifies $V_{\rm dif} = V_{\rm in} - \beta V_{\rm out}$. That is,

$$V_{\rm out} = \alpha V_{\rm dif} = \alpha V_{\rm in} - \alpha\beta V_{\rm out},$$

and the net gain g is

$$g = \frac{V_{\rm out}}{V_{\rm in}} = \frac{\alpha}{1 + \alpha\beta}. \tag{3.15}$$

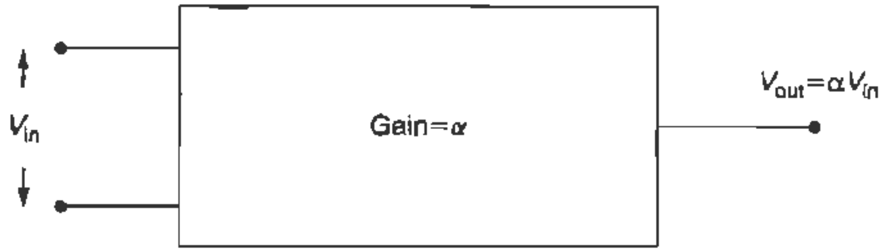


FIGURE 3.17 A generic amplifier.

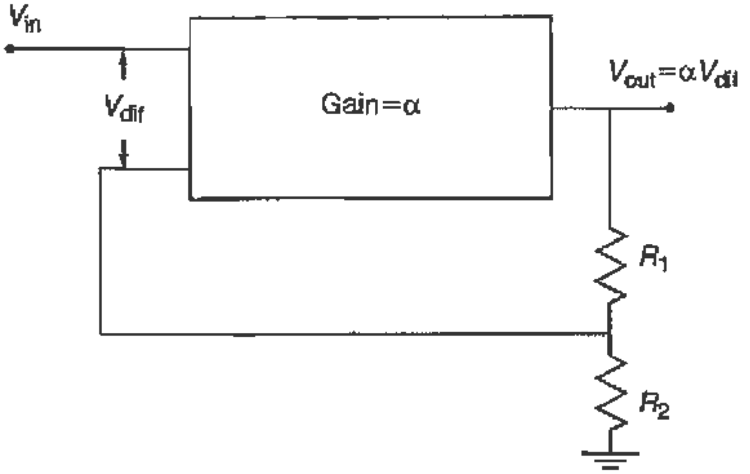


FIGURE 3.18 A generic amplifier with negative feedback.

Now's here the key point. The generic amplifier is designed so it has enormous gain. That is, α is very, very large. So large, in fact, that $\alpha\beta \gg 1$, no matter how small β is. That means that the gain is

$$g = \frac{1}{\beta} = 1 + \frac{R_1}{R_2} \qquad \text{for } \alpha\beta \gg 1. \tag{3.16}$$

The gain of the system only depends on the ratio of a pair of resistor values, and not on the gain of the generic amplifier. It is hard to get resistor values to change, so this amplifier circuit is very stable. The generic amplifier with gain α, however, is likely to depend a lot on frequency, temperature, and so on.

As you might imagine, commercial versions of the generic amplifier shown in Fig. 3.17 are available in lots of flavors. They are called *operational amplifiers* or *opamps* for short. Instead of a box, they are represented by a triangle, as shown in Fig. 3.19. The two inputs are labeled "+" and "−" for phase considerations. The $+V$ and $-V$ terminals are where you apply a voltage source to power the opamp. It is common to leave these off of schematic circuit diagrams. Opamps are cheap. Most cost less than \$1, although you can pay a lot if you want special properties. All have very large gain, i.e., α upward of 10^4 or more, up to some frequency. (Remember that capacitance kills circuits at high frequency because it becomes a short.) An old, popular opamp is the model 741, which is still widely used today. A version of the 741 in standard use today (the LF411) has a gain of at least 88 dB (i.e., $\alpha \geq 2.5 \times 10^4$) and can be used up to frequencies of tens of kilohertz or more, depending on the feedback circuit.

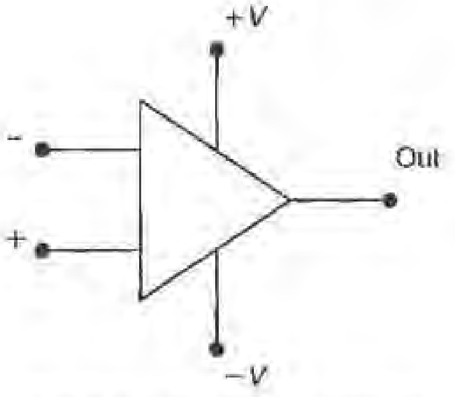


FIGURE 3.19 Opamp notation.

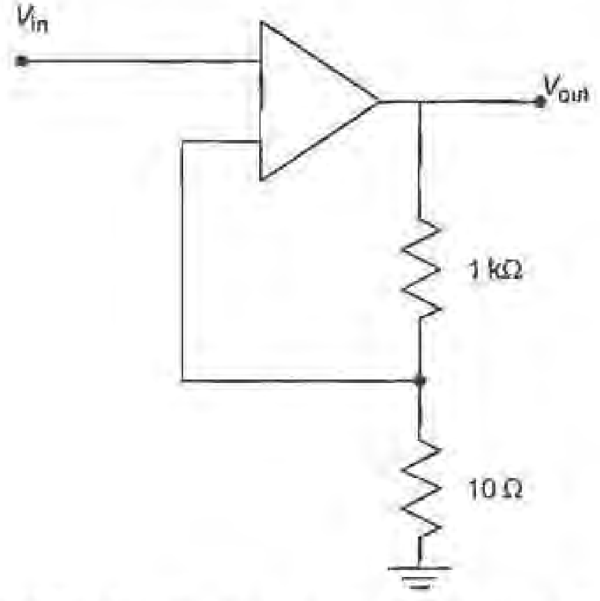


FIGURE 3.20 An amplifier circuit with gain of 100.

Horowitz and Hill (1989; see Section 3.10) tabulate the properties of your garden variety opamps. They also tell the interesting story of how opamps were developed, and why the 741 is such a mainstay. A common use of opamps, of course, is just as a negative feedback amplifier. You pick $R_1 \gg R_2$ so that the gain given by Eq. (3.16) is $g \approx R_1/R_2$. For example, to build a stable amplifier with a gain of ~100 up to a kilohertz or so, you might build the circuit shown in Fig. 3.20.

Another application of opamps connects to our discussion of passive filters. (See Section 3.1.5.) The effective input impedance of an opamp in negative feedback is huge. That is because even though you apply a voltage V_{in}, the input to the opamp is $V_{dif} = V_{in} - \beta V_{out} \approx V_{in} - \beta(V_{in}/\beta) = 0$

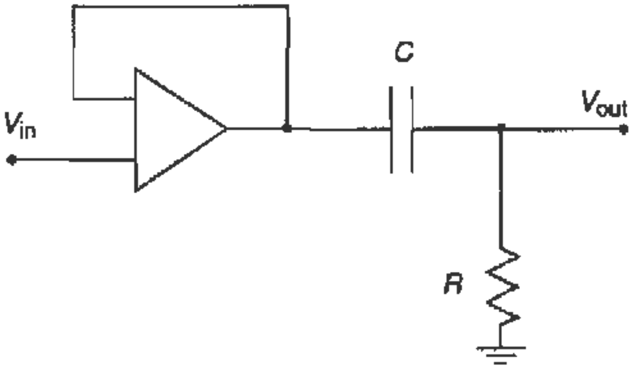


FIGURE 3.21 A high-pass filter with input load buffering.

so it draws no current. This makes the opamp ideal for "load buffering." That is, you can use it to make the input to some device (like a filter or perhaps a meter) large enough so that you can ignore its effect on the circuit that feeds it. For instance, you might build a high pass filter as shown in Fig. 3.21. *All* the output of the opamp is fed back to the input, thus $\beta = 1$ and $g = 1$. However, $Z_{\text{in}} = \infty$ (effectively) because of the opamp, so all this circuit does is cut off the output of the source for $\omega < 1/RC$ like a good high-pass filter should. If the opamp were not there, you would need to add in the filter input impedance $Z_{\text{filter}} = R + 1/i\omega C$ to the source circuit. See Dunlap (1988) for further clever variations on active filters.

3.6. MEASUREMENTS OF JOHNSON NOISE

In this experiment, we will measure a very fundamental source of noise. It has to do with the motion of electrons in a conductor and the heat energy (random motion) associated with them. This is called "Johnson noise" because it was originally measured by J. B. Johnson. Some people call it "Nyquist noise," because the phenomenon Johnson measured was first correctly explained by H. Nyquist. A more generic term is "thermal noise." Some journal articles on similar experiments are listed at the end of this chapter. You might also want to go back and look at the original work of Johnson and Nyquist, published in J. B. Johnson, "Thermal Agitation of Electricity in Conductors," *Phys. Rev.* **32**, 97 (1928), and H. Nyquist, "Thermal Agitation of Electric Charge in Conductors," *Phys. Rev.* **32**, 110 (1928).

3.6.1. Thermal Motion of Electrons

We will outline a simple model of thermal noise as presented by W. Henry (see references). The model is based on random thermal fluctuations of electrons in a one-dimensional resistor of length L and cross-sectional area A. The resistor has resistance R, and a voltage drop $V = IR$ across the ends. The current I, and therefore the voltage V, arises from the thermal fluctuations that allow more electrons to move one way than another in some short time interval t_0.

On average no current flows through the resistor, and the average value of V is zero. That is,

$$\langle V \rangle = 0.$$

On the other hand, the thermal fluctuations still give rise to a finite voltage as a function of time; in other words $V(t) \neq 0$. Therefore, the variance[4] of V is not zero; namely,

$$\sigma_V^2 = \langle (V - \langle V \rangle)^2 \rangle = \langle V^2 \rangle - \langle V \rangle^2 = \langle V^2 \rangle \neq 0.$$

This quantity $\langle V^2 \rangle = \sigma_V^2$ is called the thermal or Johnson noise voltage, and it is what we will measure in this experiment.

From Ohm's law and the definitions of current and charge, we can write

$$\begin{aligned} \sigma_V &= \sigma_i R \\ &= \frac{\sigma_q}{t_0} R \\ &= \frac{e\sigma_x / L}{t_0} R, \end{aligned}$$

where L is the length of the resistor, and σ_x is the net x motion of all the electrons in the measuring time t_0. If we can reduce this to the motion of an individual electron, then we can use a microscopic description of current and resistance. If there is a total of N independent and random electron motions (i.e., "random walks") in time t_0, then

$$\sigma_x = \sqrt{N} \sigma_d,$$

[4] The student may want to review various definitions in the theory of statistics, given in Chapter 10.

where $\sigma_{\rm d}$ is the average distance that any single electron moves. Therefore,

$$\sigma_V = \frac{e}{L}\sqrt{N}\frac{\sigma_{\rm d}}{t_0}R. \tag{3.17}$$

Now N is the total number of conduction electrons in the resistor times the number of walks in time t_0, so

$$N = (nAL) \times \frac{t_0}{\tau} = \frac{nALt_0}{\tau},$$

where n is the number density of conduction electrons and τ is the time between collisions of a *single* electron. The fluctuation in the motion of a single electron is

$$\sigma_{\rm d}^2 = \langle d^2 \rangle = \langle v_x^2 \tau^2 \rangle = \langle v_x^2 \rangle \tau^2,$$

and this is what we connect to temperature by $\langle E \rangle = \frac{1}{2}m\langle v_x^2 \rangle = \frac{1}{2}kT$, where m is the mass of an electron and we note that motion is only in one dimension. The factor k is Boltzmann's constant which defines the fundamental relationship between temperature and internal energy. Therefore

$$\sigma_{\rm d}^2 = \frac{kT\tau^2}{m}.$$

We note that (see Eq. (2.14))

$$\frac{L}{A}\frac{2m}{ne^2\tau} = \frac{L}{A}\rho = R,$$

where ρ is the resistivity.[5]

Finally, put this all into Eq. (3.17) to get

$$\begin{aligned}\sigma_V^2 &= \frac{e^2}{L^2}N\frac{\sigma_{\rm d}^2}{t_0^2}R^2 \\ &= \frac{e^2}{L^2}\frac{nALt_0}{\tau}\frac{kT\tau^2}{mt_0^2}R^2 \\ &= \frac{A}{L}\frac{ne^2\tau}{m}\frac{kT}{t_0}R^2,\end{aligned}$$

[5] The definition of τ used here differs from that used in Section 2.2 by a factor of 2. That is because we are dealing with a single electron.

or

$$\langle V^2\rangle = \frac{2kTR}{t_0}. \tag{3.18}$$

It is customary, however, to express the noise using the equivalent bandwidth $\Delta\nu = 1/2t_0$. Therefore, we have

$$\langle V^2\rangle = 4kTR\Delta\nu. \tag{3.19}$$

In order to measure the voltage V, we will need to amplify or at least process the signal in some way. Let $g(\nu)$ be the gain of this processing circuit at frequency ν. Then the output voltage fluctuation $d\langle V^2\rangle$ integrated over some small frequency range $d\nu$ is given by

$$d\langle V^2\rangle = 4kTRg^2(\nu)\,d\nu.$$

Measurements are made by integrating the signal over a relatively large bandwidth $\Delta\nu$. This bandwidth is typically determined by the gain function $g(\nu)$, which is large only over some finite frequency range. We therefore obtain the expression

$$\langle V^2\rangle = 4kTRG^2\Delta\nu, \tag{3.20}$$

where G and $\Delta\nu$ are constants defined by

$$G^2\Delta\nu \equiv \int_0^\infty g^2(\nu)\,d\nu. \tag{3.21}$$

3.6.2. Measurements

We will measure the Johnson noise in a series of resistors, and use the result to determine a value for Boltzmann's constant k.

The setup is shown schematically in Fig. 3.22. The voltage across the resistor R is immediately processed by an "amplifier," which essentially multiplies this voltage by a function $g(\nu)$. The output of the amplifier is measured using a digital oscilloscope. You will use the oscilloscope to measure $\langle V^2\rangle$, given by Eq. (3.20). By changing the value of R (simply by changing resistors), you measure $\langle V^2\rangle$ as a function of R, and the result should be a straight line. The slope of the line is just $4kTG^2\Delta\nu$, so once you have calibrated the gain function of the amplifier, you can get k. (You can assume the resistor is at room temperature.)

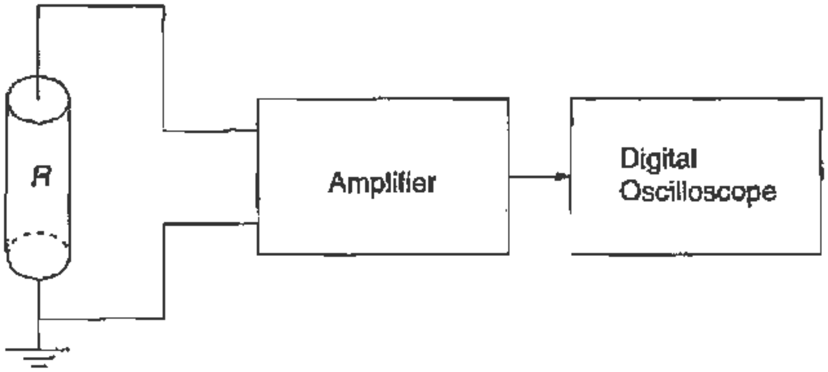


FIGURE 3.22 Schematic for measuring Johnson noise.

Let's look a little more carefully at the properties of the amplifier. We will be working in the several tens of kilohertz range, so to estimate the gain we need, take a bandwidth $\Delta\nu = 10$ kHz. The digital oscilloscope cannot make measurements much smaller than around 0.5 mV, so Eq. (3.20) implies that the nominal gain G must be on the order of 1200 or more to measure the noise in a 1-kΩ resistor. The amplifier also needs to have low noise and good stability itself, if we are going to use it on such a small signal. A high-gain opamp with negative feedback (see Section 3.5) sounds like the right solution.

The bandwidth of the amplifier also needs to be considered. In fact, if we are going to do the job right, we want to make sure that all the bandwidth limitations are given by the amplifier, and not by the oscilloscope, for example. That way, we can measure the function $g(\nu)$ of the amplifier stage only. The oscilloscope bandwidth will depend on the timebase used, that is, the time over which the output voltage is averaged and digitized. As long as the oscilloscope's bandwidth is greater than the amplifier's, you will be OK. You ensure this by putting a bandwidth filter on the output of the amplifier. In the beginning, you will use a commercial bandwidth filter with adjustable lower and upper limits.

The first "amplifier" you will use, therefore, is shown in Fig. 3.23. For now the bandwidth filter is just a box with an input and output, and with knobs you can turn. The gain-producing part of the amplifier, on the other hand, is essentially a cut-and-dry application of opamps and negative feedback. In fact, as shown in Fig. 3.23, two such negative feedback loops are cascaded to get the appropriate gain and input characteristics. The first loop uses a HA5170 opamp and a low gain, while the second stage is higher

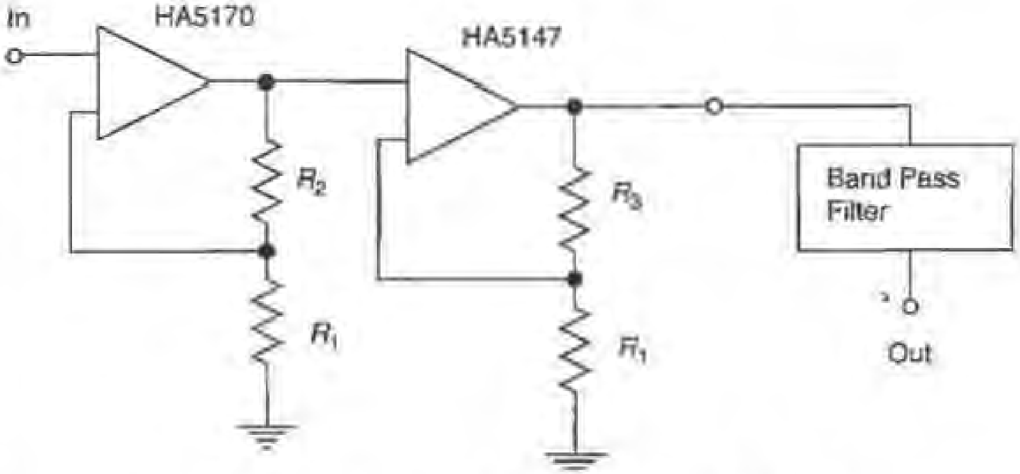


FIGURE 3.23 Amplifier stage for measurement of Johnson noise.

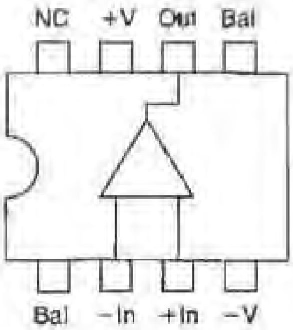


FIGURE 3.24 Pinout diagram for the opamp chips used in this experiment. We are not using the "Bal" connections. The notation "NC" means "no connection."

gain and uses a HA5147.[6] Good starting values to use are $R_1 = 10\,\Omega$, $R_2 = 100\,\Omega$, and $R_3 = 2.2\,\mathrm{k}\Omega$. This gives the first stage a gain of 11 and the second stage a gain of 221 times the bandwidth function imposed by the opamps and the bandwidth filter.

All of these components, including your input resistor R (but not the commercial bandwidth filter), are mounted on a breadboard so you can change things easily. The pinout diagram for the HA5170 and HA5147 is shown in Fig. 3.24. The opamps are powered by ±12-V levels applied in parallel with 0.1-μF capacitors to ground, to filter off noise in the power supply. Connections to the breadboard are made using wires soldered to BNC connectors.

[6] The credit for figuring out the right opamps and amplifier circuit in general goes to Jeff Pedison, RPI Class of '94. More details on this circuit design are available.

Set up the circuit shown in Fig. 3.23. Check things carefully, especially if you are not used to working with breadboards. In particular, make sure the 12-V DC levels are connected properly, before you turn the power supply on. The output from the breadboard gets connected to the bandwidth filter, and the output of the bandwidth filter goes into the oscilloscope. The lower and upper limits of the bandwidth filter are not crucial, but 5 and 20 kHz are a reasonable place to start.

First you need to measure the gain of the amplifier/bandwidth filter as a function of frequency. All you really need to do is put a sine wave input to the circuit and measure the output on an oscilloscope. The output should look the same as the input (i.e., a sine wave of the same frequency ν), but the amplitude should be bigger. The ratio of the output to input amplitudes is just the gain $g(\nu)$. There is a problem, though. You have built an amplifier of very large gain, around 2.4×10^3, and the output amplitude must be less than a few volts so the opamps do not saturate. That means that the input must be less than a couple of millivolts. That is barely enough to see on an oscilloscope, assuming your waveform generator can make a good sine wave with such a small amplitude.

You get around this problem by using the schematic shown in Fig. 3.25. The waveform generator output passes through a voltage divider, cutting the amplitude down by a known factor. This divided voltage is used as input to the amplifier. It is a good idea to measure the resistor values R_{big} and R_{small} using an ohmmeter, rather than to trust the color code (which can be off by up to 10%). Pick resistors that give you a divider ratio somewhere between 10 and 100. It is also a good idea to see the output of the waveform generator and look at it on the oscilloscope along with the amplifier/bandwidth filter output.

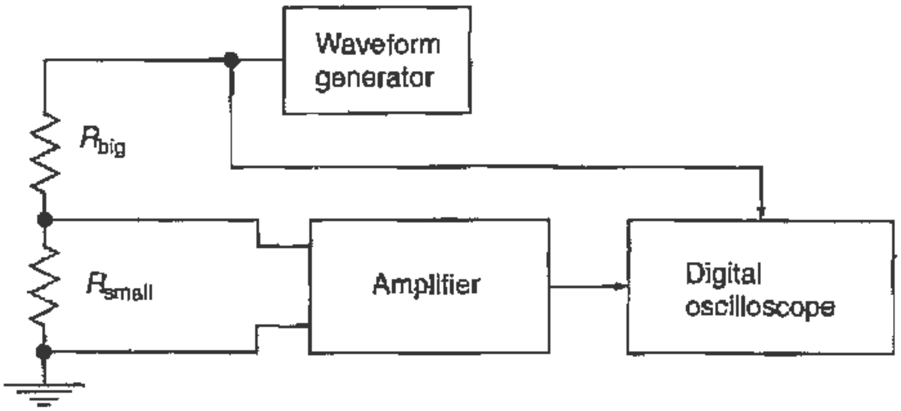


FIGURE 3.25 Calibration scheme for the noise amplifier.

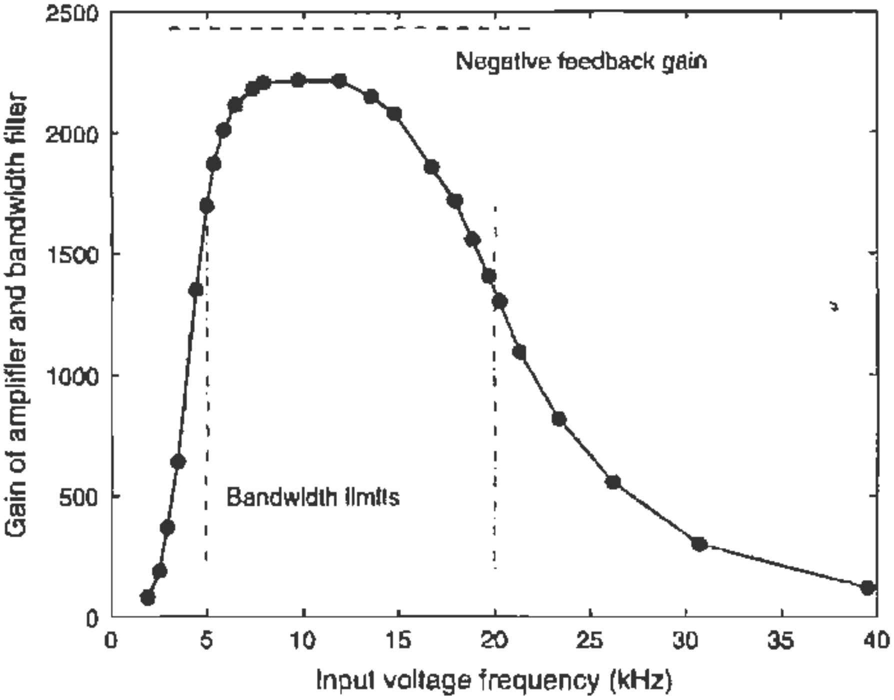


FIGURE 3.26 Sample of data used to determine $g(\nu)$ for the amplifier followed by the commercial bandwidth filter. The simple negative feedback formula gives a gain of 2431, and the bandwidth filter is set for $\nu_{LO} = 5$ kHz and $\nu_{HI} = 20$ kHz.

Make your measurements of $g(\nu)$ by varying the frequency of the waveform generator, and recording the output amplitude. Of course, you must also record the input (i.e., generator) amplitude, but if you check it every time you change ν, you can be sure it does not change during your measurement. Measure over a range of frequencies that allows you to clearly see the cutoffs from the bandwidth filter, including the shape as g approaches zero. Also make sure you confirm that the gain is relatively flat in between the limits. An example is shown in Fig. 3.26. The setup used $R_1 = 10\,\Omega$, $R_2 = 100\,\Omega$, and $R_3 = 2.2\,\mathrm{k}\Omega$, so the total gain should be 2431, and with bandwidth filter limits at 5 and 20 kHz. The main features seem to be correct, although the filter has apparently decreased the maximum gain a bit.

Now take measurements of the actual Johnson noise as a function of R. Remove the waveform generator and voltage divider inputs, and put the resistor you want to measure across the input to the amplifier. Set the time per division on the oscilloscope so that its bandwidth limit is much larger than the upper frequency you used on the bandwidth filter. For example, if there are 10,000 points (i.e., samples) per trace and you set the scope to

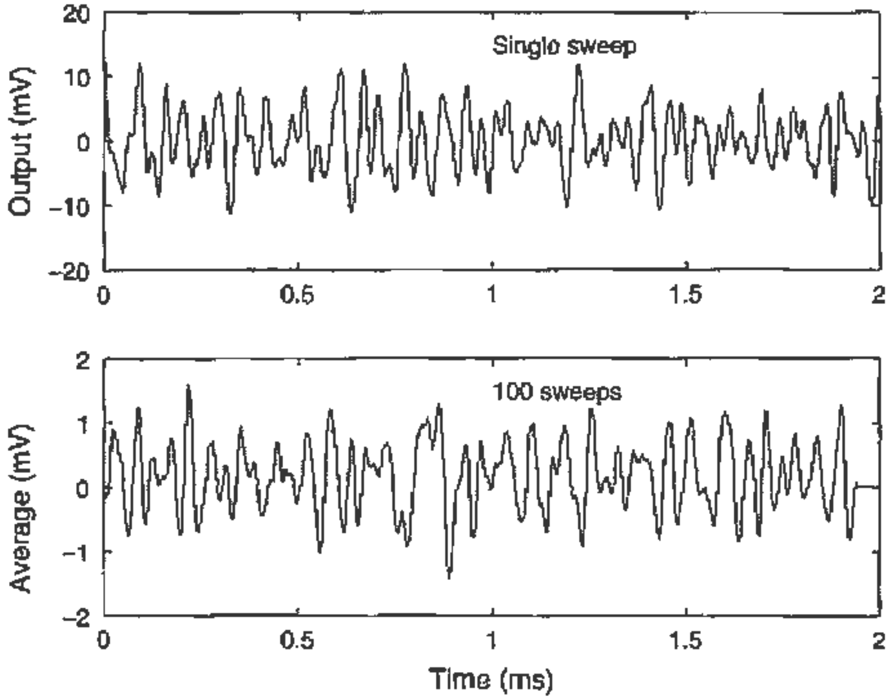


FIGURE 3.27 Oscilloscope traces of the output of the bandwidth filter, and for 100 traces averaged together by the oscilloscope. Note the difference in the vertical scales.

0.2 ms/div, then the time per sample is 0.2 μs since there are ten divisions. The bandwidth is the reciprocal of twice this time or 2.5 MHz. If the filter cuts off at 20 kHz, then this would be fine.

Use resistors with R near zero (10 Ω) and up to $R \approx 10\,\mathrm{k}\Omega$. The oscilloscope trace will look like an oscillatory signal, but that is because you are (likely) using tight bandwidth limits. What would the trace look like if the lower limit was only slightly smaller than the upper limit?

Figure 3.27 shows a single sweep trace on the scope directly from the output of the bandwidth filter, and the average (as done by the scope itself) of 100 traces. The average looks the "same" as the single sweep, but it is 10 times smaller. (Note the difference in the vertical scales.) It is clear, therefore, that the oscillations in the input signal are random in phase, even though they arc confined within the limits of the bandwidth filter. Most digital oscilloscopes have the ability to calculate and display for you the mean and variance of the trace. This will be useful for your analysis.

You need to determine the value of $G^2\Delta\nu = \int_0^\infty g^2(\nu)\,d\nu$. Make a plot of $g^2(\nu)$ as a function of ν and estimate the integral under the curve. You can try to estimate this graphically, but you can easily get an accurate answer using the MATLAB function trapz, which performs a trapezoidal

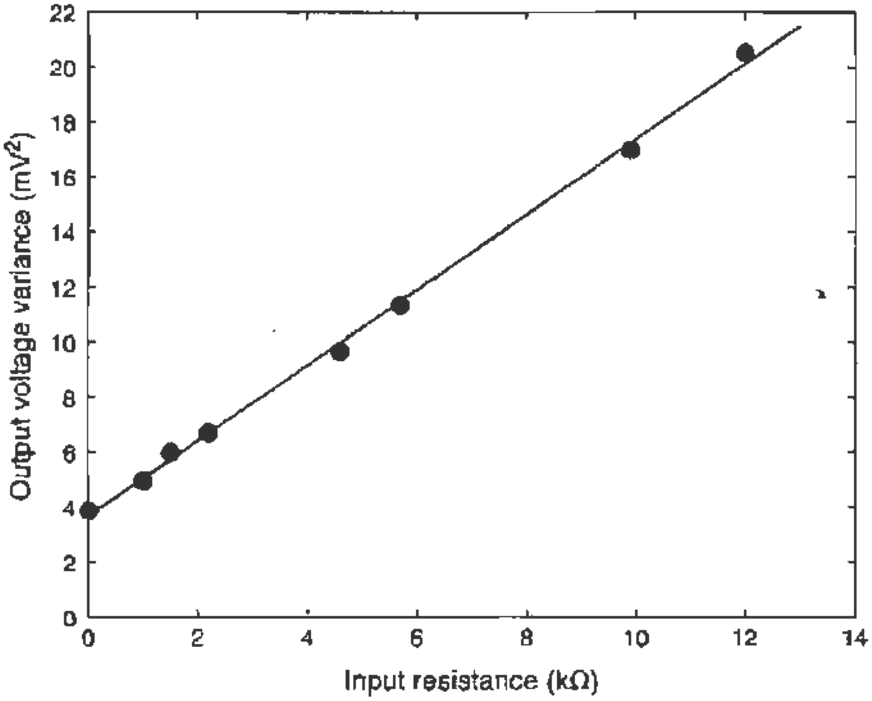


FIGURE 3.28 Data taken by measuring the standard deviation of the output voltage signal, as a function of the input resistor value. The slope gives k, while the intercept gives the equivalent input noise voltage, after correcting for the amplifier gain × bandwidth.

integration given a list of (x, y) values. For the data of Fig. 3.26 one finds that

$$G^2 \Delta\nu = (7.9 \pm 0.5) \times 10^7 \text{ kHz}.$$

Next we make a plot of $\langle (V - \langle V \rangle)^2 \rangle$ as a function of R. Note that since $\langle V \rangle = 0$, the above expression reduces to $\langle V^2 \rangle$. The plot is shown in Fig. 3.28 and a linear fit gives

$$\langle V^2 \rangle / R = (1.33 \pm 0.08) \text{ mV}^2/\text{k}\Omega$$

and an intercept at 4 mV^2.

We can now calculate Boltzmann's constant k from the above data using Eq. (3.20) and setting $T = 298$ K (room temperature). Using units of hertz, volts, and ohms, we write

$$k = \frac{\langle V^2 \rangle / R}{4TG^2\Delta\nu} = \frac{(1.33 \pm 0.08) \times 10^{-9}}{4 \times 298 \times (7.9 \pm 0.5) \times 10^{10}} = (1.42 \pm 0.13) \times 10^{-23} \text{ J/K}.$$

This result is in excellent agreement with the accepted value $k = 1.38 \times 10^{-23}$ J/K.

The intercept of the line in Fig. 3.28 is the noise at $R = 0$. You would expect this to be zero if Johnson noise in your input resistor were the only thing going on. The input opamp, however, has some noise of its own, due to internal Johnson noise, shot noise, and so on. The specification sheet for the HA5170 gives an equivalent input noisc of around 10 nV/$\sqrt{\text{Hz}}$. How does this compare to your measurement?

There are a number of variations and extensions to this experiment. For example, instead of simply using the oscilloscope to determine the standard deviation, use MATLAB and the trace data (as in Fig. 3.27) to get the values and examine their distribution. You can get the data into an array trace, and you can use mean(trace) and std(trace) to get the mean and standard deviation. The series of MATLAB commands used to plot the distribution might look like

```
bins = linspace(min(trace), max(trace), 50);
[n, x] = hist(trace, bins);
stairs(x, n);
```

The single sweep trace in Fig. 3.27 is plotted this way in Fig. 3.29. The distribution is rather Gaussian-like, as you expect, but you could test to

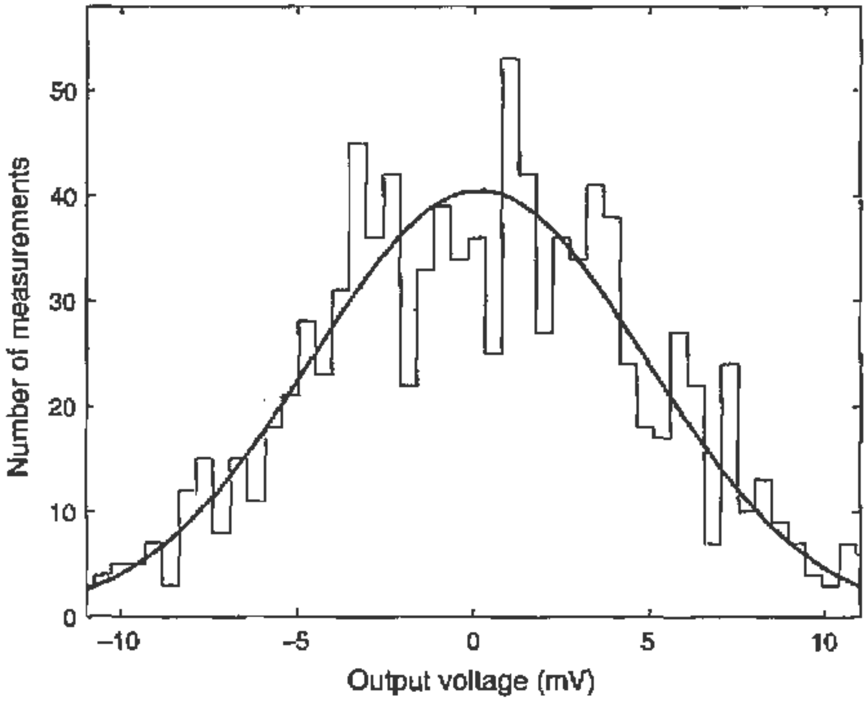


FIGURE 3.29 Histogram of the individual voltage values from a single sweep trace. The line is a Gaussian distribution, with the mean and standard deviation determined from the trace data, and normalized to the number of measurements.

see whether this is really the case by comparing it to the Gaussian with the same mean and standard deviation, and considering the χ^2. (See Chapter 10 for definitions and discussions of these quantities.) Some digital oscilloscopes have the capability of performing a real-time Fourier analysis of the input. That means that you can actually demonstrate that the noise spectrum $d\langle V^2\rangle/d\nu$ is indeed "white," that is, independent of frequency. This is straightforward data to take, but will require that you learn more about Fourier analysis to interpret it.

One nontrivial circuit modification would be to make your own bandwidth filter. For example, consider the circuit shown in Fig. 3.12.[7] Try assembling components that give you reasonable parameters for the gain integral in Eq. (3.21). A simpler kind of filter might simply be two RC filters, one high pass and one low pass, cascaded in series. If you want to do active buffering, though, be careful to use an opamp that works at these frequencies. Another interesting variation is to use a few-kiloohm resistor as input, but something mechanically large and strong enough to take some real temperature change. If you immerse the resistor in liquid nitrogen, for example, it should make a large (and predictable) change in the Johnson noise.

3.7. CHAOS

We now discuss a measurement that uses nonlinear electronic components to explore phenomena characteristic of complex physical systems.

3.7.1. The Logistic Map and Frequency Bifurcations

We are used to the notion that physical systems are described by differential equations that can be exactly solved for all times, given an appropriate set of initial conditions. This is not true in complex systems governed by nonlinear equations. A typical example is the flow of fluids. At low velocity one can identify individual "streamlines" and predict their evolution. However, when a particular combination of velocity, viscosity, and boundary dimensions is reached, *turbulence* sets in and eddies and vortices are formed. The motion becomes *chaotic*. Many chaotic systems exhibit self-similarity: that

[7] This, in fact, is what Johnson used in his 1928 paper. You might want to look it up, and compare your results to his.

is when the flow breaks into eddies, the eddies break into smaller eddies and so on. Such scaling is universal; it is observed in all chaotic systems..

A particularly simple case is that of systems that obey the *logistic map* introduced in connection with population growth. Designate by x_j the number of members of a group at the time j. Here the group may be the population on an island, the bacteria in a colony, etc. The index j labels a finite time interval (such as a day or a year) or the successive "generations" of the population. If the reproduction rate in one generation is λ, then it would hold that

$$x_{j+1} = \lambda x_j.$$

However, the population will also decrease due to deaths. In particular if the food supply on the island is finite the death rate will be proportional to x_j^2. Thus the evolution[8] is governed by the map

$$x_{j+1} = \lambda x_j - s x_j^2. \tag{3.22}$$

We use the term *map*, because given x_j we can find x_{j+1} uniquely. Both λ and s are assumed nonnegative. We see immediately that if $\lambda > 1$ and $s = 0$ the population will grow exponentially, while if $\lambda < 1$ the population will tend to 0. The map of Eq. (3.22) can be *rescaled* by introducing

$$y_j = \frac{s}{\lambda} x_j \qquad \text{for all } j.$$

Then y_j obeys the logistic map

$$y_{j+1} = \lambda y_j (1 - y_j). \tag{3.23}$$

The above map has the interesting property that if the reproduction rate for one generation is restricted in the range

$$0 < \lambda < 4,$$

then y_j remains bounded between

$$0 < y_j < 1.$$

[8] The first study of these issues is due to the English sociologist T. R. Malthus (1766–1834).

We are interested in the fate of the group after many generations, namely in the value of y_j as $j \to \infty$. We find, as already stated, that:

If $\lambda \leq 1$, as $j \to \infty$ $y_j \to 0$, the population decays to 0.
If $1 < \lambda < 3$, as $j \to \infty$ $y_j \to y \to y^*$ the population tends to a *stable point* y^*, namely

$$y^* = \lambda y^*(1 - y^*) \tag{3.24}$$

with solutions

$$y^* = 0 \qquad y^* = \left(1 - \frac{1}{\lambda}\right).$$

In this case the solution $y^* = 0$ is unstable, because if $y_0 = \epsilon$ (ϵ infinitesimal) y_∞ will tend to $(1 - 1/\lambda)$.

When $\lambda > 3$ the system behaves in a very different manner. As soon as $\lambda > 3$ but $\lambda < 3.4495\ldots$ the population alternates between 2 stable values. When $\lambda > 3.4495\ldots$ the population alternates between 4 stable values until $\lambda > 3.54\ldots$, where it alternates between 8 stable values; for $\lambda > 3.56\ldots$ the population alternates between 16 stable values, and this continues at ever more closely spaced intervals of λ. We say that there is a *bifurcation*[9] at these specific values of λ. These results can be easily checked with a pocket calculator or a simple program. Table 3.1 gives some typical results for $\lambda = 2.8$, $\lambda = 3.2$, and $\lambda = 3.5$, and the stable points are shown in the graphical construction of Fig. 3.30.

What is plotted in Fig. 3.30 is y_{final} vs y_{initial}. The continuous curve is the equation of the logistic map $y_f = \lambda y_i(1 - y_i)$. In Fig. 3.30a the curves

TABLE 3.1 Example of Stable Points of the Logistic Map

$\lambda = 2.8$	$y^* = 0.6429\ldots$
$\lambda = 3.2$	$y^* = 0.5310\ldots$
	$= 0.7995\ldots$
$\lambda = 3.5$	$y^* = 0.3828\ldots$
	$= 0.5009\ldots$
	$= 0.8269\ldots$
	$= 0.8750\ldots$

[9] Henri Poincare in 1900 had noticed such behavior in mechanical systems and named it the "exchange of stability."

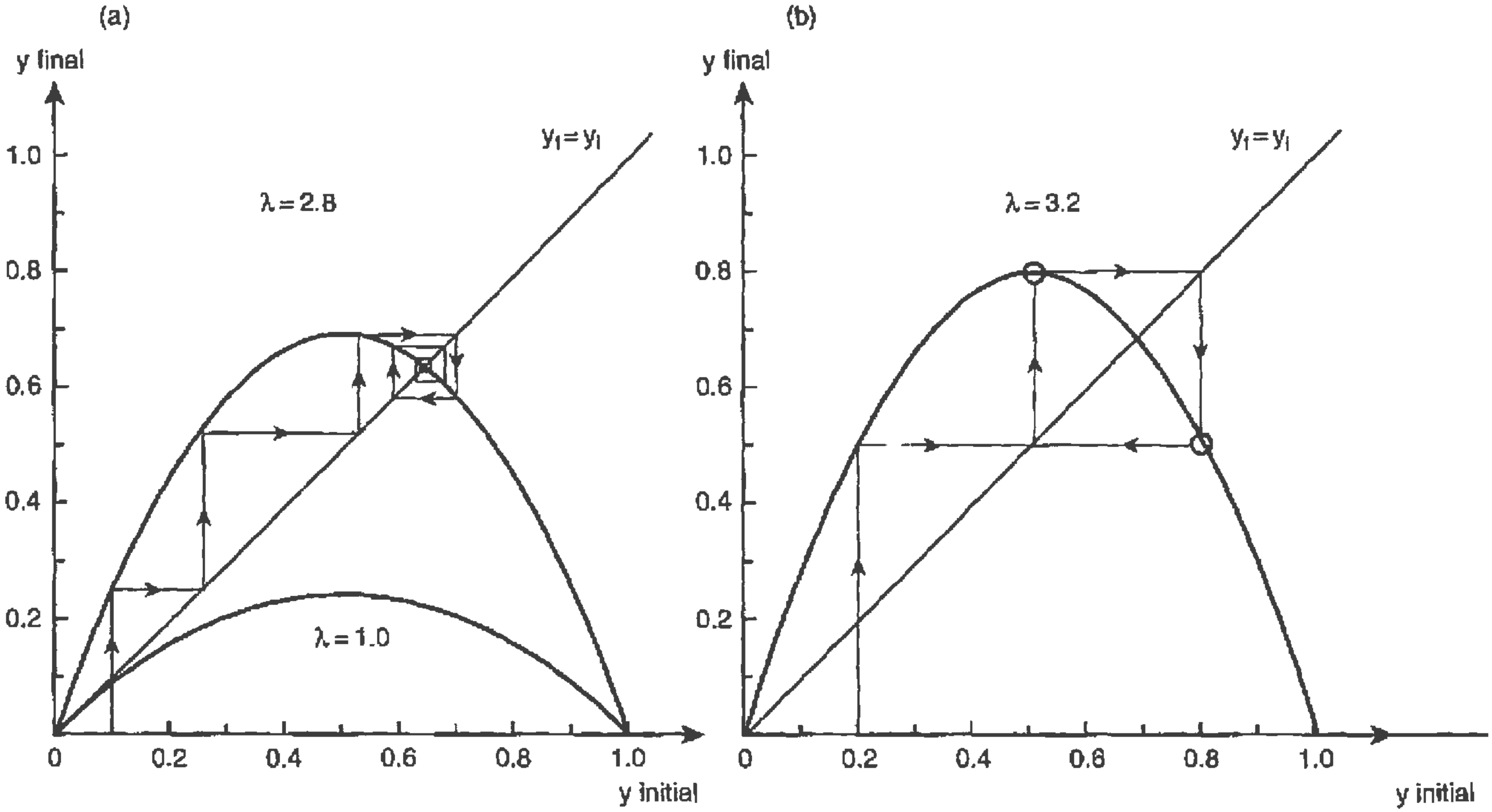


FIGURE 3.30 Plots of the logistic map: (a) for $\lambda = 1.0$ and $\lambda = 2.8$; for $\lambda = 2.8$ there is one stable point at $y^* = 0.6429\ldots$. (b) For $\lambda = 3.2$; there are now two stable points at $y^* = 0.7995\ldots$ and $y^* = 0.5130$. See the text for details of the path leading to the stable points.

for $\lambda = 2.8$ and $\lambda = 1.0$ are shown, while in Fig. 3.30b the curve for $\lambda = 3.2$. The lines for $y_f = y_i$ are also drawn. We can follow the path from some initial value $y_0 = 0.1$ in Fig. 3.30a to the stable point (indicated by a circle). Given y_0 we find $y_1 = y_f$ at the intersection with the curve. However, y_1 must now be used as an input, y_i, so we use the $y_f = y_i$ line to locate y_i and proceed to find y_2 and so on. The process converges to the circled point at $y^* = 0.6429\ldots$.

It is also evident that the same construction for the $\lambda = 1$'curve will lead to $y^* = 0.0$. In Fig. 3.30b we start (for more rapid convergence) from $y_0 = 0.2$. We now find the two stable points at $y^* = 0.7995$ and $y^* = 0.5130$. The map requires that one stable point leads to the next and vice versa.

When $\lambda > 3.5699\ldots$ the population no longer reaches a stable point but takes on an infinity of values in the range $0 < y_\infty < 1$. We say that the system behaves *chaotically*. This persists in the remainder of the range $3.5699 \cdots < \lambda < 4.0$, but one finds regions of stability where an odd number of stable points exist. The dependence of the bifurcations on λ is shown in Fig. 3.31 where the λ-scale is highly nonlinear in order to show enough detail; the vertical scale gives the values $y_j^*(j \to \infty)$ of the stable points.

The remarkable discovery by M. Feigenbaum in 1975 was that all systems that exhibit chaos follow the same (*universal*) behavior and that the difference $\Delta_n = \lambda_{n+1} - \lambda_n$ of the values of the *parameter* at which bifurcations (period doubling) occur converges rapidly as $n \to \infty$. In particular as $n \to \infty$ the ratio

$$\frac{\lambda_{n+1} - \lambda_n}{\lambda_{n+2} - \lambda_{n+1}} \to \delta = 4.669201660910\ldots \tag{3.25}$$

tends to the universal number δ. For instance in our previous example λ_n is the value of the reproduction rate λ at the nth bifurcation.

However, also the amplitude of the population at the stable points exhibits universal behavior and scales according to a different universal number α. Let $y_n^{*(1)}$ and $y_n^{*(2)}$ be two stable points of a given branch at the bifurcation value λ_n. We define

$$\Delta y_n^* = y_n^{*(1)} - y_n^{*(2)},$$

and it holds that as $n \to \infty$

$$\frac{\Delta y_n^*}{\Delta y_{n+1}^*} \to \alpha = 2.5029078\ldots. \tag{3.26}$$

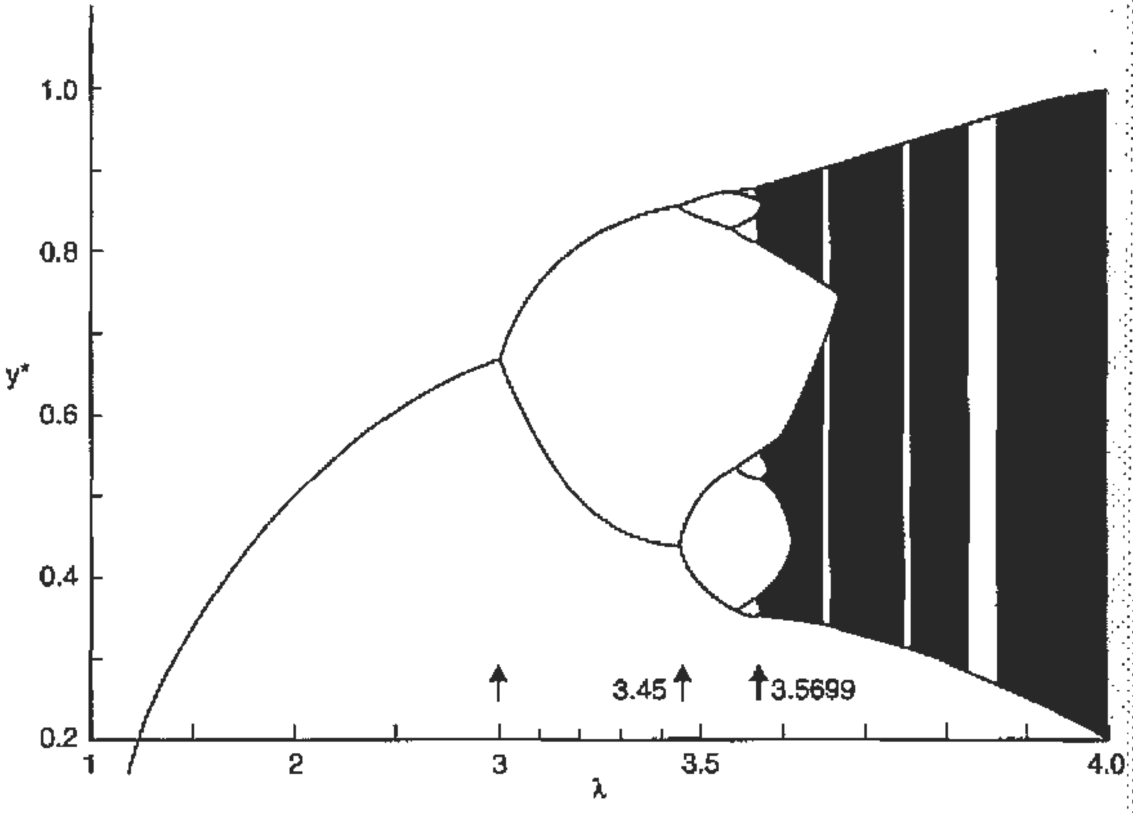


FIGURE 3.31 The stable points of the logistic map as a function of λ. The λ-scale is highly nonlinear in order to clearly show the bifurcations. The black parts of the plot indicate the chaotic region. Note, however, the thin white lines, which indicate islands of stability.

This indicates that as λ increases the system replicates itself after rescaling by a factor $1/\alpha$, as shown in Fig. 3.31; typical intervals[10] Δy_2^* and Δy_3^* are indicated.

The numbers δ and α are named after Feigenbaum and are obtained by numerical calculations of maps or equations that lead to chaotic behavior. They are always found to be the same for all problems. We will verify this to the accuracy that can be reached in the experiment described below.

3.7.2. The Diode–R–L Circuit

A simple R–L–C circuit where the capacitor is replaced by a diode, driven at its resonant frequency, exhibits bifurcations and eventually chaotic behavior. This is not so surprising because the diode is a nonlinear device. The

[10] The intervals Δy_n^* must be chosen appropriately as is also evident from Fig. 3.31.

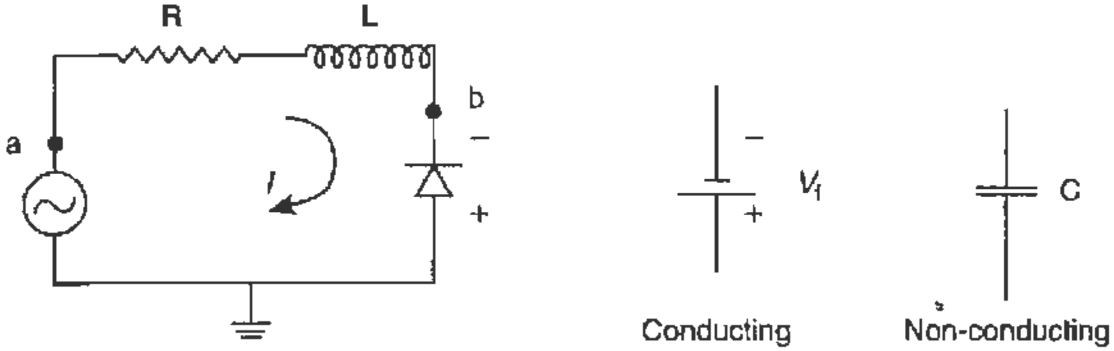


FIGURE 3.32 The diode–R–L circuit. The equivalent behavior of the diode in its two states.

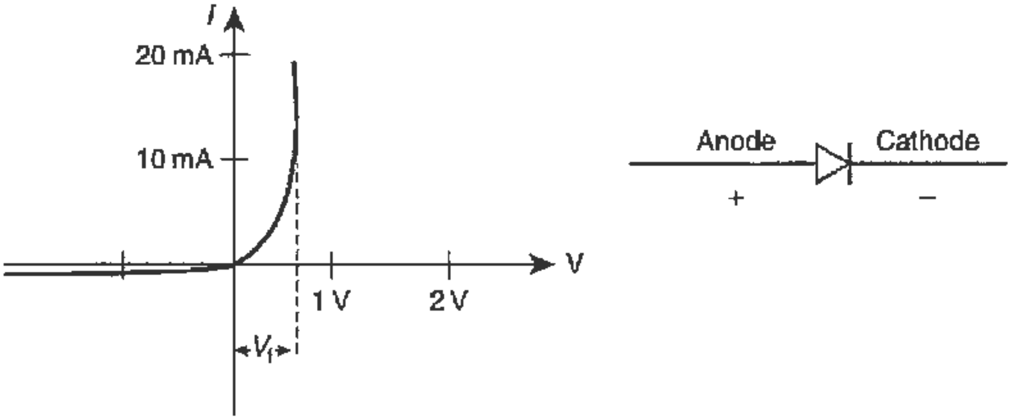


FIGURE 3.33 The *I–V* characteristic of a diode.

effect was first reported by Linsay[11] and was analyzed in detail by Rollins and Hunt.[12]

The circuit is shown in Fig. 3.32, and the *I–V* curve of the diode in Fig. 3.33. When there is a positive voltage across the diode it conducts and appears as an EMF of magnitude $-V_f$, i.e., as a voltage drop. In its nonconducting state the diode behaves as a capacitor C and will draw a charging current. These two states are shown schematically in Fig. 3.32 where we also indicate our convention for positive current flow.

The source is assumed sinusoidal of amplitude V_0, so that the voltage at point a of the circuit is

$$V_0 = V_0 \cos \omega t. \tag{3.27}$$

[11] P. S. Linsay, *Phys. Rev. Lett.* **47**, 1349 (1981).

[12] R. W. Rollins and E. R. Hunt, *Phys. Rev. Lett.* **49**, 1295 (1982); R. W. Rollins and E. R. Hunt, *Phys. Rev. A* **29**, 3327 (1984).

current in the circuit and the voltage at point b (i.e., across the diode) be calculated straightforwardly from the discussion in Section 3.1.[13] In conducting state, obviously $V_b = -V_f$, whereas in the nonconducting ͻ the voltage will follow the frequency of the source. However, the ›litude need not be the same for every cycle. This happens because the le does not stop conducting as soon as the current goes to zero but has emory; it continues to conduct for a time interval

$$\tau_r = \tau_m\left(1 - e^{-|I_m|/I_c}\right). \tag{3.28}$$

his expression $|I_m|$ is the maximum current during the current cycle; τ_m I_c are constants. If $|I_m|$ is zero then the recovery time τ_r is also zero. $I_m| \gg I_c$ then $\tau_r = \tau_m$. Thus the maximum current in the following le depends on the value of I_m in the previous cycle in a *noninvertible* ıion. We have a mapping

$$|I_m|_n \Longrightarrow |I_m|_{n+1}.$$

ͻ behavior of the current and voltage are shown in Fig. 3.34.

Γhe period of the source $T_0 = 2\pi/\omega_0$ defines the cycles or generations he system. The source voltage sets the reproduction parameter through $= V_0/V_f$. The voltage across the diode V_b (in the nonconducting state) elated to the population y_j in the jth cycle. Depending on λ, the voltage repeats with the period, T_0, of the source, or with period $2T_0$, $4T_0$, and on until V_b becomes completely chaotic. A numerical analysis of the ›ve circuit is given in the papers of Rollins and Hunt.

[3] In the conducting state we find that

$$\begin{aligned} I(t) &= \left(V_0/\sqrt{R^2 + L^2\omega^2}\right)\cos(\omega t - \theta_a) + Ae^{-(R/L)t} + \tfrac{V_f}{R} \\ V_b(t) &= -V_f. \end{aligned}$$

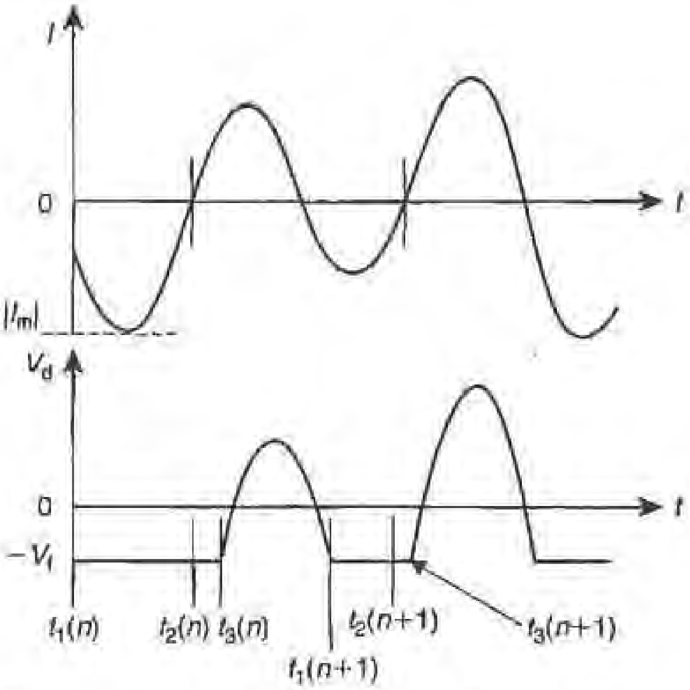
ıe nonconducting state we find that

$$\begin{aligned} (t) &= V_0/\sqrt{R^2 + L^2(\omega^2 - \omega_0^2)^2/\omega^2}\cos(\omega t - \theta_b) + Be^{-(2R/L)t}\cos(\omega_b t + \phi) \\ {}_b(t) &= V_0 \cos\omega t - I(t)R - L(dI/dt) \end{aligned}$$

ı

$$\omega_0 = 1/\sqrt{LC} \qquad \omega_b^2 = \omega_0^2 - (R/2L)^2$$

A, B, ϕ are constants.



GURE 3.34 The current and voltage in the diode–R–L circuit shown as a function of ne.

7.3. Experimental Results

e circuit is set up as shown in Fig. 3.32. A Hewlett–Packard function nerator HP3325 is used to drive the circuit. A fairly hefty variable inducnce ($L = 10$ mH) is used since the diode capacity is small. The series sistance was $R \simeq 50\ \Omega$. The diode should not be too slow (such as are recier diodes) nor too fast. Good results were obtained with a 1N4007 diode; her diodes, namely 1N4001 and 1N5404, gave qualitatively similar (but antitatively different) results.

The first step is to tune the inductor to find the resonant frequency of the rcuit. In this case it was found that $\omega_0/2\pi = 71.5$ kHz $\simeq 1/(2\pi\sqrt{LC})$.

Figs. 3.35a–3.35d are shown the voltage across the diode V_b and the iving voltage V_0. For $V_0 < 0.875$ V, V_b has the same periodicity as . However just above $V_0 = 0.875$ V, V_b alternates between two difrent values as shown in Fig. 3.35a. The effect is clear, but not very onounced, because the data have been taken only slightly above the first furcation. Figure 3.35b corresponds to $V_0 = 2.033$ V where the secd bifurcation sets in. The period of V_b is now four times that of V_0. gain the difference between the two high-level states is very small and

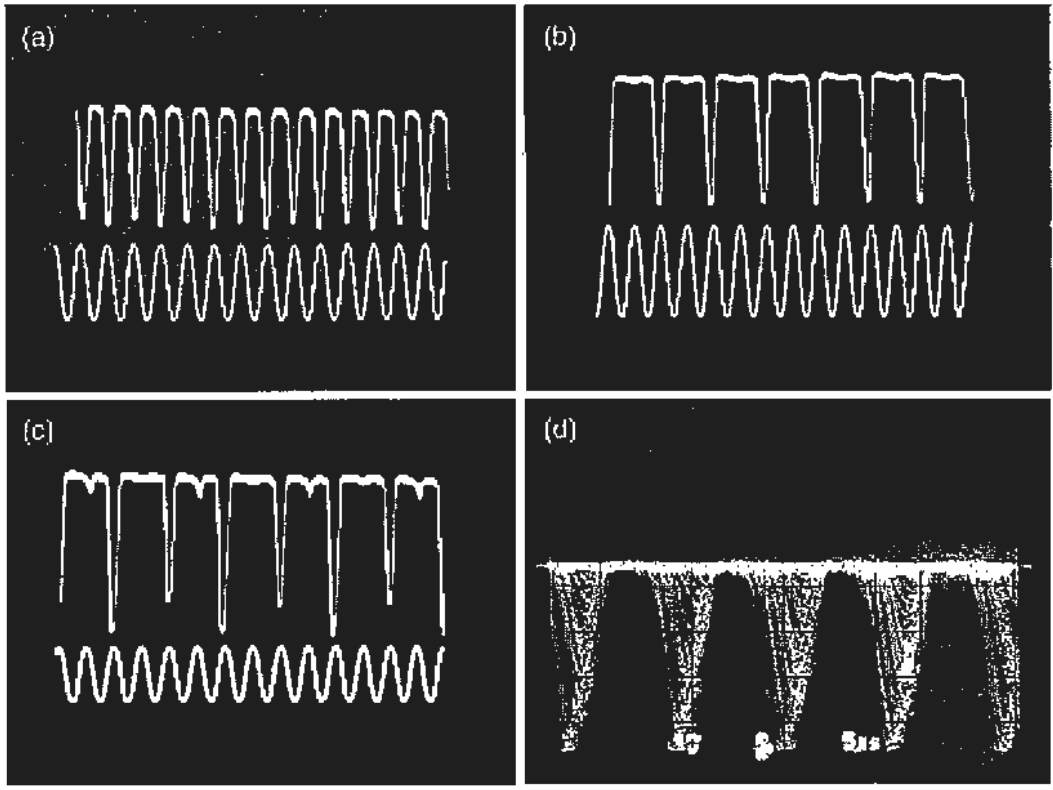


FIGURE 3.35 Oscilloscope traces of the voltage, V_b, across the diode (upper trace) and of the driving voltage V_0 (lower trace). The driving frequency is 71.5 kHz. (a) Immediately after the first bifurcation. Note that the upper trace is bimodal and has period $2T_0$. (b) Immediately after the second bifurcation. Note that the large peaks are bimodal; the period is $4T_0$. (c) Immediately after the third bifurcation; the period is now $8T_0$. (d) Chaotic behavior.

that between the two low-level states is not observable. The next scope traces, Fig. 3.35c, correspond to $V_0 = 2.280$ V and were taken right after the third bifurcation. The period of V_b is now eight times that of V_0 and similar comments apply as to the distinguishability of the different states. A fourth bifurcation was observed at $V_0 = 2.340$ V. Finally Fig. 3.35d shows V_b when $V_0 \gtrsim 2.355$ V where chaos was observed to set in.

A plot of the bifurcations obtained for this diode is shown in Fig. 3.36. The error in determining the exact bifurcation voltage[14] is ±5 mV. We summarize the results in Table 3.2. From these data we calculate the

[14] A more precise determination of the voltage at which bifurcation occurs can be made when a signal analyzer (FFT) is available. In this case the onset of period doubling is evident from the appearance of subharmonics in the frequency spectrum.

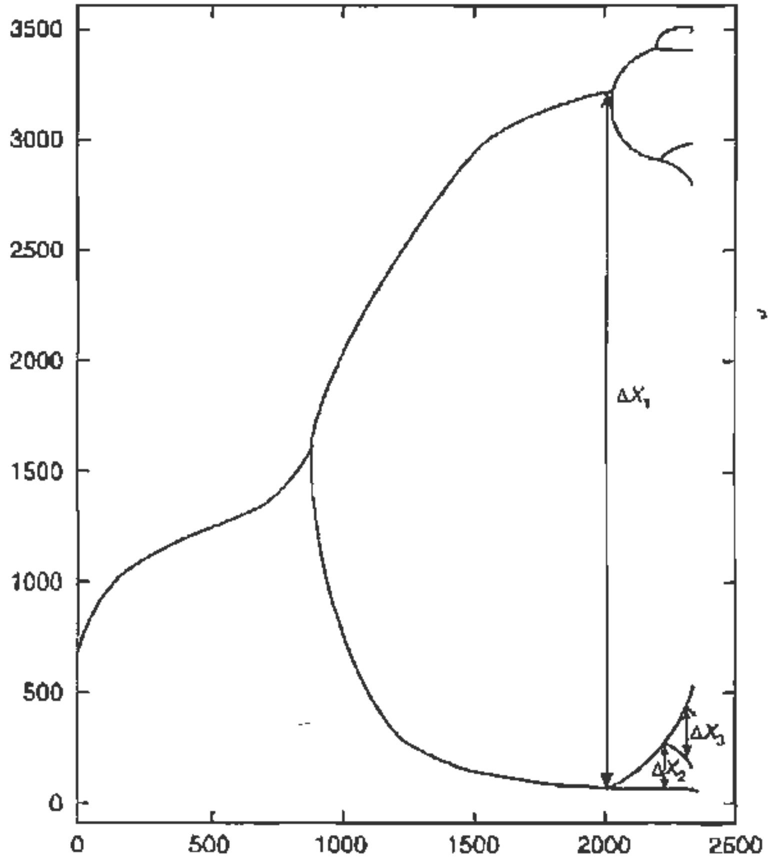


FIGURE 3.36 Plot of V_b vs V_0 as measured for the 1N4007 diode. The bifurcations are clearly observed. Some ΔV_b spacings are also indicated. Chaos sets in at $V_0 = 2.355$ V.

TABLE 3.2 Bifurcation Data from Measurements of Chaos

Bifurcation	V_0 (mV)
1st	875
2nd	2033
3rd	2280
4th	2340
Chaos	2355

Feigenbaum number δ. We have

$$\lambda_2 - \lambda_1 = 1158 \pm 7 \text{ mV}$$
$$\lambda_3 - \lambda_2 = 247 \pm 7 \text{ mV}$$
$$\lambda_4 - \lambda_3 = 60 \pm 7 \text{ mV}$$

and therefore

$$\delta^{(1)} = \frac{\lambda_2 - \lambda_1}{\lambda_3 - \lambda_2} = 4.688 \pm 0.13$$

$$\delta^{(2)} = \frac{\lambda_3 - \lambda_2}{\lambda_4 - \lambda_3} = 4.117 \pm 0.49.$$

These results are consistent with the asymptotic value given in Eq. (3.25), even though input from only the first four bifurcations was used.

The determination of the second Feigenbaum number α is not possible with the present data. As pointed out previously, the intervals Δy_n^* must be selected appropriately, but even then (see Fig. 3.36) the ratios of Δy_n^* seem much larger than α. This is due in part to the fact that one has not reached the asymptotic regime of Eq. (3.26) and in part to discontinuous jumps in V_b at certain values[15] of V_0. However, it is evident from the data that the system replicates itself after each bifurcation. Furthermore, the spacing between stable points in every branch decreases in subsequent bifurcations by a multiplicative factor; this factor seems to converge toward Feigenbaum's α. We also note that for the 1N4001 diode it was possible to observe islands of stability in the chaotic region.

3.8. LOCK-IN DETECTION

Suppose one is studying a signal, amid noise, that comes at a specific frequency. We can use this to pick the signal out of the noise. Furthermore, we can be sensitive to the *phase* of the signal as well as its frequency, and that can make a huge improvement. The technique that does all this is called *phase-sensitive detection*. Tbe device that you do it with is called a *lock-in amplifier*.

There are two inputs to a lock-in amplifier. One input carries the signal (and the noise). The signal, remember, is varying at some specific frequency which you are aware of. It may be completely buried in noise, however, so you would not see it on an oscilloscope, for example. The other input carries a reference that varies at the frequency of the signal. The signal oscillates because you make it do so, and the way you do that also gives you the reference. For example, your experiment measures a response to a

[15] Some diodes show marked hysteresis associated with these discontinuities.

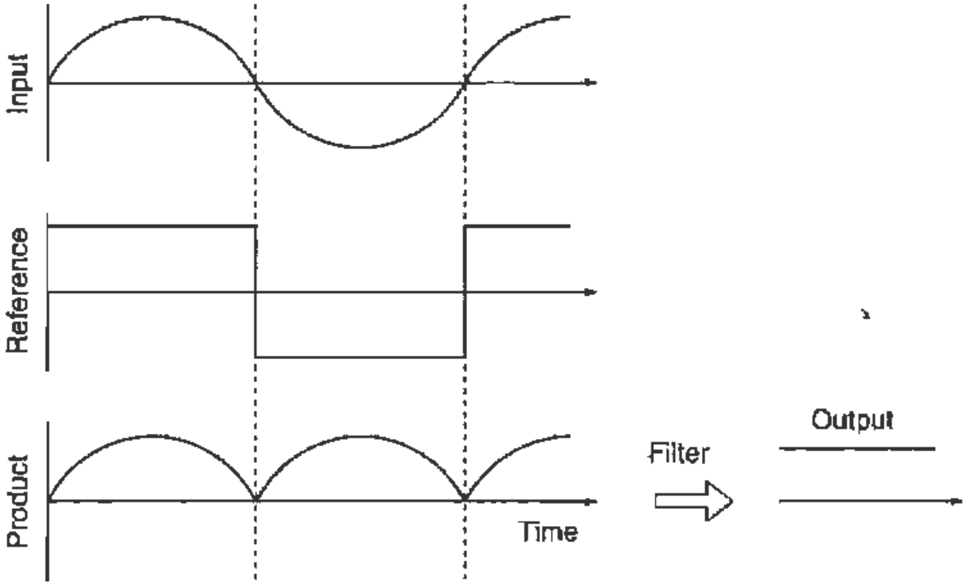


FIGURE 3.37 The lock-in amplifier acting on an in-phase signal.

laser, so you turn the laser on and off rapidly with a mechanical chopper. The motor drive for the chopper gives you the reference signal.

The lock-in amplifier takes the reference signal and uses it as a switch. For half the period, the switch is "up," and it lets the signal input pass through it with no change. For the other half, the switch is "down," and it reverses the sign of the signal (i.e., multiplies it by -1) before it passes. This is shown in Fig. 3.37. The result of this is a modified signal that is always positive, instead of oscillating around 0 like the input signal. A low-pass filter takes out the remaining oscillation and lets the DC level pass through. This DC level is read off a meter, presented at some output connector, or digitized by some computer, depending on the lock-in amplifier.

Now consider what happens if the signal is *out of phase* by 90° with respect to the reference. This situation is shown in Fig. 3.38. Now the output of the multiply stage is still something that oscillates about 0. The average DC level is 0, and that is the output of the lock-in amplifier. So, as promised, the lock-in amplifier only detects signals that are *in phase* with the reference. Most lock-ins have a "phase adjustment" knob on the front that allows you to maximize the output signal. If you have the phase 180° away, the output signal should reverse sign.

Now consider what the lock-in amplifier does to noise that has some frequency other than the frequency of the signal. The answer is obvious. The output of the multiply stage will just be a jumble of noise like the input stage since the reference is essentially just randomly flipping amplitudes.

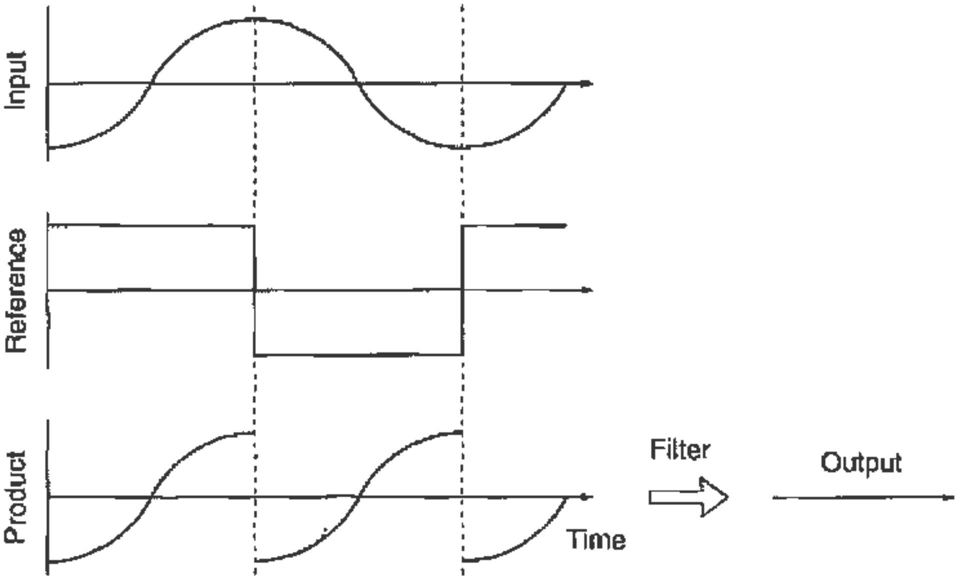


FIGURE 3.38 The lock-in amplifier acting on an out-of-phase signal.

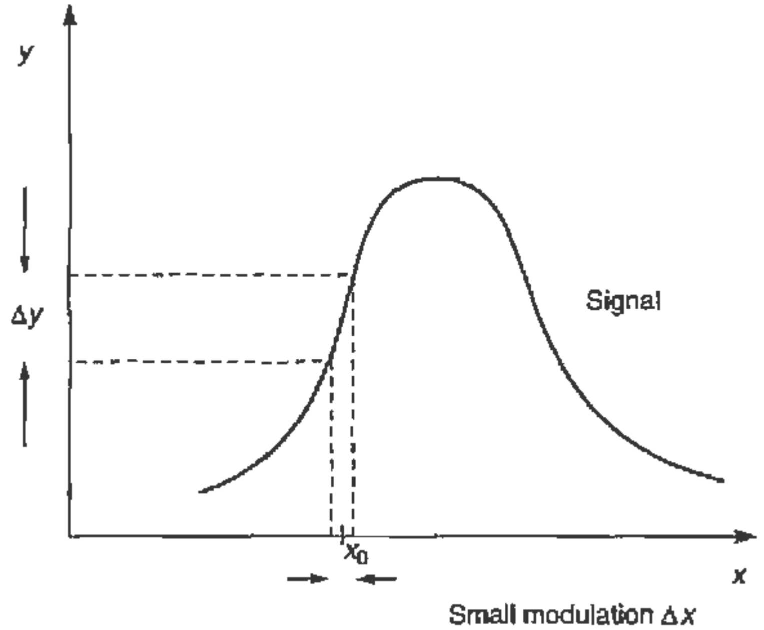


FIGURE 3.39 Using a lock-in amplifier for modulation spectroscopy.

The output of the low-pass filter will *average* to 0 over some time determined by the RC time constant of the filter.

The lock-in amplifier is actually quite a versatile instrument. One of its uses beyond noise rejection is as a spectroscopy tool. Let's say you have a signal y that is a function of some parameter x. For example, you might have an NMR signal as a function of the large magnetic field that polarizes the sample. Such a thing is graphed in Fig. 3.39. Now assume the signal is modulated (i.e., made to oscillate) by setting x to some central value x_0 and making it oscillate about x_0 by a small amount Δx. Then the amplitude Δy of the modulated signal is given by

$$\Delta y = \left.\frac{dy}{dx}\right|_{x_0} \Delta x.$$

In other words, the output of the lock-in is the derivative of the line shape $y(x)$. It does this, of course, while throwing out any noise that gets in its way. One common technique, described in detail by Dunlap (1988), is to sweep the value of x many times and record the output in a multichannel analyzer. This uses signal averaging to get rid of any remaining noise.

3.9. COMPUTER INTERFACES

Many of the experiments described in this book, as well as in many undergraduate instructional laboratories, can be done without the use of sophisticated computerized data acquisition. Indeed, in experiments such as the Balmer series in hydrogen (Section 1.5.3), the Faraday effect (Section 5.7), and the γ–γ angular correlation in ^{60}Co (Section 9.5.4), for example, there is much instructional value in taking, recording, and analyzing data "by hand."

Nevertheless, directly interfacing a computer with the experiment makes it possible to take data much more quickly in many cases, and this also has much instructional value. Furthermore, some experiments that had once been very difficult, if not impossible, in the instructional laboratory, can now be done with relatively simple and inexpensive computer interfaces. A wide variety of commercial interfaces exist, and it is not possible to cover all of them in this textbook. Indeed, the market moves quickly and different options appear and disappear very regularly. A recent publication, available free from Keithley at http://www.keithley.com/, is the "Data Acquisition and Control Handbook." However, a number of standard situations apply.

The simplest computer interface is a "serial" interface using an RS232 standard communications port on the computer. The electronics on your computer and in the data acquisition device to which you wish to interface support a standard "handshake" protocol for moving instructions and data back and forth between the two devices. All that you need is a four-wire telephone cable to connect the two, and software to make it work. This software is very often available using a free download from the vendor of the data acquisition device. For example, for their line of digital oscilloscopes, LeCroy Corporation (http://www.lecroy.com/) provides a program called ScopeExplorer for this purpose. There are many other examples.

It is a good idea to consider "middleman" computer interfaces, so that your computer and software can talk to one specific device, and then this device can be connected to any number of other instruments that acquire different kinds of data. This cuts down on the "interface programs" that must run on your computer and with which you need to become familiar, and gives you more flexibility for your experiments, at the cost of a bit more expense. For example, Vernier Software & Technology (http://www.vernier.com/) sells the "Universal Laboratory Interface" (ULI), a serial computer interface that then connects to experiments through a variety of analog and digital inputs for measuring voltages, currents, scaler counts, and so forth. The company also sells inexpensive computer programs for controlling the ULI from any number of a variety of computers.

Serial interfaces are simple, but they are slow. They transfer data one bit at a time ("serial"), and the number of bits per second (the "baud" rate) is limited by the simple cabling and connection standards to some 56,000 bits per second (56 kbits). This is fast enough for many applications, but the experimenter can quickly be needing (or wishing for) a higher data rate.

Faster data rates are provided by parallel interfaces, where many lines connect the computer to the data acquisition apparatus, or possibly through the network connections to the computer using an ethernet connection and TCP/IP protocol. At this point, the number of hardware and software options increases enormously, including interfaces designed and built in the laboratory itself. Some companies that sell such interfaces and software include Agilent Technologies (http://agilent.com/), Keithley Instruments (http://www.keithley.com/), and National Instruments (http://www.ni.com/), among others. LabVIEW from National Instruments is a very popular software tool for laboratory interfaces which features a graphical programming environment, but which can be difficult to use in an undergraduate laboratory setting without the necessary support.

Probably the most popular standard parallel interface is GPIB or "General Purpose Interface Bus." Also known as the IEEE-488 standard, or as

HPIB by people at Hewlett–Packard Corporation (now Agilent Technologies), GPIB uses an ASCII code to communicate, very similar to most serial line communication systems, but uses a 24-pin connector, allowing data to be transferred in parallel at some level. It can transmit up to 1 MByte per second, within this communication protocol. In order to communicate with a data acquisition device equipped with a GPIB port, some sort of computer port is also necessary, generally provided using a plug-in card, available from several manufacturers depending on the type of computer. Virtually all commercial general use data acquisition software packages provide for communication through GPIB, including ScopeExplorer and LabVIEW.

One thing to keep in mind is that the next step after data acquisition is data analysis. Depending on what software you may use for analyzing your data, you should try to acquire the data in a way that is amenable to your analysis tools. Once again, this can be solved with commercial products if you have the resources. In this book, for example, we use MATLAB for data analysis, and it is possible to purchase from The Mathworks (http://www.mathworks.com/) toolboxes for MATLAB for instrument control and for data acquisition, although we are not making use of these specialized toolboxes in this book.

Depending on the local expertise and available resources, the variety of computer interfaces can become quite large and complicated. We will use a number of different options for the experiments in this book[16] including

- a LeCroy Digital oscilloscope and ScopeExplorer to measure the decays of eddy currents in metals (Section 2.2).
- a plug-in board for control and voltage readout, operated with LabVIEW, for a high-resolution optical monochromator (Section 6.3.3).
- a Vernier ULI and LoggerPro software to count and record Geiger counter signals to measure nuclear decay rates (Section 8.6).
- a Canberra multichannel analyzer and a GPIB interface to measure gamma ray spectra, including an experiment on Compton scattering (Sections 8.4 and 9.2).
- a home-built time-to-analog measurement system for determining the mean life of the muon (Section 9.4.3).

[16] The reader should be aware that it is unlikely (and unnecessary) that these options be duplicated exactly in your own laboratory.

3.10. REFERENCES

An excellent, popular, and up-to-date text and reference book on electronics is:

P. Horowitz and W. Hill, *The Art of Electronics*, second ed., Cambridge Univ. Press, Cambridge, UK, 1989.

A student manual for this book is also available. A good book with introductory chapters on solid-state electronics, including the physics behind diodes and transistors, is

R. A. Dunlap, *Experimental Physics: Modern Methods*, Oxford Univ. Press, Oxford, 1988.

Some good articles that discuss the physics and experimentation of thermal noise can be found in

R. W. Henry, Random walk model of thermal noise for students in elementary physics, *Am. J. Phys.* **41**, 1361 (1973).

P. Kittel, W. R. Hackerman, and R. J. Donnelly, Undergraduate experiment in noise thermometry, *Am. J. Phys.* **46**, 94 (1978).

D. L. Livesey and D. L. McLeod, An experiment on electronic noise in the freshman laboratory, *Am. J. Phys.* **41**, 1364 (1973).

There exists by now an extensive literature on chaos. Some suggested references are:

M. J. Feigenbaum, *J. Stat. Phys.* **19**, 25 (1978); **21**, 669 (1979).

L. P. Kadanoff, *Phys. Today* **36** (12), 46 (Dec. 1983).

G. L. Baker and J. P. Gollub, *Chaotic Dynamics: An Introduction*, Cambridge Univ. Press, Cambridge, UK, 1990.

H. Nagashima and Y. Baba, *Introduction to Chaos*, English transl., Institute of Physics, Bristol, 1999.

CHAPTER 4

Lasers

Lasers are a source of intense, highly monochromatic, coherent beams of light in the visible and infrared parts of the spectrum. "Laser" is an acronym for "light amplification by stimulated emission of radiation" introduced in 1958 by its inventors, C. H. Townes and A. L. Schawlow. The first successful operation of a laser was achieved by Maiman in 1960 using a ruby rod as the lasing material. Fundamentally a laser consists of the lasing medium, of means for exciting the medium either through an electrical discharge or by an external light source, and of an optical cavity made out of a pair of high-reflectivity mirrors. The lasing medium can be a gas, a transparent solid, a liquid (dye), or a semiconductor. The relative simplicity and low cost of lasers have contributed to a wide range of applications. Semiconductor lasers, also referred to as diode lasers, can be found in almost every modern piece of equipment.

Because of the coherence, intensity, and monochromaticity of the light emitted by a laser, such beams are ideal for demonstrating the properties of light and of optical elements. Experiments that without a laser were tedious and required great skill can now be performed routinely. We begin with a brief discussion of the laser equations and a description of the HeNe laser. As the first application we show how a laser beam can be expanded with a pair of lenses and how to measure its spatial profile. We then discuss the two most familiar types of interferometers, the Michelson and the Fabry–Perot. We demonstrate how they can be used to measure the wavelength of the HeNe line. Before using a laser the reader should carefully consult Appendix C on *laser safety*.

4.1. THE PRINCIPLE OF LASER OPERATION

Light is emitted when an atom (or molecule) makes a transition from an excited state to a state of lower energy. The frequency of the light is given by

$$\nu = \frac{\omega}{2\pi} = \frac{1}{2\pi}\,\frac{E_2 - E_1}{\hbar}, \tag{4.1}$$

where $\Delta E = E_2 - E_1$ is the energy difference between the upper and lower states (also referred to as levels) involved in the transition. Here $\hbar$ is Planck's constant divided by 2π

$$\hbar = 1.054 \times 10^{-34}\ \text{J-s}, \tag{4.2}$$

and therefore

$$\Delta E = \hbar\omega. \tag{4.3}$$

The reason for the factors of 2π is that they simplify the writing of the equations.

The transition from an upper state to a state of lower energy will occur *spontaneously* and we designate by A the probability per unit time for such an occurrence. However, the transition can also be *stimulated* (induced) by the presence of light of angular frequency ω, satisfying Eq. (4.3). We designate by B the probability, per unit time per unit energy density in unit frequency interval, for stimulated emission. It is reasonable that the presence of the electromagnetic (EM) field (light) at the resonant frequency will not only induce transitions from $2 \rightarrow 1$ but also from $1 \rightarrow 2$ with the same probability. It is equally reasonable that the photons arising from stimulated emission will have exactly the *same frequency* and *same direction* as those in the incident EM wave. These arguments were first proposed

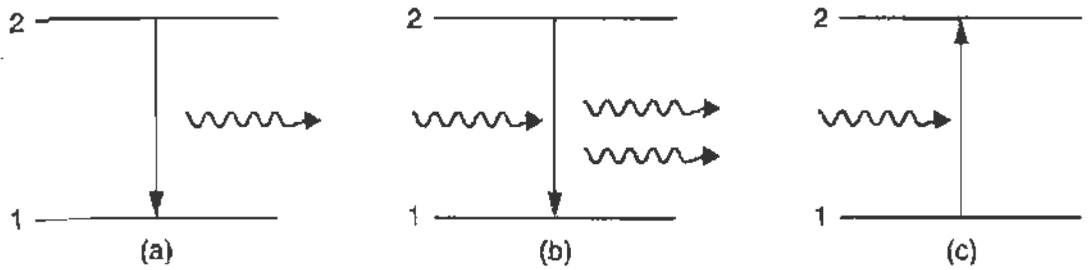


FIGURE 4.1 Emission and absorption of radiation between two atomic levels: (a) Spontaneous emission with transition from state 2 to state 1, (b) stimulated emission with transition 2 → 1, and (c) absorption with transition 1 → 2.

by A. Einstein, and the coefficients A, B are related and can be calculated from a knowledge of the structure (wave function) of the atomic states

$$A = \frac{\hbar\omega^3}{\pi^2 c^3} B = \frac{\omega^3}{\epsilon_0 \pi \hbar c^3} \, |\langle f|\mu \cdot \epsilon\,|i\rangle|^2. \tag{4.4}$$

Here $\langle f|\mu \cdot \epsilon\,|i\rangle$ is the matrix element of the electric dipole moment operator μ between the initial and final states and ϵ the polarization of the EM field. Loosely speaking, the matrix element is a measure of the average value of the distance of the electron from the nucleus.

The three possibilities are shown schematically in Fig. 4.1 and correspond to spontaneous emission, stimulated emission, and absorption of radiation. We remark that the probability for absorption is identical to that for stimulated emission for the same external field, a consequence of the reversibility in time of elementary physical processes. So far we have talked about *single* atoms, whereas in reality the lasing medium consists of a collection of N atoms. It is then important to know how many of the atoms are in the (excited) state 2 and how many in the (lower) state 1; we will designate these numbers by N_2, N_1, where $N_2 + N_1 = N$. The number of transitions per unit of time from 2 → 1 is then given by

$$R_{2\to 1} = N_2(A + Bu(\omega)), \tag{4.5}$$

whereas from 1 → 2

$$R_{1\to 2} = N_1 B u(\omega). \tag{4.6}$$

Here $u(\omega) = du/d\omega$ is the energy density of the EM field per unit frequency interval. Normally the relative population N_2/N_1 is governed by the

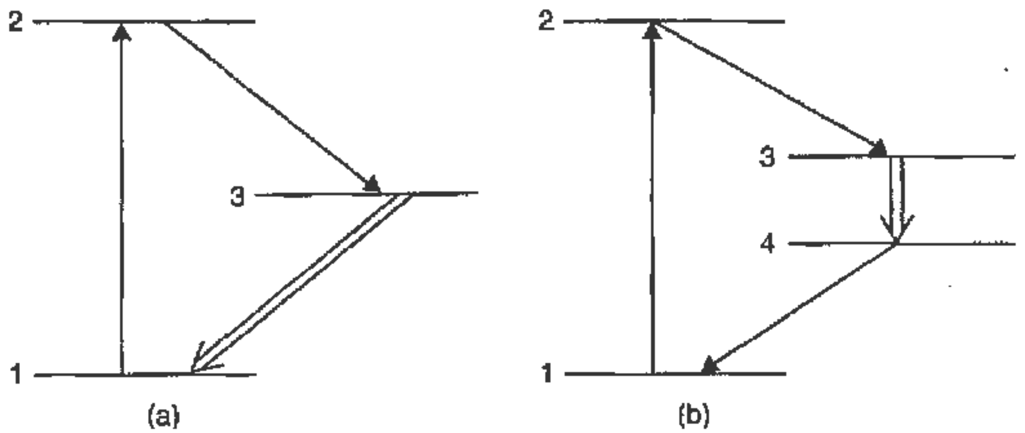


FIGURE 4.2 Creating population inversion (a) in a 3-level system and (b) in a 4-level system. The double arrow indicates the lasing transition, while the up-going arrow 1 → 2 is the pump. Level 3 must have a relatively long lifetime, whereas levels 2 and 4 should have a fast spontaneous decay along the indicated arrows.

Boltzmann distribution

$$\frac{N_2}{N_1} = \frac{Ne^{-E_2/kT}}{Ne^{-E_1/kT}} = e^{-[(E_2-E_1)/kT]}. \tag{4.7}$$

We can see from Eqs. (4.5) and (4.6) that for stimulated emission to occur in preference over absorption, we must have $N_2 > N_1$. Usually the opposite is true because ΔE for atomic levels is of order of a few electronvolts, whereas at room temperature $kT = 0.025$ eV, and from Eq. (4.7) $N_2 \ll N_1$. It is therefore necessary to create a *population inversion*, namely to increase N_2 while maintaining N_1 small. This can be achieved by involving three or four atomic levels as shown in Fig. 4.2. In the three-level laser, atoms are pumped from the ground state 1 to the excited state 2 and quickly decay to state 3 by spontaneous emission. If N_3 exceeds N_1 lasing can take place in the 3 → 1 transition. It is, however, easier to use a four-level scheme. In this case atoms are pumped from the ground state to level 2 and spontaneously decay to populate level 3. Now level 4 is practically empty because of a rapid transition to the ground state. Thus it is much easier to maintain $N_3 > N_4$ and achieve lasing in the 3 → 4 transition.

If the lasing medium is placed in an optical cavity (Fig. 4.3) we can assume that photons emitted along the cavity axis are trapped in the cavity and interact with the lasing medium, only a small fraction being lost. We consider a four-level laser and can set $N_2 = 0$ and $N_4 = 0$ because the transitions 2 → 3 and 4 → 1 are presumed fast. The total number of photons in the cavity is N_γ and n_1, n_3 are the atoms per unit volume in

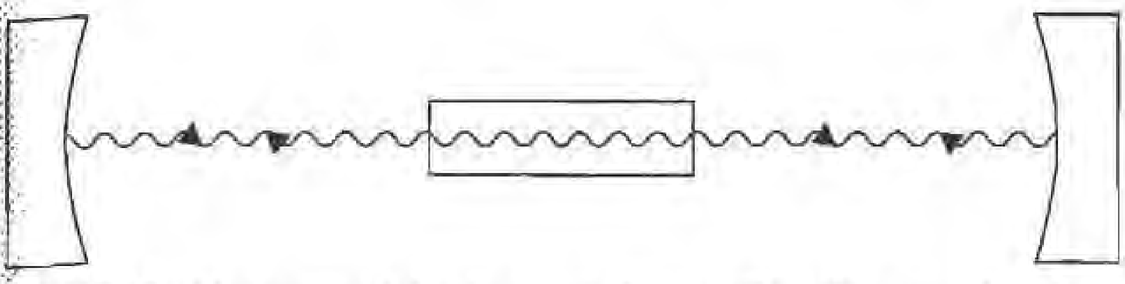


FIGURE 4.3 A lasing material placed in an optical cavity will lase if it is pumped to achieve sufficient gain.

states 1 and 3. It then holds that

$$n_1 + n_3 = n,$$

n being the atomic density, and

$$\frac{dn_3}{dt} = W_p n_1 - \overline{B} N_\gamma n_3 - \frac{n_3}{\tau} \tag{4.8}$$

$$\frac{dN_\gamma}{dt} = V\overline{B} N_\gamma n_3 - \frac{N_\gamma}{\tau_c}. \tag{4.9}$$

Here W_p is the probability per unit time for pumping $1 \to 2 \to 3$ (transferring atoms from state 1 to state 3) and $\overline{B}$ is the probability that one photon in the cavity will induce a lasing transition in unit time. The lifetime due to spontaneous transitions is τ and due to cavity losses τ_c. The (mode) volume in which the photons interact inside the lasing medium is designated by V. In all cases the spontaneous transition rate $1/\tau \ll \overline{B} N_\gamma$, so we can neglect this term. With this assumption, the *steady-state* solution of the rate equations (i.e., $dn_3/dt = dN_\gamma/dt = 0$) is

$$N_\gamma \simeq W_p n_1 V \tau_c. \tag{4.10}$$

In the steady state, the cavity losses per pass equal the gain per pass; the laser output depends linearly on the pump power, lasing medium density, and mode volume. Note that $V\overline{B} = c\sigma$ where σ is the cross section for the absorption of photons in the lasing medium.

The (logarithmic) gain per unit length of the lasing medium is found from Eq. (4.9) if we neglect the cavity losses. Then

$$\frac{dN_\gamma}{N_\gamma} = V\overline{B} n_3\, dt = V\overline{B} n_3 \frac{d\ell}{c} = \sigma n_3\, d\ell$$

and

$$g = \frac{1}{N_\gamma}\,\frac{dN_\gamma}{d\ell} = \sigma n_3.$$

Thus in a finite length ℓ, a number of incident photons $N_\gamma(0)$ will grow to

$$N_\gamma(\ell) = N_\gamma(0)e^{g\ell}.$$

Often $e^{g\ell} = G$ is designated as the gain per pass through the lasing medium.

4.2. PROPERTIES OF LASER BEAMS

Lasers emit a "beam" of light, the properties of the beam being determined primarily by the optical cavity. In the cavity shown in Fig. 4.3 the radiation travels in both directions and the electric and magnetic fields of the wave must satisfy boundary conditions at the two mirrors. Standing waves will exist in the cavity as shown in Fig. 4.4, and only frequencies such that the cavity length is an integral number of half-wavelengths are allowed. If the cavity length is ℓ, then

$$\frac{2\ell}{\lambda} = q \qquad \text{and} \qquad \nu = q\,\frac{c}{2\ell}, \tag{4.11}$$

where q is an integer. The frequency difference between two such adjacent *longitudinal modes* is

$$\nu_{q+1} - \nu_q = \frac{c}{2\ell} \equiv \text{FSR} \tag{4.12}$$

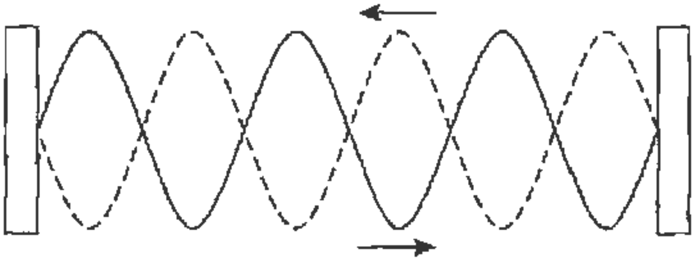


FIGURE 4.4 A laser cavity must support standing waves.

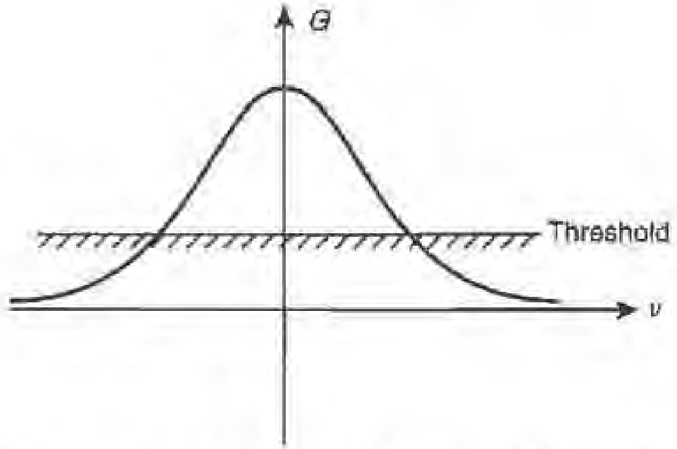


FIGURE 4.5 The gain curve of a typical lasing material as a function of frequency. Only lines with gain larger than the threshold will lase.

and is referred to as the *free spectral range* (FSR) of the cavity. As an example if we take $\ell = 0.5$ m, we find that $\mathrm{FSR} = 3 \times 10^8$ Hz. This spacing is very narrow as compared to the frequency of optical lines, i.e., for $\lambda = 600$ nm, $\nu = 5 \times 10^{14}$ Hz, and

$$q = \frac{2\ell}{c}\,\nu = \frac{\nu}{\mathrm{FSR}} \sim 1.6 \times 10^6.$$

Only a limited number of longitudinal modes are present in the emitted radiation. This is so because the lasing levels have finite energy width; this width determines the range of possible frequencies as shown in Fig. 4.5 and is referred to as the *gain curve*. The width of the individual longitudinal modes is determined by the number of round trips the light makes in the cavity before being attenuated; this is referred to as the finesse F of the cavity. The finesse depends on the losses in the cavity. If we consider only the losses at the mirrors that have a reflectivity $R < 1$, we find (see Section 4.6)

$$\Delta\nu = \frac{\mathrm{FSR}}{F} = \frac{c}{2\ell}\,\frac{(1-R)}{\pi\sqrt{R}} \tag{4.13}$$

or to a good approximation

$$\Delta\nu = \frac{c}{2\pi\ell}\,(1-R). \tag{4.14}$$

For $R = 0.99$ and $\ell = 0.5$ m, we find $\Delta\nu = 10^6$ Hz $= 1$ MHz. In contrast, the gain curve has a width of several gigahertz.

In the transverse direction the optical cavity is not bounded but is open. However, the beam is confined near the axis and its transverse structure is determined by the focal properties of the mirrors and the length of the cavity. A simple example is the confocal resonator, where both (spherical) mirrors have equal radii of curvature,[1] R, and R equals the distance, ℓ, between them; note that $f = R/2$, so that the focus is in the center of the cavity. The transverse beam distribution can assume any of the *transverse modes* characterized by the indices m, n. The electric field at a longitudinal distance z from the center of the cavity and at the transverse coordinates x, y is given by

$$E(x, y) = E_0 \frac{w_0}{w(z)} H_m\left(\frac{\sqrt{2}\,x}{w(z)}\right) H_n\left(\frac{\sqrt{2}\,y}{w(z)}\right) e^{-\left(\frac{x^2+y^2}{w^2(z)}\right)}, \qquad (4.15)$$

H_m, H_n are the m, n Hermite polynomials, and E_0 is the peak field value. For simplicity we have omitted the phase of the field.

Of particular interest is the lowest mode where $m = n = 0$, the TEM$_{00}$ mode. Since $H_0 = 1$, the field distribution is a Gaussian

$$E(x, y) = A\,\frac{w_0}{w(z)}\,e^{-\left(\frac{x^2+y^2}{w^2(z)}\right)}. \qquad (4.16)$$

The field falls to $1/e$ of its peak value, and the intensity to $1/e^2$, at a radius $r = w(z)$. We refer to $w(z)$ as the "beam radius" at the distance z. The smallest beam radius is at $z = 0$, where the wavefront is plane and normal to the cavity axis; we speak of a *beam waist* and for the confocal resonator

$$w_0\ \text{(confocal)}\ = \sqrt{\frac{\ell\lambda}{2\pi}}. \qquad (4.17)$$

The beam radius at the distance z is given by

$$w(z) = w_0\sqrt{1 + (z/z_0)^2}, \qquad (4.18)$$

where z_0 is the confocal parameter, or *Rayleigh range*. It is related to the beam waist through

$$z_0 = \frac{\pi w_0^2}{\lambda}. \qquad (4.19)$$

[1] It is unfortunate that the same symbol R is used for the reflectivity and the radius of curvature of a spherical mirror or lens.

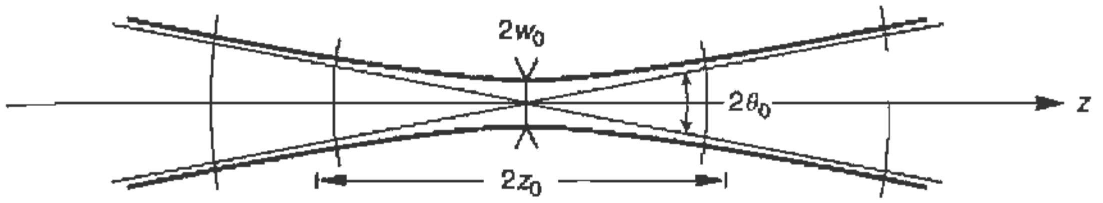


FIGURE 4.6 Focal properties of a TEM_{00} Gaussian beam propagating along z. At the waist the amplitude falls to $1/e$ of its on axis value at a distance w_0 from the axis. Note the wavefronts (surfaces of constant phase). The Rayleigh length z_0 and the divergence angle θ_0 are also indicated.

Thus for the confocal resonator, where Eq. (4.17) is applicable, we find that

$$z_0 \text{ (confocal) } = \ell/2.$$

In this case the beam radius at the mirrors has grown by $\sqrt{2}$ over the value at the waist.

At large distances $z \gg z_0$ the beam divergence is given by

$$\theta = \frac{w(z)}{z} \simeq \frac{w_0}{z_0} = \frac{\lambda}{\pi w_0}, \tag{4.20}$$

and for the confocal cavity

$$\theta \text{ (confocal) } = \sqrt{2\lambda/\pi\ell},$$

which is typically of order 10^{-3} or smaller. Figure 4.6 shows the rays, wavefronts, and beam waist in a confocal cavity. The fact that the beam cannot be focused to a point but instead forms a waist is due to the wave nature of the EM field.

Not all mirror combinations lead to stable cavities. The confocal resonator in particular is at the limit of the stable range and is not used in practice. Instead, most laser cavities consist of one perfectly reflecting flat mirror and of a curved mirror with radius $R > l$. Usually the curved mirror has a finite reflectivity, for instance 95%, and thus serves as the *output coupler*, by transmitting some fraction, say 5%, of the beam stored in the cavity.

4.3. THE HeNe LASER

The helium–neon gas laser is the most commonly used laser for simple laboratory work, alignment, and other low-power applications. The first

HeNe was built by A. Javan at Bell Labs in 1961 and now HeNe lasers are available at low cost from many manufacturers. A thin tube is filled with helium at a pressure of a few Torr and approximately 10% of neon gas is added. An electric discharge is established in the rarefied gas by the application of few kilovolts between the two electrodes. The electrons in the discharge excite the helium atoms to the $2S$ levels, which lie about 20 eV above the ground state. By a fortuitous coincidence these levels coincide with the $4S$ and $5S$ levels of neon. Through collisional exchange the neon atoms are excited to these levels, resulting in population inversion. Lasing takes place as indicated in Fig. 4.7, corresponding to the wavelengths

$$\begin{aligned} 5S &\to 3P \qquad \lambda = 632.8 \text{ and } 543 \text{ nm} \\ 4S &\to 3P \qquad \lambda = 1523 \text{ nm} \\ 5S &\to 4P \qquad \lambda = 3391 \text{ nm}. \end{aligned}$$

The $3P$ level de-excites quickly to the $3S$ state from where the atoms return to the ground state by colliding with the walls. By coating the mirrors for

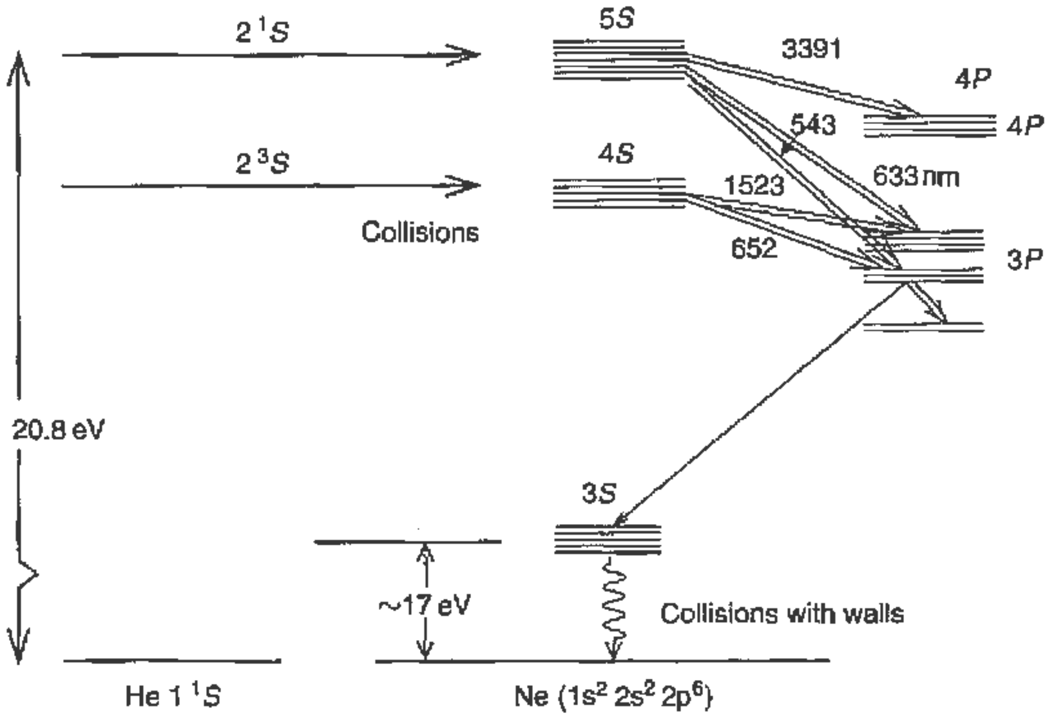


FIGURE 4.7 Energy levels of helium and neon. The principal lasing transitions are indicated by double arrows. Note that the ground state is at a much lower energy.

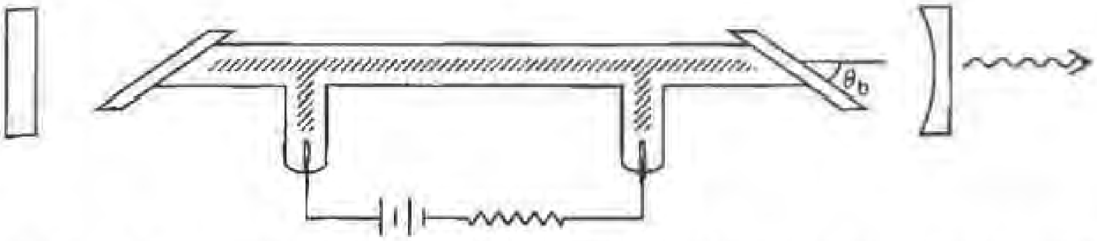


FIGURE 4.8 Schematic of a HeNe laser showing the discharge tube and the cavity mirrors.

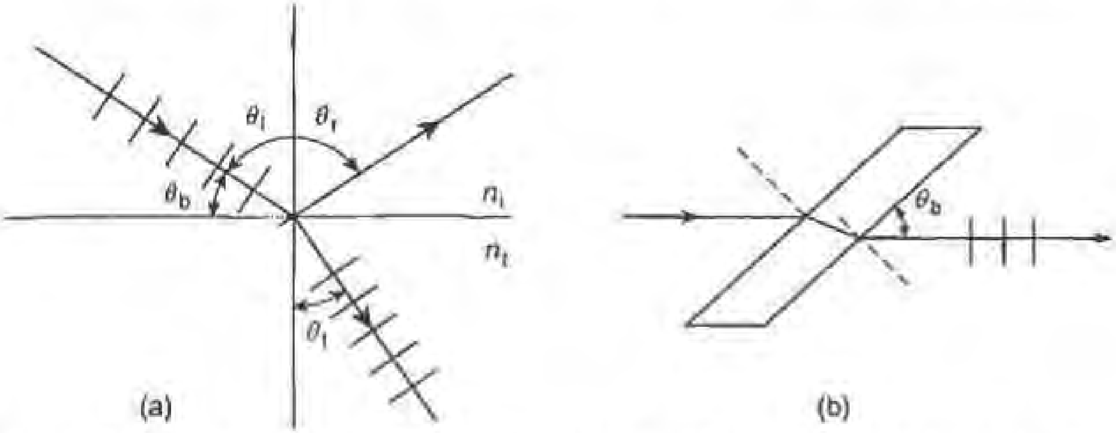


FIGURE 4.9 (a) Definition of Brewster's angle θ_b. (b) Transmission of a p-polarized ray at Brewster angle without attenuation.

reflectivity at a given wavelength, a particular laser line, most often the red line at 632.8 nm, can be selected.

A sketch of a HeNe laser is shown in Fig. 4.8. The tube diameter is chosen so as to maximize the population inversion of the neon atoms, an empirical formula relating the pressure (in Torr) to the tube diameter (in mm) being $pD \sim 4$ Torr-mm, usually $D \sim 2$ mm. The length of the optical cavity ranges from 20 to 50 cm. As shown in the sketch the electrodes are recessed. The gain in the low-pressure gas is relatively low, resulting in amplification $g \sim 0.10\ \text{m}^{-1}$. As a result the power level is also low, in the range of a few milliwatts. The width of the gain curve is dominated by Doppler broadening and is of order of 1.5 GHz.

A special feature in the sketch of Fig. 4.8 is the exit windows of the tube, which are set at the "Brewster" angle θ_b. As shown in Fig. 4.9, light polarized in the plane of incidence (p-light) and incident at θ_b is not reflected. If the refractive index of the window is n_t, the Brewster condition is

$$\frac{\pi}{2} - \theta_t = \theta_r = \theta_i$$

but from Snell's law

$$\sin\theta_t = \frac{n_i}{n_t}\sin\theta_i.$$

Therefore we must satisfy

$$\frac{\sin\theta_i}{\cos\theta_i} = \frac{n_t}{n_i}.$$

For $n_i = 1.0$ and $n_t = 1.5$, $\theta_i = 56.3°$ and the Brewster angle, which is the complement of θ_i, is $\theta_b = 33.7°$. Light polarized normal to the plane of incidence (s-light) is partially reflected from the windows and the higher losses prevent s-light from lasing.

In Eq. (4.12) of the previous section we showed that the spacing between the longitudinal modes is FSR $= c/2\ell$. One can demonstrate the presence of these modes by a simple experiment using a HeNe laser. Since $\ell = 0.3$ m, the FSR $= 500$ MHz, whereas the width of the gain curve is of order 1.5 GHz. Thus we can expect that three to four longitudinal modes could be lasing simultaneously. One way of observing these modes is to use a fast diode to record the intensity of the laser light. Because the diode detects the intensity, i.e., the square of the amplitude of the laser field, its signal will contain frequency components at the difference between the frequencies of the modes present in the light.

To explain this let us consider just two modes at frequencies ω_1 and ω_2. Then the amplitude (the electric field) is

$$A = A_1\cos\omega_1 t + A_2\cos\omega_2 t, \tag{4.21}$$

and the intensity (assuming A_1, A_2 real)

$$I = |A|^2 = A_1^2\cos^2\omega_1 t + 2A_1A_2\cos\omega_1 t\cos\omega_2 t + A_2^2\cos^2\omega_2 t. \tag{4.22}$$

The terms in $\cos^2\omega_1 t = \frac{1}{2}(1+\cos 2\omega_1 t)$ and $\cos^2\omega_2 t = \frac{1}{2}(1+\cos 2\omega_2 t)$ oscillate so fast that the diode will respond only to the constant part $|A_1|^2/2$ and $|A_2|^2/2$. However, the cross term can be expanded to give

$$2A_1A_2\cos\omega_1 t\cos\omega_2 t = A_1A_2\left\{\cos[(\omega_1+\omega_2)t] + \cos[(\omega_1-\omega_2)t]\right\}. \tag{4.23}$$

As before the term in $\cos[(\omega_1+\omega_2)t]$ will average to 0, but the diode can respond to the term in the difference frequency

$$A_1A_2\cos[(\omega_1-\omega_2)t]. \tag{4.24}$$

If there are more than two modes present we expect to see not only the fundamental difference frequency

$$\frac{1}{2\pi}(\omega_{q+1} - \omega_q) = \text{FSR},$$

but also higher harmonics arising from

$$\frac{1}{2\pi}(\omega_{q+2} - \omega_q) = 2\,\text{FSR}$$

and so on. Data obtained by using a fast diode connected to a microwave spectrum analyzer are shown in Fig. 4.10. The central peak is at 550 MHz, and there is a second peak at twice that frequency. (The peak on the left is just the DC level.) This indicates the presence of at least three longitudinal modes.

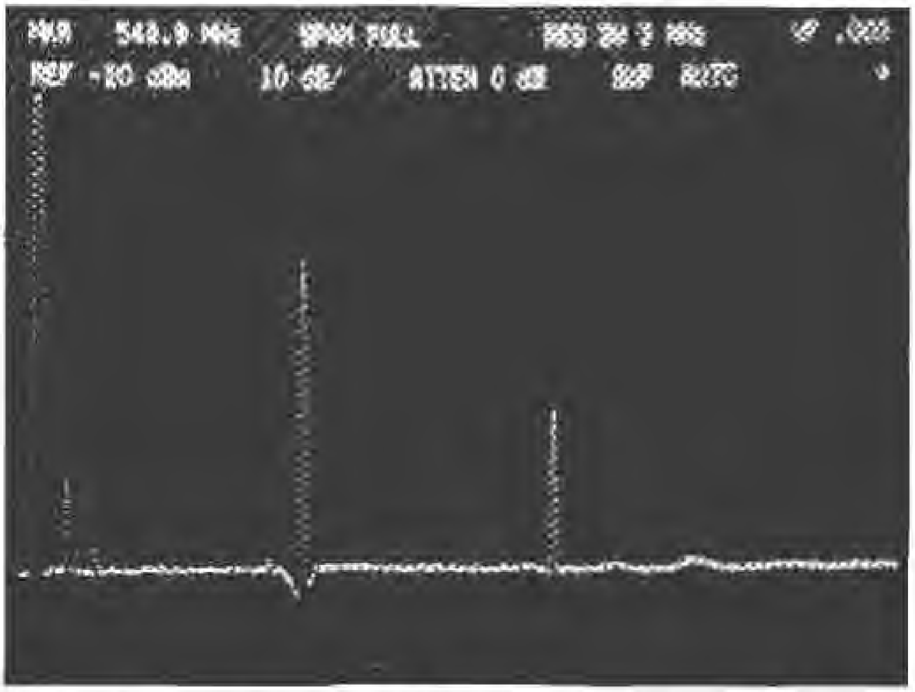


FIGURE 4.10 Microwave spectrum of the signal from a fast diode viewing a HeNe beam. The frequency scale is 184 MHz/cm. The line at 550 MHz (at 1100 MHz) gives the separation in frequency between adjacent longitudinal modes (modes differing by two integers). This spectrum indicates the presence of at least three longitudinal modes.

4.4. MEASUREMENT OF THE TRANSVERSE BEAM PROFILE

Often it is desired to expand or reduce the diameter of a laser beam while maintaining the parallelism, the *collimation*, of the beam. This can be achieved with a pair of lenses arranged as a "telescope." Telescopes were invented by Galileo and by Newton who used them to achieve *angular* magnification; the same arrangements are still in use and are named after their discoverers. To calculate the magnification of the beam we will use only geometrical optics; this is sufficient for our present considerations, even though a Gaussian laser beam diverges due to diffraction effects (see Fig. 4.6).

Figure 4.11 shows the *Newtonian* (or astronomical) telescope consisting of two focusing lenses of focal length f_1 and f_2. As shown in the sketch the lenses are converging (plano-convex), and the distance between them is

$$\ell = f_1 + f_2. \tag{4.25}$$

If a collimated beam of diameter d_1 is incident from the left, parallel to the telescope axis, it will exit with a diameter d_2, where

$$d_2 = d_1 \frac{f_2}{f_1}. \tag{4.26}$$

By appropriate choice of f_1, f_2 we can magnify or demagnify the beam.

The *Galilean* telescope is shown in Fig. 4.12; a diverging (plano-concave) lens of focal length f_1 and a converging lens of focal length f_2 are used. To preserve collimation the distance between the lenses must be

$$\ell = f_2 - f_1, \tag{4.27}$$

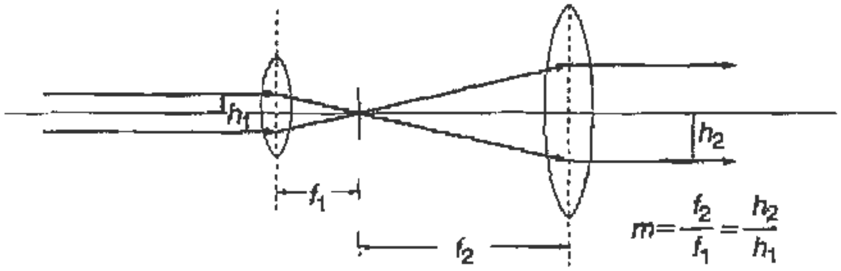


FIGURE 4.11 A Newtonian telescope with magnification f_2/f_1.

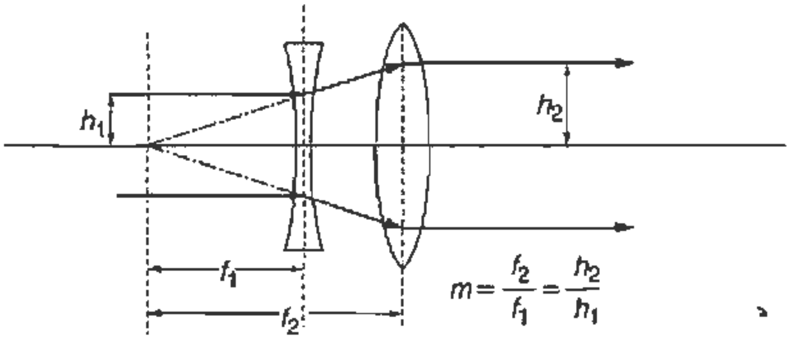


FIGURE 4.12 A Galilean telescope with magnification f_2/f_1.

and the spatial magnification is given by

$$d_2 = d_1 \frac{f_2}{f_1}. \tag{4.28}$$

The curved surface of the lens, whether convex or concave, is spherical, and the focal length is related to the radius of curvature, R, through the lens-makers equation. When the second surface is plane,

$$\frac{1}{f} = \frac{n-1}{R}, \tag{4.29}$$

where n is the index of refraction of the lens material. For most glasses used in lens manufacture and for visible light we can approximate $n \simeq 1.5$, so that $f \sim 2R$.

In setting up a telescope certain “alignment tricks” are useful. The beam must pass through the center of both lenses. Thus the lenses must be set on the optical table at the same height as the laser. In the horizontal direction one can be helped by noting that a beam that is passing through a lens off-center is steered. Furthermore, the surface of the lens must be perpendicular to the beam axis; this is most easily achieved by back-reflecting the beam.

The transmitted intensity of the beam is measured by a photodiode. (See Appendix E.) Since the photodiode area is small, it is often necessary to focus the beam on it, especially if it has been expanded. The diode is backward biased, usually with a low-voltage battery as shown in Fig. 4.13. With no incident light R_D is infinite. When light is incident some carriers are liberated and the resistance R_D of the diode decreases. Therefore the voltage across the load varies as

$$V_{out} = V_B \frac{R_L}{R_L + R_D}. \tag{4.30}$$

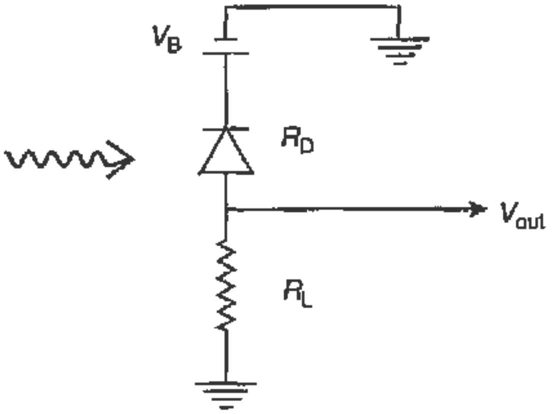


FIGURE 4.13 A photodiode reverse biased by a source V_B and working into a load R_L.

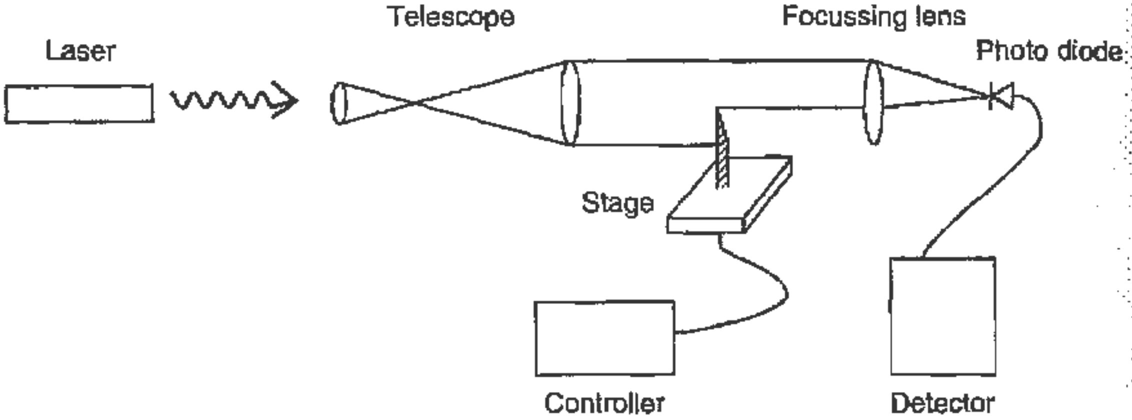


FIGURE 4.14 Arrangement for measuring the transverse profile of a laser beam.

At low light levels (i.e., where R_D is large) a digital voltmeter ($R_L \simeq 10\,\text{M}\Omega$) is adequate to read V_{out}. At high light levels the diode may become saturated, and it is desirable to use a shunt resistor; the signal can also be viewed on an oscilloscope, but when fast response is desired, as in Fig. 4.10, a 50-Ω impedance must be maintained throughout. When working at low light levels the photodiode must be shielded against room light.

One way of measuring the beam profile is to record the intensity received at the photodiode as a sharp edge (i.e., a razor blade) is moved through the beam. The blade is mounted on a translation stage that can be positioned with a resolution of few micrometers. The arrangement is sketched in Fig. 4.14, and the recorded intensity as a function of position gives the integral of the beam profile. If the beam profile in the direction of the blade motion is

$$I(x) = I_0\, g(x) \tag{4.31}$$

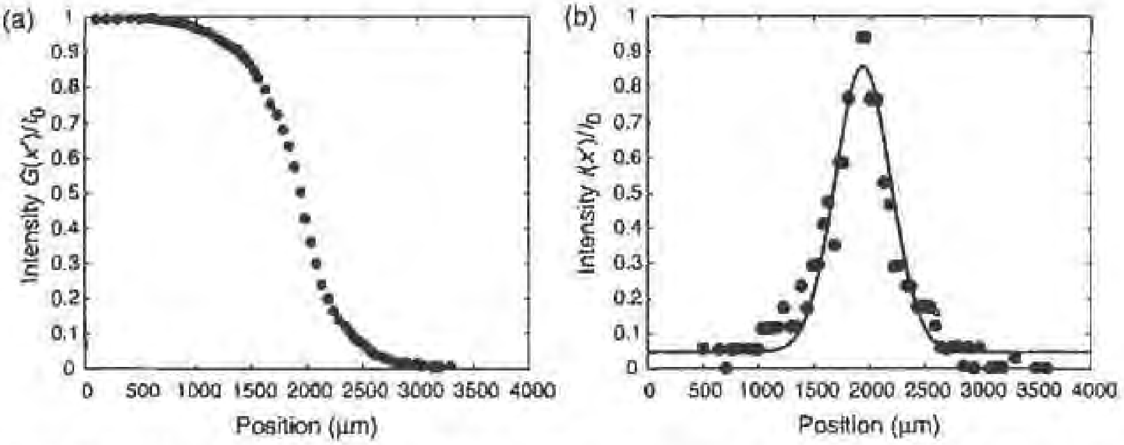


FIGURE 4.15 (a) The transmitted intensity as a function of the position of the obstacle (razor blade), which is moved across the beam. (b) The derivative of (a) gives the transverse profile of the beam intensity. It is fitted by a Gaussian.

with $\int_{-\infty}^{\infty} g(x)\,dx = 1$, the transmitted intensity when the blade is at position x' is

$$G(x') = I_0 \int_{x'}^{\infty} g(x)\,dx \tag{4.32}$$

(when the beam is fully unmasked, $x' \to -\infty$).

A typical result for the laser beam is shown in Fig. 4.15a where measurements were taken every 100 μm. By differentiating $G(x')$ we recover the intensity profile

$$-\frac{d}{dx'}\,G(x') = I(x'). \tag{4.33}$$

Performing this operation on the data of Fig. 4.15a we obtain the result shown in Fig. 4.15b, which can be adequately fitted by a Gaussian. The $1/e^2$ points of the Gaussian define the beam diameter, which in this case is $2w = 1000$ μm.

4.5. THE MICHELSON INTERFEROMETER

We are familiar with the fact that wave phenomena exhibit interference; namely at every point in space the amplitudes of two waves are added linearly (they are superimposed), whereas the intensity is determined by the square of the resultant amplitude. For interference to take place the two waves must retain their relative phase relationship over the time and space of the observation: they must be *coherent.* Laser beams are coherent in the plane normal to the direction of propagation but also over considerable length along the direction of propagation. For instance, a simple HeNe laser has a coherence length ℓ_c of order of meters. It is therefore possible to demonstrate interference with relative ease.

The arrangement of the Michelson interferometer is shown in Fig. 4.16. The HeNe beam is expanded (for convenience of observation) and is incident from the left on the "*beam-splitter*" B set at 45° with respect to the incident beam. A beam-splitter is a half-silvered glass plate or similar optical element that allows half of the beam to propagate through it toward $M1$ and reflects the other half toward $M2$. This technique of producing two coherent light beams is referred to as "amplitude division." The mirrors $M1$ and $M2$ reflect the corresponding beams that return to B. Half of the beam returning from $M1$ is transmitted through B, and the other half is reflected toward the screen; the same is true for the beam returning from $M2$. If B is set exactly at 45° and $M1$ and $M2$ are exactly normal to the beam direction, the two beams arriving at the screen are exactly parallel and their amplitudes will be superimposed.

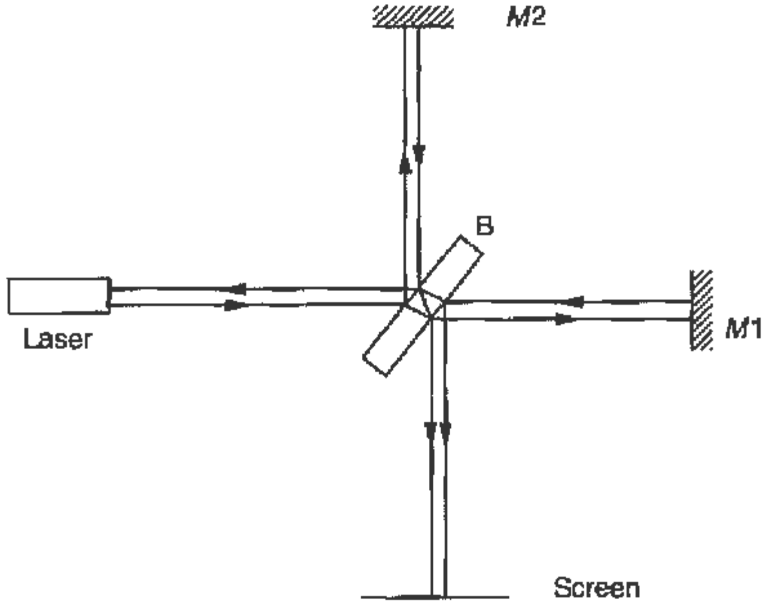


FIGURE 4.16 Outline of the Michelson interferometer. B is a beam splitter, $M1$ and $M2$ are the mirrors in the two arms and the interference pattern is observed on the screen.

If the intensity on the beam splitter is I, the wave amplitude[2] is

$$A_0(z,t) = E_0 \cos(\omega t - kz) \tag{4.34}$$

and

$$I_0 = \langle |A_0|^2 \rangle = E_0^2/2. \tag{4.35}$$

We set $z = 0$ at the beam splitter, and the amplitude of the wave is reduced by $\sqrt{2}$ each time it traverses (or is reflected from) the beam splitter. Thus the amplitudes coming from the two arms 1 and 2, when arriving at the screen, are

$$A_1(z_s,t) = \frac{E_0}{2}\cos(\omega t - 2k\ell_1 - k\ell_s)$$

$$A_2(z_s,t) = \frac{E_0}{2}\cos(\omega t - 2k\ell_2 - k\ell_s),$$

where ℓ_1, ℓ_2, and ℓ_s are the distances from the beam splitter to $M1$, $M2$, and the screen, respectively. The resultant amplitude at the screen is

$$\begin{aligned} A_s(z_s,t) &= \frac{E_0}{2}\left[\cos(\omega t - 2k\ell_1 - k\ell_s) + \cos(\omega t - 2k\ell_2 - k\ell_s)\right] \\ &= E_0 \cos[\omega t - k(\ell_1 + \ell_2 + \ell_s)]\cos[k(\ell_1 - \ell_2)], \end{aligned} \tag{4.36}$$

and the resultant intensity

$$I_s = \langle |A_s|^2 \rangle = \left(E_0^2/2\right)\cos^2[k(\ell_1 - \ell_2)]. \tag{4.37}$$

In Eqs. (4.35) and (4.37) we used the fact that $\langle \cos^2(\omega t) \rangle = \frac{1}{2}$. Note that the light reflected toward the source also forms an interference pattern of intensity

$$I_b = \left(E_0^2/2\right)\sin^2[k(\ell_1 - \ell_2)], \tag{4.38}$$

so that

$$I_s + I_b = I_0.$$

It is much more difficult to observe I_b than I_s.

[2] In this section we use trigonometric rather than exponential notation.

From the above analysis we conclude that the intensity at the screen will vary as $\cos^2[k(\ell_1 - \ell_2)]$. Since $k = 2\pi/\lambda$, it follows that when

$$\Delta\ell = |\ell_1 - \ell_2| = n\,\frac{\lambda}{2} \qquad n = 0, 1, 2 \ldots, \tag{4.39}$$

the screen will be bright (bright field), and when

$$\Delta\ell = |\ell_1 - \ell_2| = \left(n + \frac{1}{2}\right)\frac{\lambda}{2} \qquad n = 0, 1, 2, \ldots, \tag{4.40}$$

it should be completely dark (dark field). For intermediate values of $\Delta\ell$ the screen will be partially illuminated as indicated by Eq. (4.37). The idealized situation described by Eqs. (4.39) and (4.40) is very difficult to obtain in practice: very slight misalignment of the mirrors and even small air currents are sufficient to change the relative phase of different parts of the wavefront. Imperfections or nonflatness of the mirrors or beam-splitter at the level of a fraction of a wavelength distort the wavefront and modify the interference pattern.

Nonparallelism between the mirrors $M1$ and $M2$ gives rise to "interference fringes" at the screen. We assume that the two mirrors are set so that their normals are in the plane of incidence (the plane of the paper in Fig. 4.17), but $M2$ is misaligned by an angle α with respect to the axis of the beam as shown. Because the rays returning from $M1$ are reflected by 90° at B, we *can think* of $M1$ as located at $M1'$, and that the reflected rays propagate in exact parallelism with the z axis. The z axis is defined from the screen toward $M2$ and the x axis is in the direction of the screen as indicated in the figure. For a small misalignment angle α, a well-collimated beam, and for ℓ_2, ℓ_s sufficiently large we need consider only rays from $M2$ that propagate parallel to the z axis. Then the rays reaching the point x on the screen have traversed path lengths

$$\begin{aligned} z_1 &= 2\ell_1 + \ell_s \\ z_2 &= 2(\ell_2 + x\tan\alpha) + \ell_s, \end{aligned} \tag{4.41}$$

and their path difference is

$$(z_1 - z_2) = 2(\ell_1 - \ell_2) - 2x\tan\alpha. \tag{4.42}$$

Bright fringes perpendicular to the plane of incidence will appear on the screen when $(z_1 - z_2) = n\lambda$. Consequently, the fringes are separated on

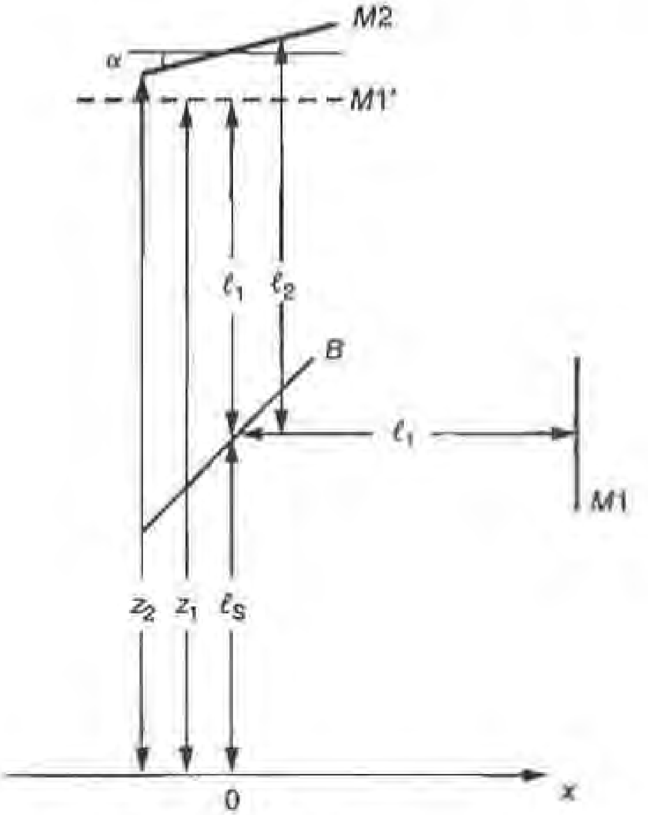


FIGURE 4.17 Schematic of the Michelson interferometer with one mirror slightly misaligned. To calculate the interference pattern $M1$ can be relocated at the dotted line $M1'$. Vertical (to the plane of the paper) fringes appear on the screen separated by $\Delta x = \lambda/(2\tan\alpha)$.

the screen by a distance

$$\Delta x = \frac{\lambda}{2\tan\alpha}. \tag{4.43}$$

For example, for the HeNe, $\lambda = 633$ nm and if we take $\alpha = 10^{-4}$, we find $\Delta x \simeq 3$ mm. As the angle α is increased the fringes crowd together and eventually the interference pattern is lost.

In the previous discussion we have implicitly assumed that the expanded HeNe was collimated; for a noncollimated beam the fringes form a circular pattern. Some residual curvature is observed even with a collimated beam when the interferometer is not perfectly aligned or when the optics have aberrations.

In the laboratory we set up the mirror $M1$ on a translation stage (the same as used for the beam profile measurements). The mirrors are carefully aligned until an interference pattern is achieved. When the translation stage is moved, the interference pattern changes: for a stage motion $\Delta z = \lambda/4$ bright fringes become dark and vice versa, and the original pattern reappears for $\Delta z = \lambda/2$. When the motion is continuous the fringe pattern appears to "walk" across the screen and one can count how many fringes have passed by, for a given amount of motion. It is convenient to measure Δz for ~25 fringes at a time; this corresponds to motion of ~8 μm, which can be adequately resolved by the counter on the translation stage. The wavelength is immediately obtained from

$$\lambda = 2(\Delta z/N), \tag{4.44}$$

where Δz is the motion of the stage and N the number of fringes that passed by.

4.6. THE FABRY–PEROT INTERFEROMETER

In the Michelson interferometer, two coherent waves were made to interfere. In the arrangement introduced by Fabry and Perot a very large (in theory infinite) number of waves are made to interfere. Because of the participation of many waves, very sharp contrast between bright and dark fringes can be obtained and this results in excellent wavelength resolution.

The Fabry–Perot consists of two mirrors, often parallel plates coated on their inner surface to have good reflectivity at the wavelength of interest. The spacing, t, between the plates is maintained by precision spacers, forming an assembly referred to sometimes as an *étalon*. This is shown schematically in Fig. 4.18, where for simplicity we have shown the plates as infinitely thin. A ray incoming at an angle θ with respect to the normal after traversing plate 1 will undergo repeated reflections. We label the rays emerging from plate 2 by AB, CD, EF, etc. The path difference between two adjacent rays, say AB and CD, is

$$\Delta\ell = BC + CK$$

with BK normal to CD. The finite thickness of the plate does not modify the above relation. It follows that

$$\Delta\ell = 2t\cos\theta. \tag{4.45}$$

Note that $CK = BC\cos 2\theta$ and $BC\cos\theta = t$; thus, $\Delta\ell = BC(1 + \cos 2\theta) = 2BC\cos^2\theta = 2t\cos\theta$. Therefore, constructive interference

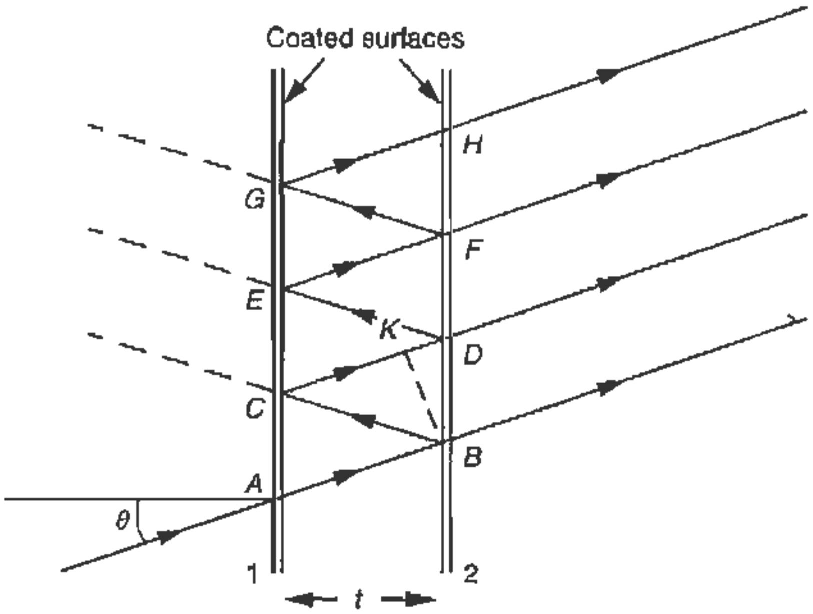


FIGURE 4.18 The Fabry–Perot interferometer. A ray incident at an angle θ is shown. For simplicity the mirrors are indicated as infinitely thin. Note that an infinite number of reflections contribute to the transmitted intensity at angle θ.

will occur when the path difference is a multiple of a wavelength

$$2t\cos\theta_n = n\lambda. \tag{4.46}$$

Since θ_n is a small angle, n is a large number of order $n \sim 2t/\lambda$.

The above constructive interference condition holds provided the distance form the étalon to the point of observation is the same for all rays, namely when the observation point is at infinity. To achieve this we use a lens to focus the rays emerging from the étalon onto a screen. For a slightly diverging incident beam one observes a set of rings of radius

$$r_n = f\tan\theta_n \sim f\theta_n, \tag{4.47}$$

where θ_n is determined by Eq. (4.46) and f is the focal length of the lens. Note that the incident beam should not be perfectly collimated but should contain enough angular divergence to support the angles θ_n.

To obtain the spacing between consecutive maxima (fringes) we first note that for $\theta = 0$, the path difference between adjacent beams, measured in wavelengths, is

$$n_0 = 2t/\lambda, \tag{4.48}$$

which in general is *not* an integer. The first observable ring is formed at an angle θ_1 where n_1 is the integer closest to (smaller than) n_0. Thus

$$n_1 = n_0 - \epsilon \qquad 0 \le \epsilon \le 1$$

and

$$\epsilon = \frac{2t}{\lambda}(1 - \cos\theta_1) = \frac{4t}{\lambda}\sin^2\left(\frac{\theta_1}{2}\right). \tag{4.49}$$

As we move out from the center, the pth ring corresponds to

$$n_p = (n_0 - \epsilon) - (p - 1). \tag{4.50}$$

Using Eq. (4.50) in Eq. (4.46), recalling the definition of Eq. (4.48), dropping ϵ with respect to $(p-1)$ and replacing $2\sin^2(\theta/2)$ by $\theta^2/2$ we find that the angle of the pth ring is

$$\theta_p \simeq \sqrt{(p-1)\frac{\lambda}{t}}, \tag{4.51}$$

applicable for moderately large values of p, $p \gtrsim 5$. As an example, if $t = 1$ cm and $\lambda = 633$ nm we have $\lambda/t \simeq 6.3 \times 10^{-5}$ and the $p = 11$ ring will appear at $\theta = 25 \times 10^{-3}$ rads; for a lens with focal length $f = 40$ cm the radius is 1 cm.

Next we calculate the intensity of the rings (fringes) and the contrast between bright and dark fringes. We designate by T the power transmission coefficient of the inner surfaces of the étalon. For simplicity we also assume that both surfaces have the same transmission and reflection coefficients. The power reflection coefficient is R, so that in the *absence* of absorption

$$R + T = 1.$$

The amplitude transmission and reflection coefficients are designated by

$$t = \sqrt{T} \qquad \text{and} \qquad r = \sqrt{R}.$$

We also designate the incident intensity by I_0 and the amplitude by A_0, where $I_0 = \frac{1}{2}A_0^2$. The transmitted ray B will have amplitude

$$A_B = A_0 t^2 e^{i\phi}, \tag{4.52}$$

where ϕ is a phase acquired in traversing both plates and the space between them. Ray D will have amplitude

$$A_D = A_B r^2 e^{i2\delta}, \tag{4.53}$$

ray F

$$A_F = A_D r^2 e^{i2\delta}, \tag{4.53'}$$

and so on. Here the phase angle 2δ is due to the path difference of adjacent rays as they travel between the plates. It follows from Eq. (4.45) that

$$2\delta = 2\pi \frac{2t \cos\theta}{\lambda}. \tag{4.54}$$

From Eqs. (4.53) we see that the amplitude of successive rays decreases by $r^2 = R$, but there is an infinite number of such rays. The amplitude of the transmitted light is

$$A_T = A_0 t^2 e^{i\phi} \sum_{q=1}^{\infty} \left[1 + r^{2q} e^{iq2\delta}\right]. \tag{4.55}$$

This geometric series can be easily summed

$$A_T = A_0 t^2 e^{i\phi} \frac{1}{1 - r^2 e^{i2\delta}},$$

and the transmitted intensity

$$I_T = \frac{1}{2} |A_T|^2 = I_0 \frac{T^2}{(1-R)^2 + 4R \sin^2\delta}. \tag{4.56}$$

Maxima occur when δ is an integral multiple of π, whereas minima occur when δ is a half-integral multiple of π. At the maxima

$$I_T = \frac{I_0 T^2}{(1-R)^2}. \tag{4.57}$$

We see that in the absence of absorption I_T (max) $= I_0$. At the minima

$$I_T = \frac{I_0 T^2}{(1+R)^2} = I_0 \frac{(1-R)^2}{(1+R)^2}, \tag{4.58}$$

showing that very good contrast can be achieved if R is close to 1. Equation (4.56) is plotted in Fig. 4.19 for different values of R.

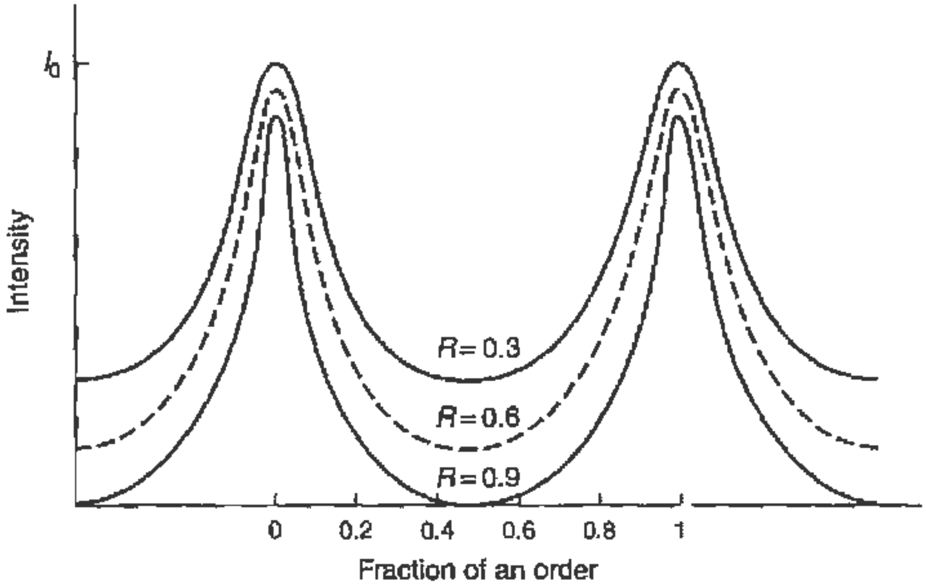


FIGURE 4.19 The width of the Fabry–Perot fringes as a function of mirror reflectivity. The two peaks are separated in frequency by 1 FSR $= c/2t$.

The bright fringe will reach half its peak intensity when

$$4R\sin^2(\delta_{1/2}) = (1-R)^2$$

or when

$$\delta_{1/2} = \frac{(1-R)}{2\sqrt{R}}, \tag{4.59}$$

where the small angle approximation was used. The full-width at half-maximum (FWHM) of the fringe is $2\delta_{1/2}$. The spacing between adjacent fringes corresponds to a phase angle difference of 2π, and we define the *finesse* of the Fabry–Perot interferometer as the ratio between fringe spacing and the HWHM of the fringe

$$F = \frac{2\pi}{\delta_{1/2}} = \frac{\pi\sqrt{R}}{1-R}. \tag{4.60}$$

For a typical reflectivity $R = 0.98$, the finesse is $F = 155$.

The spacing between bright fringes defines the free *spectral range* of the interferometer. Let the wavelength λ_1 form its pth ring at angle θ, and wavelength λ_2 form its $(p-1)$ ring at the same angle. Since these two rings overlap,

$$(n-1)\lambda_2 = n\lambda_1 \qquad \text{or} \qquad \lambda_2 - \lambda_1 = \lambda_2/n.$$

However, $n\lambda_1 \sim n\lambda_2 \sim 2t$, so we obtain that

$$\lambda_2 - \lambda_1 = \lambda_2^2/2t. \tag{4.61}$$

If we express Eq. (4.61) in terms of frequency, $\nu = c/\lambda$ we find that

$$\nu_1 - \nu_2 = \frac{c}{2t}.$$

Namely overlapping rings correspond to the free spectral range already introduced in Eq. (4.12). For instance for $\lambda = 633$ nm and $t = 1$ cm the wavelength spacing is $\delta\lambda = \lambda^2/2t = 0.02$ nm. However, lines between fringes can be resolved if they do not exactly overlap and this depends on the line width, i.e., the finesse of the instrument. Thus, the wavelength resolution is given by

$$d\lambda = \frac{1}{F}(\lambda_2 - \lambda_1) = \frac{1}{F}\frac{\lambda^2}{2t}. \tag{4.62}$$

For the above example and for $F = 155$, $d\lambda/\lambda \sim 2 \times 10^{-7}$, showing that extremely high resolution can be achieved with a relatively simple apparatus.

Fabry–Perot etalons used in conjunction with lasers are frequently made with two focusing mirrors rather than flat plates. This facilitates the alignment but fixes the free spectral range. They serve as high-resolution filters to select specific wavelengths and as optical "spectrum analyzers," which are in essence high-resolution scanning spectrometers.

CHAPTER 5

Optics Experiments

5.1. INTRODUCTION

The wide use of lasers in so many applications has increased the need for high-quality optics and for good optical designs. We address some of these questions in this chapter where we discuss the diffraction of light and rotation of the optical polarization, as well as propagation in optical fibers.

When a collimated beam of light passes through an aperture, or if it encounters an obstacle, it spreads out and the resulting pattern contains bright and dark regions. This effect is called *diffraction*, and is characteristic of all wave phenomena. It can be understood by considering the interference between different parts of the wavefront, which was altered in passing through the aperture. The angle of diffraction is of order λ/d with λ the wavelength and d the dimension of the aperture. Thus, for visible light, apertures in the range 10–100 μm produce easily resolved diffraction patterns.

Very different patterns are formed near and far from the aperture. In the near field we speak of *Fresnel diffraction*, and to observe the pattern it is convenient to form an *image* of it on a screen. In the far field we obtain the *Fraunhofer diffraction* pattern, which can be observed by simply placing a screen at some distance from the aperture; more precisely a lens should be used and the pattern observed in the focal plane. In the following three sections we discuss Fraunhofer diffraction from a slit and a circular aperture. The results shown were obtained with a CCD camera.

The diffraction grating was already introduced in Chapter 1. In Section 5.5 we derive the grating equation and show a modern setup that can be readily digitized; also included are results on the Hg spectrum. Next we introduce the concept of "spatial frequency" components in a beam of light. This allows us to manipulate an image by imposing suitable spatial filters, in the focal plane, a procedure also referred to as "Fourier optics." We have kept the mathematics simple and emphasized the physical principles instead. In Section 5.7 we discuss the Faraday effect, namely the rotation of the linear polarization of light when traversing a medium immersed in an axial magnetic field. The power of the lock-in detection technique is evident in this experiment. The last section is a demonstration and measurement of "Berry's phase." This is the rotation of polarization due to a topological change in the direction of propagation of the light. It is demonstrated by injecting the light in an optical fiber that is wound as a helix.

5.2. DIFFRACTION FROM A SLIT

We can find the minima of the diffraction pattern with the help of the sketches shown in Fig. 5.1. Consider a plane wave of visible light incident on a vertical slit of width d. The incident "rays" are normal to the screen that contains the slit; i.e., the *wavefront* is parallel to the screen. We can "divide" the slit in half (i.e., $AB = BC = d/2$) and consider the rays 1 and 2 emerging at an angle θ with respect to the direction of incidence (Fig. 5.1a). The path difference between these rays is $BD = AB \sin\theta = (d/2)\sin\theta$. If the path difference is $\lambda/2$, then at *large distances* from the slit, rays 1 and 2 will interfere destructively. However, this will also happen for rays $1'$ and $2'$, $1''$ and $2''$, and so on, so that there will be no light in the direction θ_1, where

$$\frac{d}{2}\sin\theta_1 = \frac{\lambda}{2}. \tag{5.1}$$

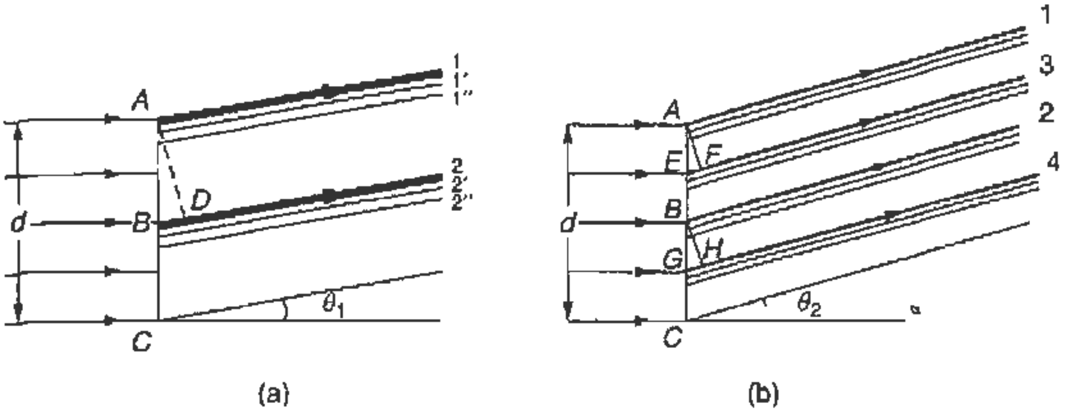


FIGURE 5.1 Finding the minima of a diffraction pattern (a) the slit of width d is "divided" in half and (b) into quarters. The rays are focused at infinity and the path difference is indicated.

In contrast, at $\theta = 0$ the path length (out to a large distance) of all rays is equal and the resultant amplitude is maximal.

To find the next zero, let us "divide" the slit into quarters as shown in Fig. 5.1b. In this case ray 1 will interfere destructively with ray 3 when $EF = AE \sin\theta = (d/4)\sin\theta = \lambda/2$. However, ray 2 also will interfere destructively with ray 4 and similarly for all intermediate rays. Thus there will be no light in the direction θ_2, where

$$\frac{d}{4}\ \sin\theta_2 = \frac{\lambda}{2}. \tag{5.2}$$

This argument can be continued by subdividing the slit into an (integral) number of smaller and smaller segments. By analogy with Eqs. (5.1) and (5.2) we find the generalized expression for the minima

$$d \sin\theta_n = \pm n\lambda, \qquad n = 1, 2, 3, \ldots. \tag{5.3}$$

For small angles $\sin\theta \sim \theta$ and

$$\theta_n = n\,\frac{\lambda}{d}, \qquad n = 1, 2, 3, \ldots. \tag{5.4}$$

The complete expression for the intensity distribution of the diffracted light is derived in the next section. It is

$$I(\theta) = I_0 \left[\frac{\sin\left(\frac{\pi d}{\lambda}\sin\theta\right)}{\frac{\pi}{\lambda}\sin\theta}\right]^2, \tag{5.5}$$

where I_0 is the intensity (into a small angular interval $d\theta$) in the forward direction ($\theta = 0$). Intensity is the energy traversing unit area in unit time

$$I = |\mathbf{S}| = |\mathbf{E} \times \mathbf{H}| = c\epsilon_0 |E|^2, \tag{5.6}$$

where $\mathbf{E}$, $\mathbf{H}$ are the electric and magnetic fields of the light wave, and we usually take the time average, which introduces a further factor of $\frac{1}{2}$. Equation (5.5) has zeros in agreement with Eq. (5.3); maxima occur (to a good approximation) when

$$d \sin\theta_m = \pm\left(m + \frac{1}{2}\right)\lambda, \qquad m = 1, 2, 3, \ldots. \tag{5.7}$$

The intensity at the secondary maxima decreases as m increases. Equation (5.5) is of the general form

$$f(x) = \frac{\sin^2 x}{x^2}, \tag{5.8}$$

which is graphed in Fig. 5.2. Note that as $x \to 0$, $f(x) \to 1$.

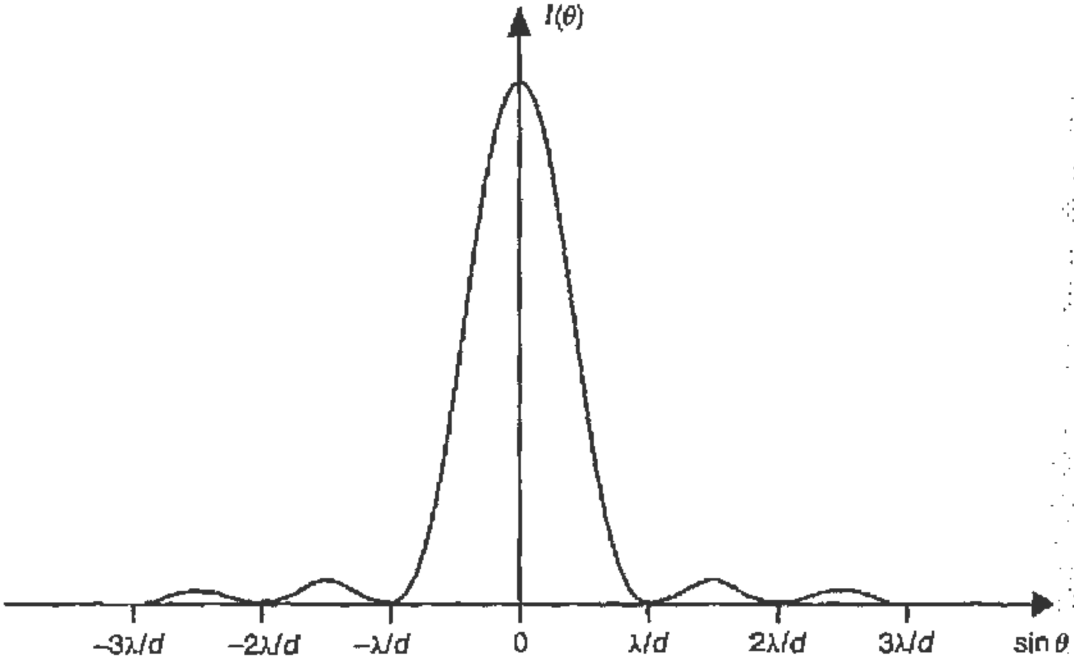


FIGURE 5.2 Plot of the Fraunhofer diffraction pattern $\sin^2 x/x^2$.

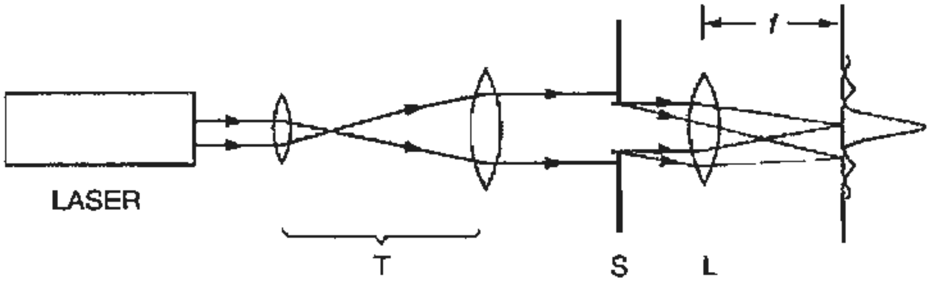


FIGURE 5.3 Schematic of a simple layout to observe Fraunhofer diffraction.

The experimental setup is shown in Fig. 5.3. The laser beam is expanded in a 4:1 telescope T, to better approximate a plane wave and is then incident on the slit D. The diffraction pattern is observed on the screen S, which is in the focal plane of the lens L. Thus, we observe the image of the pattern formed at infinity. The slit width was $d = 200$ μm and the focal length $f = 50$ cm, so that the first minimum appears at a distance x from the principal maximum

$$x = f\theta \sim f \sin\theta = f(\lambda/d) = 1.68 \text{ mm},$$

where we used $\lambda = 633$ nm. A picture of the diffraction pattern taken with a CCD camera is shown in Fig. 5.4. The central spot saturates the camera.

Instead of using a slit, we can observe the same diffraction pattern by placing a thin wire of width d in the path of the beam. Since it is easier to

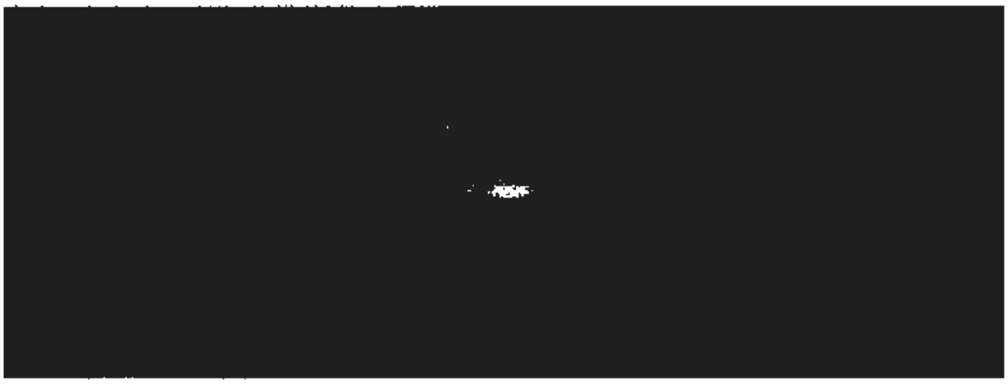


FIGURE 5.4 Diffraction pattern from a thin slit observed in the focal plane obtained with a CCD camera.

obtain thin wires (or hairs) than to manufacture thin slits, the former are often used for demonstrating diffraction. That the two patterns are equivalent (*except* in the forward direction) is known as *Babinet's principle*. To illustrate the principle we assume that the incident plane wave is "uniform and infinite" in extent in the x direction. Thus, the amplitude of the wave is independent of x, $A(x) = A$. Immediately after the *slit*, the amplitude B is given by

$$\begin{aligned} B(x) &= A \qquad -d/2 < x < d/2 \\ B(x) &= 0 \qquad\qquad |x| > d/2. \end{aligned} \tag{5.9}$$

In the presence of the obstacle the amplitude C of the wave, just past the obstacle, is given by

$$\begin{aligned} C(x) &= 0 \qquad -d/2 < x < d/2 \\ C(x) &= A \qquad\qquad |x| > d/2. \end{aligned} \tag{5.10}$$

Combining Eqs. (5.9) and (5.10) we can write

$$C(x) = A - B(x) \qquad \text{(valid for all } x\text{).} \tag{5.11}$$

We know that when the amplitude is constant for all x, the wave propagates only in the $\theta = 0$ direction. Thus for angles $\theta \neq 0$, the constant amplitude does not contribute, and Eq. (5.11) becomes

$$C(x) = -B(x) \qquad \text{(valid for } \theta \neq 0\text{).} \tag{5.12}$$

It follows that the diffraction pattern, which is proportional to the square of the amplitude

$$|C(x)|^2 = |B(x)|^2,$$

is the same in both cases. Equation (5.5) remains valid but with the direction $\theta = 0$ excluded.

Another case of interest arises when instead of a slit a square aperture is used. The result is shown in Fig. 5.5 and consists, primarily, of two single-slit diffraction patterns along the x and y directions. The intensity of the maxima in directions differing from the x or y axes decreases very rapidly.

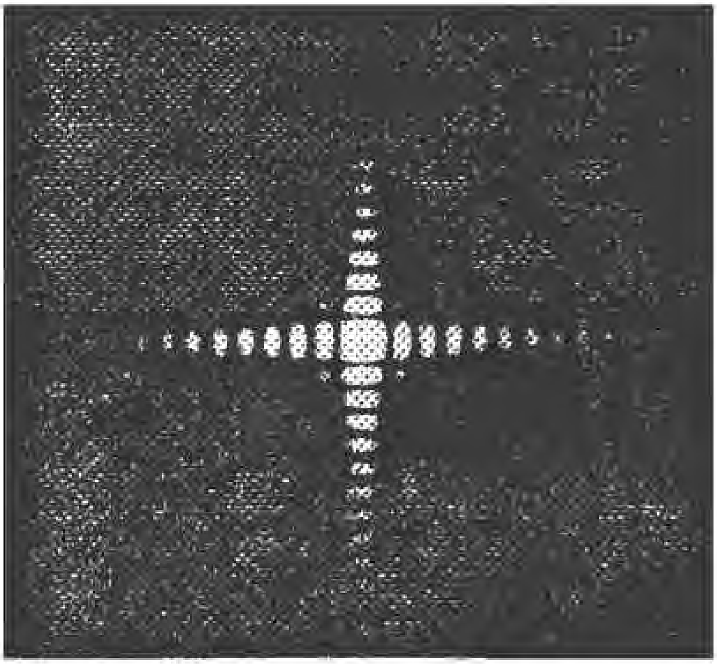


FIGURE 5.5 Diffraction from a square aperture.

5.3. CALCULATION OF THE DIFFRACTION PATTERN

To obtain an expression for the diffraction pattern formed by an aperture, we will make use of the Huygens–Fresnel principle. The principle states that every point on the aperture D is a source of *spherical* wavelets with amplitude and phase determined by the incident wave. These "secondary" wavelets propagate at all angles and interfere at every point in the observation plane to determine the diffracted wave amplitude. We take the incident wavefront parallel to the aperture plane, whereas the observation plane is located at infinity (Fraunhofer diffraction). This is approximated in the sketch of Fig. 5.6 where we show both the aperture and observation planes and two typical rays to the observation point P'. We need be concerned only with the transverse coordinates. In the aperture plane, the point M is specified by the coordinates ζ, η, whereas in the observation plane, the point P' is specified by x', y'.

Because we observe at *infinity*, R, the distance from O to the observation point is very large (and equals OS') as compared to the dimensions of the aperture; therefore, rays OP' and MP' are to be considered as *parallel*. Then the path difference between the ray OP' (from the coordinate origin to P') and the ray MP' (from the source point to P') is the length OB

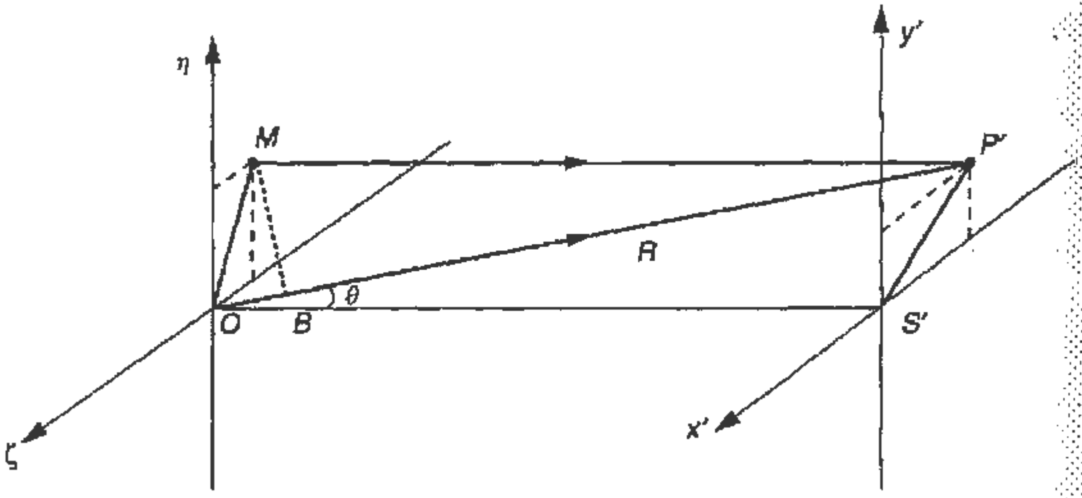


FIGURE 5.6 Coordinate systems in the aperture and observation planes for calculating diffraction.

where *MB* is perpendicular to *OP'*. If we designate by **q** the unit vector along the ray *OP'* we obtain for *OB*

$$OB = \mathbf{OM} \cdot \mathbf{q} = \zeta \left(\frac{x'}{R}\right) + \eta \left(\frac{y'}{R}\right) = [\zeta x' + \eta y']/R. \qquad (5.13)$$

The *direction cosines* of the vector *OP'* are

$$(x'/R) = u \qquad \text{and} \qquad (y'/R) = v$$

and are well defined whether *R* is finite or tends to infinity. The phase difference between the two rays is

$$\Delta\phi = \frac{2\pi}{\lambda}\,[\zeta u + \eta v]. \qquad (5.14)$$

Without loss of generality we set the reference phase of the ray *O P'* equal to 0; we can then write for the contribution of the source point *M* to the amplitude at the observation point *P'*

$$dA'(x', y') = e^{i\frac{2\pi}{\lambda}[\zeta u + \eta v]}\,d\zeta\,d\eta. \qquad (5.15)$$

For simplicity we dropped the time-dependence $e^{-i\omega t}$; $d\zeta\,d\eta$ is the differential element of the aperture at the point *M*.

To obtain the amplitude at the point P' we must integrate the contribution from all source points. If the amplitude and phase of the incident wave are *constant* over the aperture, we directly integrate Eq. (5.15). For the case of a square aperture with dimensions $2\zeta_0$ and $2\eta_0$ the integral is elementary,

$$\begin{aligned} A'(x', y') &= \int_{-\eta_0}^{\eta_0} \int_{-\zeta_0}^{\zeta_0} e^{i\frac{2\pi}{\lambda}[\zeta u + \eta v]} d\zeta\, d\eta \\ &= 4\zeta_0\eta_0 \left[\frac{\sin\left(\frac{2\pi}{\lambda}\zeta_0 u\right)}{\frac{2\pi}{\lambda}\zeta_0 u} \right] \left[\frac{\sin\left(\frac{2\pi}{\lambda}\eta_0 v\right)}{\frac{2\pi}{\lambda}\eta_0 v} \right]. \end{aligned} \tag{5.16}$$

The intensity is given by the square of the amplitude

$$I(x', y') = 16\zeta_0^2\eta_0^2 \left[\frac{\sin\left(\frac{2\pi}{\lambda}\zeta_0 u\right)}{\frac{2\pi}{\lambda}\zeta_0 u} \right]^2 \left[\frac{\sin\left(\frac{2\pi}{\lambda}\eta_0 v\right)}{\frac{2\pi}{\lambda}\eta_0 v} \right]^2 \tag{5.17}$$

and is proportional to the *square* of the illuminated (aperture) area. This is typical of diffraction phenomena, as compared to incoherent illumination, which is simply proportional to the area.

In the case of a long *vertical* slit, $\eta_0 \gg \zeta_0$, the intensity vanishes very rapidly for $v \neq 0$. (Note that ηv becomes large and the exponential in Eq. (5.15) oscillates rapidly, its average value tending to zero.) Thus we observe a *horizontal* diffraction pattern confined to the x' axis, as shown in Fig. 5.4. Exactly on the x' axis, $v = 0$ and Eq. (5.17) reduces to

$$I(x', y' = 0) = 16\zeta_0^2\eta_0^2 \left[\frac{\sin\left(\frac{2\pi}{\lambda}\zeta_0 u\right)}{\frac{2\pi}{\lambda}\zeta_0 u} \right]^2 = I_0 \left[\frac{\sin\left(\frac{\pi d}{\lambda}\sin\theta\right)}{\frac{\pi d}{\lambda}\sin\theta} \right]^2. \tag{5.18}$$

In the last step we made use of the relations $u = x'/R = \sin\theta$ (valid for $y' = 0$), where θ is the angle from the z axis and $\zeta_0 = d/2$; we also set $16\zeta_0^2\eta_0^2 = I_0$ to represent the intensity at $\theta = 0$. Note that the above result is exactly that given in Eq. (5.5).

We now consider the case where the amplitude of the incident wave is not constant over the aperture. Such variation can be introduced deliberately by placing a suitable mask over the aperture, or because the incident wave is modulated in magnitude and/or phase. In this case we express the amplitude on the aperture by

$$F(\zeta, \eta),$$

and the amplitude at the observation point P' (Eq. (5.16)) takes the form

$$A'(x', y') = \int\int F(\zeta, \eta) e^{i\frac{2\pi}{\lambda}(\zeta u + \eta v)} d\zeta d\eta. \tag{5.19}$$

The integration in Eq. (5.19) is over the aperture; however, since the amplitude of the transmitted wave vanishes beyond the aperture boundary, we can extend the integration limits to infinity. With this modification, we see that the far-field amplitude, i.e., in the focal plane of the lens, is the *Fourier transform* of the amplitude in the near field. To explain this statement note that Eq. (5.19) is very similar to the more familiar Fourier transform between the frequency and time domains. If $F(t)$ describes the time dependence of a pulse, then $A(\omega)$ describes the spectrum of the frequencies contained in the pulse

$$A(\omega) = \int_{-\infty}^{+\infty} F(t) e^{-i\omega t} dt. \tag{5.20}$$

Similarly, in Eq. (5.19) $F(\zeta, \eta)$ describes the spatial dependence in the aperture plane and $A'(x', y')$ can be thought of as describing the spectrum of "spatial frequencies" $2\pi u/\lambda$ and $2\pi v/\lambda$. We will make use of these concepts in Section 5.4.

5.4. DIFFRACTION FROM A CIRCULAR APERTURE

Instead of a slit we shall now use a circular aperture. Some skill is required in manufacturing such small apertures, but they can also be purchased commercially. The smaller the diameter, the larger the diffraction angle, but the transmitted intensity decreases at the fourth power of the diameter, making observation of the pattern correspondingly more difficult. In the present experiment the aperture diameter was $d = 150$ μm, which is a good compromise for the HeNe wavelength.

To obtain the diffraction pattern, we integrate Eq. (5.15) over the circular aperture. To do so we rewrite Eq. (5.15) in terms of polar coordinates as

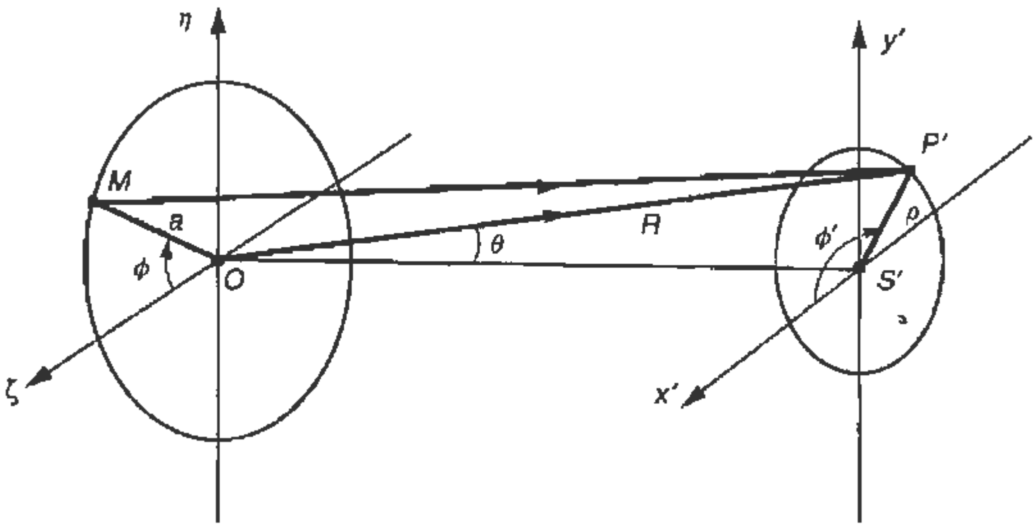


FIGURE 5.7 Coordinate systems in polar coordinates for calculating diffraction.

shown in Fig. 5.7. In the source plane we use the coordinates a, ϕ so that

$$\zeta = a\cos\phi \qquad \eta = a\sin\phi, \tag{5.21}$$

whereas in the observation plane we use ρ, ϕ' so that

$$u = \frac{x'}{R} = \frac{\rho\cos\phi'}{R} = \alpha\cos\phi', \qquad v = \frac{y'}{R} = \frac{\rho\sin\phi'}{R} = \alpha\sin\phi', \tag{5.22}$$

where $\alpha = \rho/R$ is the sine of the radial diffraction angle. Expressed in terms of these new coordinates, the argument of the exponential in Eq. (5.15) becomes

$$i\,\frac{2\pi}{\lambda}\,[\zeta u + \eta v] = i\,\frac{2\pi}{\lambda}\,a\alpha\cos(\phi - \phi'). \tag{5.23}$$

Since the origin of the angles is arbitrary we can set $\phi' = 0$ (the pattern must be rotationally symmetric about the axis). Thus the amplitude at the angle θ, where $\sin\theta = \alpha$, is given by

$$A'(\alpha) = \int_0^{a_0}\int_0^{2\pi} e^{i\frac{2\pi}{\lambda}a\alpha\cos\phi}a da\, d\phi. \tag{5.24}$$

Here a_0 is the radius of the circular aperture, and we have assumed uniform illumination.

The integral in Eq. (5.24) cannot be performed in terms of trigonometric functions but is well known. One finds that

$$A'(\alpha) = \pi a_0^2 \frac{2J_1\left(\frac{2\pi}{\lambda}a_0\alpha\right)}{\frac{2\pi}{\lambda}a_0\alpha}, \tag{5.25}$$

where J_1 is the Bessel function of order 1. The intensity is given by the square of the amplitude

$$I(\alpha) = \left(\pi a_0^2\right)^2 \left[\frac{2J_1\left(\frac{2\pi}{\lambda}a_0\alpha\right)}{\frac{2\pi}{\lambda}a_0\alpha}\right]^2 . \tag{5.26}$$

We recognize that the intensity is proportional to the *square* of the illuminated area. Since $\alpha = \sin\theta$ is the diffraction angle, Eq. (5.26) is similar to Eq. (5.5) with the replacement of the sine by the J_1 Bessel function.

Equation (5.26) is plotted as a function of its argument, $x = (2\pi/\lambda)a_0\alpha$, in Fig. 5.8. The zeros occur at the following values of x,

$$x_1 = 3.83, \qquad x_2 = 7.02, \qquad x_3 = 10.17, \text{ etc.},$$

whereas the maxima fall in between. The pattern is that of an intense central disk surrounded by alternating dark and bright rings, as shown in Fig. 5.9. The first dark ring occurs at an angle

$$\theta_1 \simeq \sin\theta_1 = \frac{3.83}{\pi}\frac{\lambda}{2a_0} = 1.22\frac{\lambda}{D}, \tag{5.27}$$

where D is the diameter of the aperture. If the lens used has a focal length f, the radius of the first dark ring on the screen occurs at

$$\rho_1 = 1.22\left(\frac{f}{D}\right)\lambda. \tag{5.28}$$

Equation (5.27), first obtained by Airy, gives the smallest radius that can be obtained by focusing a beam of wavelength λ with optics specified by the *f-number* (f/D). The shorter the focal length, and the larger the aperture, the smaller the focal spot and thus the resolution of the instrument. The central disk contains 76% of the total intensity.

The experimental setup is the same as shown in Fig. 5.3, except that the slit is replaced by the circular "pinhole." Figure 5.9 is a CCD picture obtained with a 150-μm diameter pinhole. Three dark rings could be

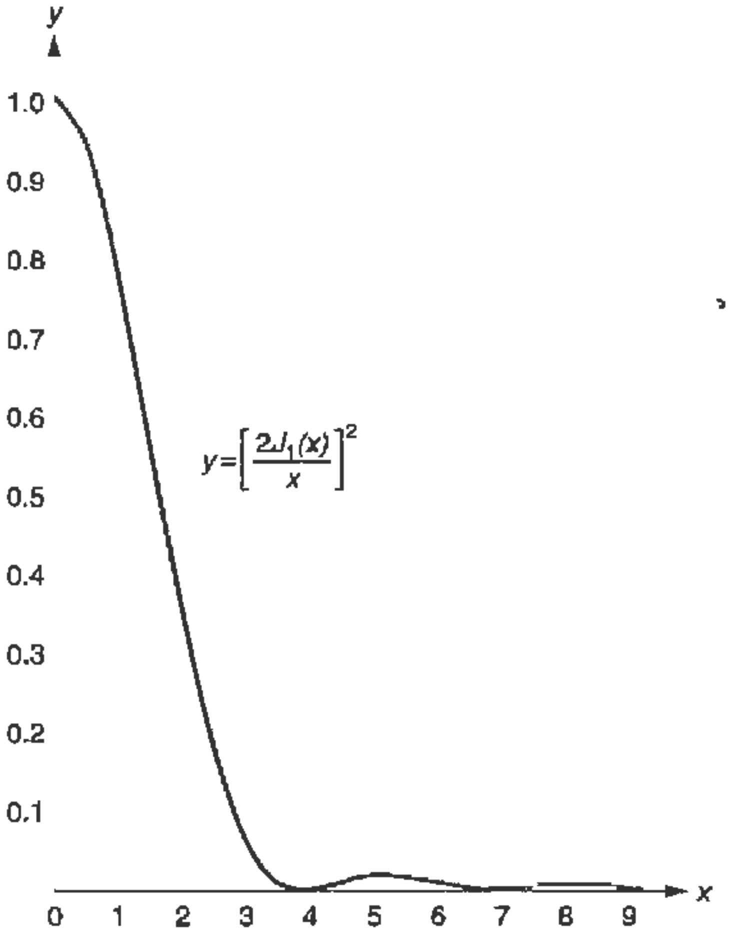


FIGURE 5.8 The intensity distribution for Fraunhofer diffraction from a circular aperture as a function of $x = (2\pi/\lambda)a_0 \sin\theta$; θ is the diffraction angle and a_0 the aperture radius.

measured at the angles

$$\alpha_1 = (5.25 \pm 1) \times 10^{-3} \text{ radians}$$

$$\alpha_2 = (10.5 \pm 2) \times 10^{-3} \text{ radians}$$

$$\alpha_3 = (14.5 \pm 2) \times 10^{-3} \text{ radians.}$$

Using the values for the zeros of J_1 as given previously, we obtain the corresponding values for λ/D

$$(\lambda/D) = 4.3 \times 10^{-3},\ 4.7 \times 10^{-3},\ 4.5 \times 10^{-3}.$$

These results are self-consistent and predict a pinhole diameter $D = 140\ \mu\text{m}$, in good agreement with the "nominal" value.

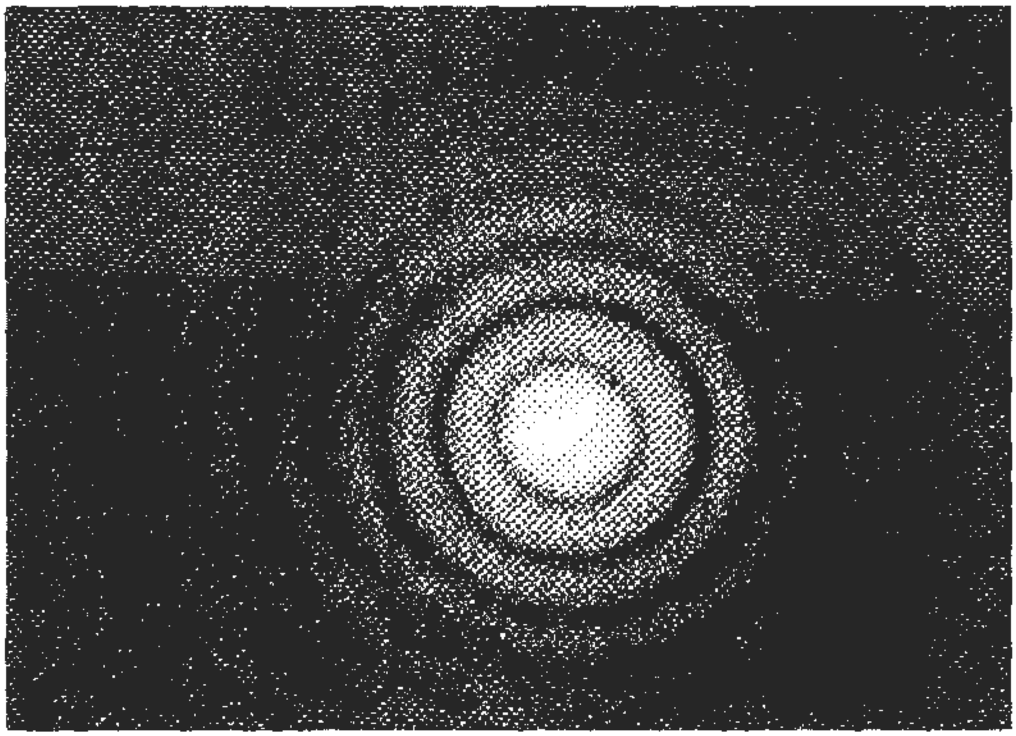


FIGURE 5.9 Observed diffraction pattern of a HeNe beam from a pinhole of diameter 150 μm.

5.5. THE DIFFRACTION GRATING

We have already made use of the diffraction grating and discussed the physical principles in Chapter 1. Here we will carry out a more detailed analysis and demonstrate a compact spectrometer using a digital readout.

If instead of a single slit, two slits are illuminated by a plane wavefront, a series of interference fringes parallel to the slits will appear on a far screen. This is the classical experiment of Thomas Young (1800) shown in Fig. 5.10a. If the spacing between the slits is d, the intensity distribution on the screen is[1]

$$I(\theta) = 4I_0 \cos^2\left(\frac{\pi d}{\lambda}\sin\theta\right). \tag{5.29}$$

The angle θ is measured, as usual, with respect to the normal to the plane containing the slits. If one of the slits is blocked, the fringes disappear and the transmitted intensity is I_0.

[1] We take the wavefront parallel to the plane in which lie the slits.

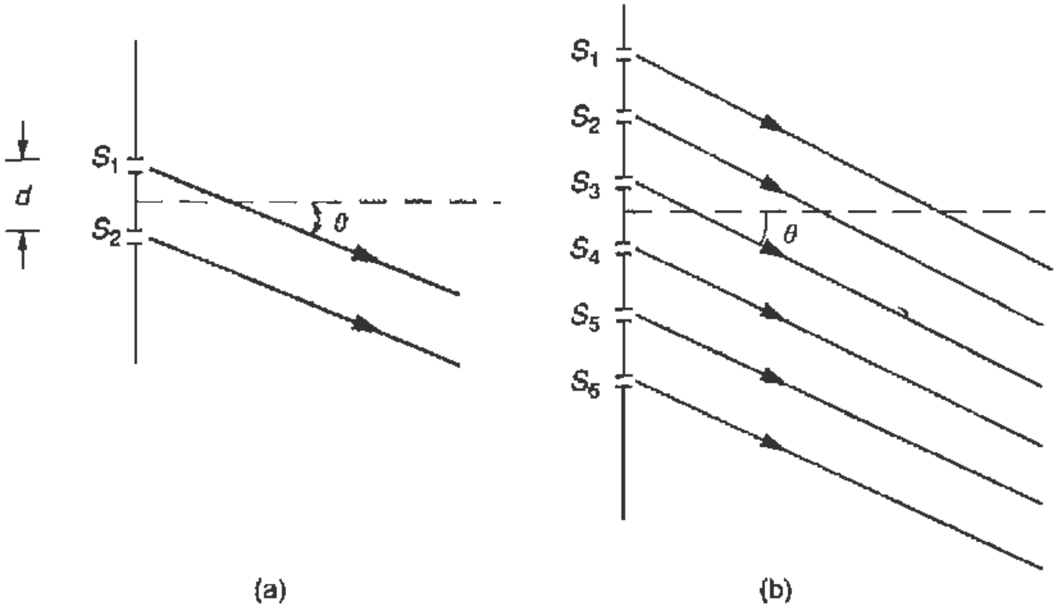


FIGURE 5.10 (a) Young's two-slit experiment. (b) Multiple-slit interference, the diffraction grating.

In Eq. (5.29) we have not included the effects of diffraction due to the width of the slits. Let the slit width be a. Then Eq. (5.29) is modulated by the diffraction pattern of Eq. (5.5), and we obtain for the intensity distribution

$$I(\theta) = 4I_0 \cos^2\left(\frac{\pi d}{\lambda}\sin\theta\right)\left[\frac{\sin\left(\frac{\pi a}{\lambda}\sin\theta\right)}{\frac{\pi a}{\lambda}\sin\theta}\right]^2. \tag{5.30}$$

If instead of two slits, several equidistantly spaced slits are illuminated by the wavefront, the interference maxima become much sharper, and the interference pattern is given by

$$I(\theta) = I_0\left[\frac{\sin\left(N\frac{\pi d}{\lambda}\sin\theta\right)}{\sin\left(\frac{\pi d}{\lambda}\sin\theta\right)}\right]^2. \tag{5.31}$$

Here d is the spacing between the slits, and N the total number of slits; we have not included the effects of diffraction because in practical applications the slits are so narrow that the modulation is not important. Note that Eq. (5.31) reduces to Eq. (5.29) for $N = 2$, as it must.

What is of particular interest is that the pattern contains *principal* maxima when the denominator of Eq. (5.31) becomes zero, namely when

$$\sin\theta = \pm n\lambda/d, \qquad n = 0, 1, 2, \ldots. \tag{5.32}$$

The intensity at the principal maxima can be found as follows. Near a principal maximum $(\pi d/\lambda)\sin\theta = n\pi + \epsilon$ and therefore $\sin[(\pi d/\lambda)\sin\theta] \simeq \epsilon$ so that Eq. (5.31) can be written as

$$I(\theta)_{\max} = I_0 \left[\frac{\sin[N(n\pi+\epsilon)]}{\epsilon}\right]^2 = I_0 N^2 \left[\frac{\sin(N\epsilon)}{N\epsilon}\right]^2 = I_0 N^2 \left(\frac{\sin^2 x}{x}\right), \tag{5.33}$$

where $x = N\epsilon = N\pi(d/\lambda)\Delta\theta$, and $\Delta\theta$ is the departure of θ from the condition of Eq. (5.32). Since the function $(\sin x/x)^2 \to 1$ as $x \to 0$, the intensity at the principal maxima $[\Delta\theta = 0]$ is

$$I_{\max} = N^2 I_0. \tag{5.34}$$

This pattern is shown in Fig. 5.11.

The width of the principal maxima is given by the first minimum of the function $(\sin x/x)$, which occurs when $x = \pm\pi$, namely

$$\Delta\theta = \pm\frac{\lambda}{Nd}. \tag{5.35}$$

Note that (Nd) is the total extent of the region covered by the slits. Thus the principal maxima are as narrow as if the wavefront diffracted from a slit of width Nd. By combining Eq. (5.35) with Eq. (5.32) we can express

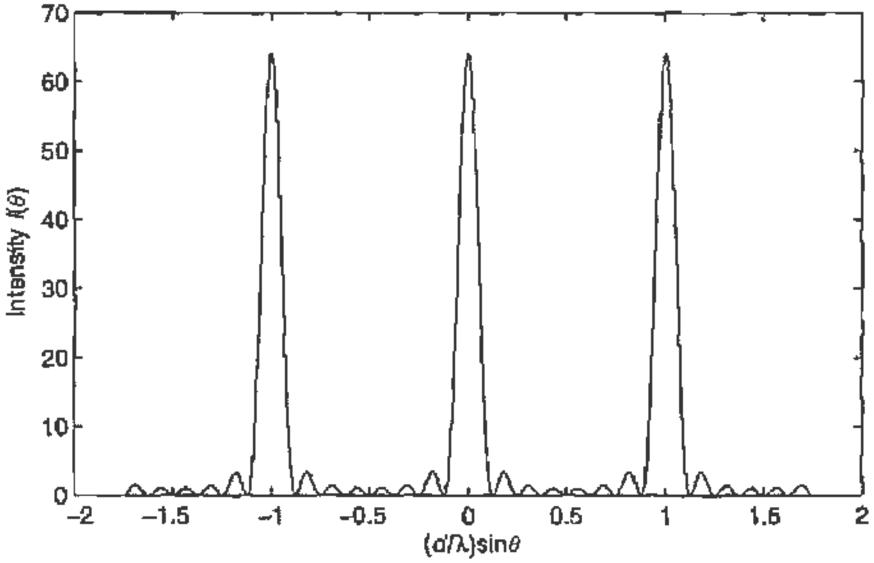


FIGURE 5.11 Different orders of monochromatic light scattered from a grating. Note that the principal maxima are very narrow peaks, whereas the secondary maxima are suppressed. Plotted for $N = 5$.

the resolution of the system of N slits by

$$\frac{\Delta\lambda}{\lambda} = \frac{1}{Nn}\cos\theta \simeq \frac{1}{Nn}. \tag{5.36}$$

A *diffraction grating* is equivalent to such a system of many slits and can be used either in transmission or in reflection. The angle of incidence θ_i can be different from the normal to the grating, in which case Eq. (5.32) must be modified to read

$$\sin\theta_i - \sin\theta_r = \pm n\,\frac{\lambda}{d}, \qquad n = 0, 1, 2, \ldots. \tag{5.37}$$

The diffraction angle is θ_r and is taken positive if it is *opposite* from θ_i with respect to the normal. These definitions are shown in Fig. 5.12. For a reflection grating $n = 0$ corresponds to specular reflection ($\sin\theta_r = \sin\theta_i$). Reflection gratings are often manufactured so as to enhance reflection at particular angles. Recall that Eq. (5.37) was already used in Chapter 1 (see Eq. (1.16)).

The arrangement used in the laboratory is shown in Fig. 5.13. The light source is focused on the slit and the emerging beam is made parallel by lens L1, which has focal length $f_1 = 20$ cm. The parallel beam is incident on the 4×4 cm^2 grating, which has 1200 lines/mm. The angle of incidence was chosen to be $\theta_i = 55.7°$. The beam diffracted in first order was focused with lens L2, identical to L1, onto the "reticon" where it formed an image of the slit.

The reticon is a linear array of pixels, which can be read out on an oscilloscope. In the present case the array contained 128 pixels; the clock speed was 80 kHz so that a pixel is read out every $\Delta t = 12.5$ μs. The pixel

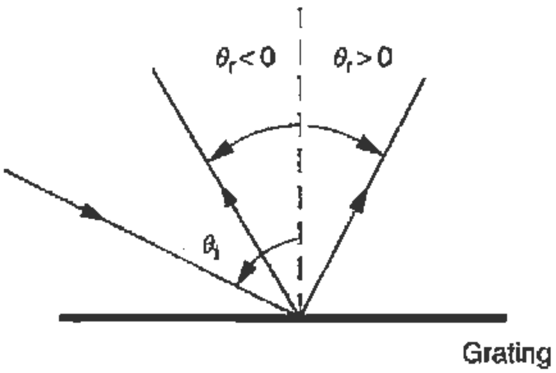


FIGURE 5.12 The convention used for labeling the incidence and reflection angles for a reflection grating.

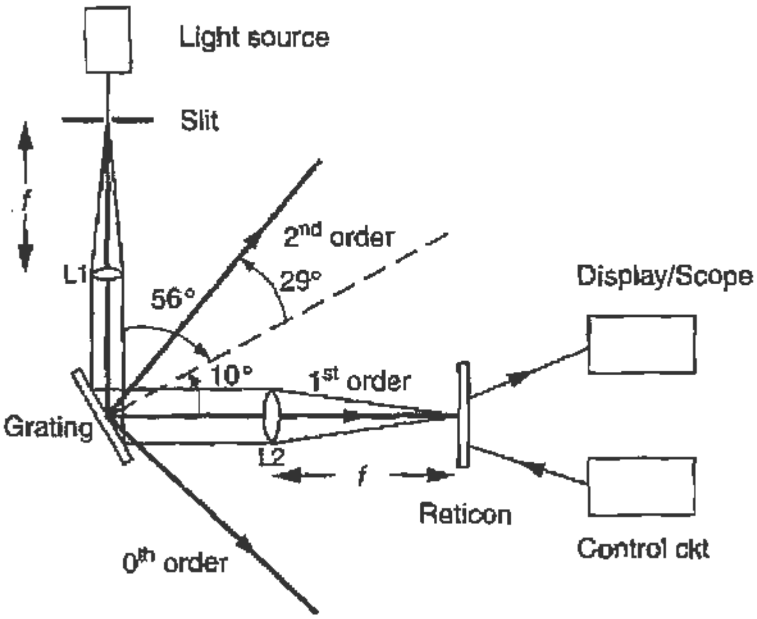


FIGURE 5.13 Layout of a simple grating spectrometer read out by a reticon (a one-dimensional solid-state detector array).

size was $\Delta x_0 = 100\ \mu\text{m}$ for a total array length of 1.28 cm. Thus we have the conversion factor

$$\Delta x = 8\ \mu\text{m}/\mu\text{s}, \tag{5.38}$$

and since $\Delta x = f\,\Delta\theta$, $f = 20$ cm

$$\Delta\theta = 0.04\ \text{mrad}/\mu\text{s}. \tag{5.39}$$

The spectrum of a Hg arc lamp is shown in Fig. 5.14a. The horizontal scale (sweep speed) corresponds to 200 μs/cm. The spectrum was observed in first order, and from Eq. (5.37) with $\theta_i = 55.7°$, $d = 10^{-3}/(1.2\times10^3)$ m we find that for the green line of Hg ($\lambda_g = 546.1$ nm)

$$\sin\theta_r = \sin\theta_i - \frac{\lambda}{d} = 0.170,$$

namely $\theta_r = 9.8°$ in the quadrant opposite to the incident beam. The second order appears in the same quadrant as the incident beam at $\theta_r = 29°$ as shown in Fig. 5.13.

The green line corresponds to the peak on the right-hand side of the graph (Fig. 5.14a), whereas the doublet on the left corresponds to the yellow lines ($\lambda_1 = 577.7$ nm and $\lambda_2 = 579.1$ nm). Knowledge of these wavelengths allows us to make a more precise calibration of the spectrometer, including

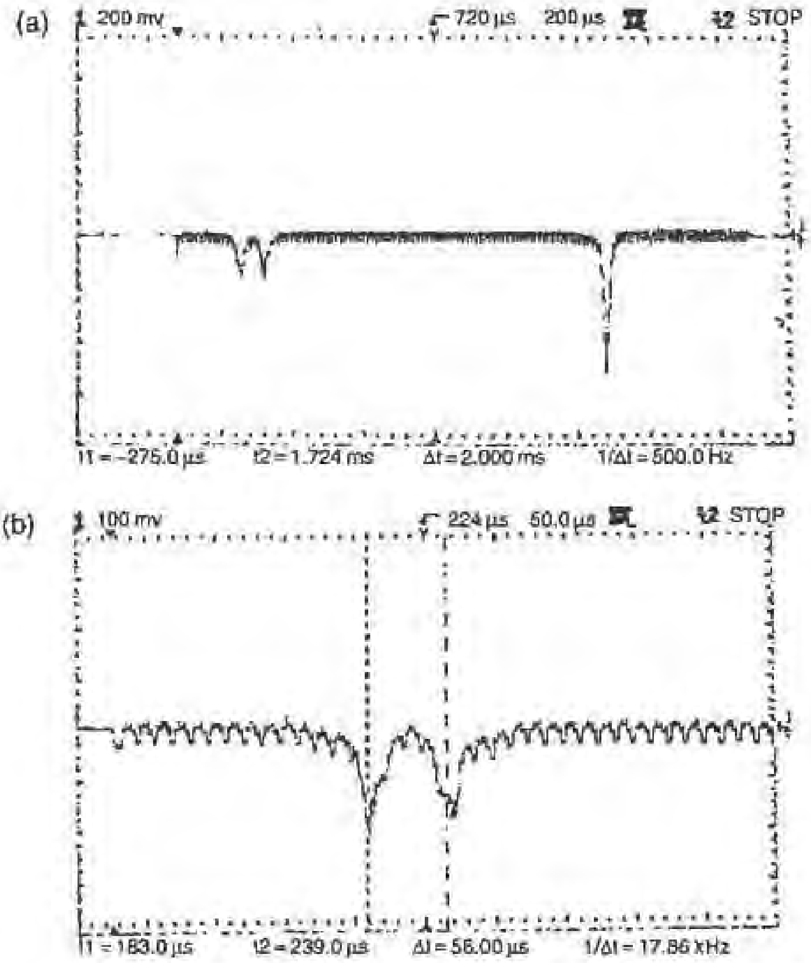


FIGURE 5.14 The observed spectrum (a) of the green line and yellow doublet of the Hg spectrum obtained with the spectrometer of Fig. 5.13b. (b) The yellow doublet on an expanded scale.

misalignment and other instrumental effects. Differentiating Eq. (5.37) with θ_i fixed, we obtain

$$n\Delta\lambda/d = -\cos\theta_r \Delta\theta_r. \tag{5.40}$$

In our case $n = 1$, $\cos\theta_r = 0.99$ and $\Delta\lambda$, from the first yellow line λ_2 to the green line λ_g, is $\Delta\lambda = 33$ nm, or $\Delta\theta = 40.0$ mrad. The time interval between these lines as measured off Fig. 5.14a is $\Delta t = 1030$ µs, and thus the calibration

$$\Delta\theta = 39 \times 10^{-3} \text{ mrad}/\mu\text{s} \tag{5.41}$$

in close agreement with the direct calculation.

To measure the fine structure of the yellow doublet the sweep speed is increased so that the scale factor is 50 µs/cm as shown in Fig. 5.14b.

One can now recognize the response of individual pixels. The separation of the two lines is 56 μs; using Eq. (5.41) we find $\Delta\theta = 2.18$ mrad and from Eq. (5.40)

$$\Delta\lambda = 1.8 \text{ nm}.$$

Our result is only in modest agreement with the accepted value of $\Delta\lambda = 1.4$ nm. This is not surprising because one pixel, the ultimate resolution of our detector in this configuration, contributes an uncertainty of $\delta\lambda = 0.42$ nm. Thus one must be cautious when using digital techniques, which often do not have the advantages of the high resolution of photographic film or of visual observation.

5.6. FOURIER OPTICS

In Eq. (5.19), we showed that the amplitude of the electric field in the focal plane of a lens is the Fourier transform of the near-field amplitude incident on the lens. We will now give a physical discussion of this result and show how it can be used in practice. These considerations were first introduced by E. Abbe in Jena, Germany, but found much wider use as lasers became available.

A transmission grating is a repetition of regions in space that alternatively transmit/absorb the incident wavefront; we can represent the transmission of the grating by the "square-wave" function shown in Fig. 5.15a. We are immediately reminded of the analogous square-wave function of time that has period T, and thus frequency $\nu = 1/T$. Therefore we can assign to the

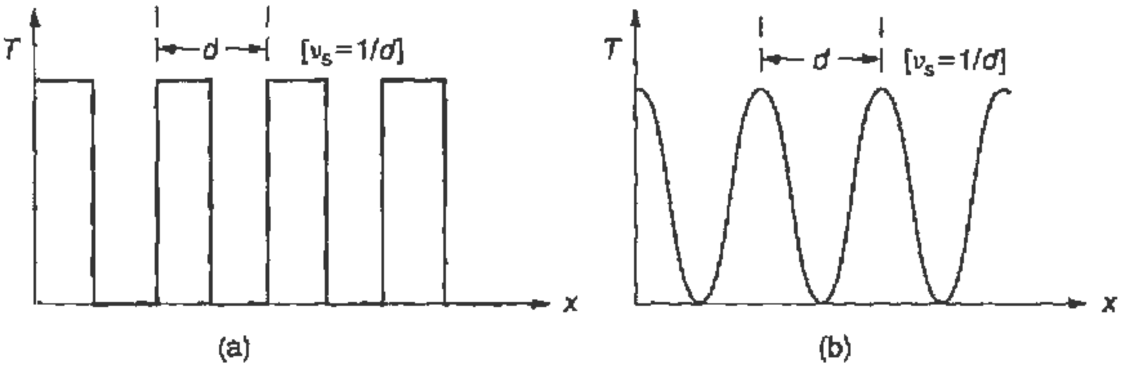


FIGURE 5.15 (a) Representation of the transmission of a grating; the spatial spectrum contains the fundamental frequency $\nu_s = 1/d$ and its higher harmonics. (b) If the transmission is sinusoidal, only the frequency $\nu_s = 1/d$ is present in the scattered wave.

grating a *spatial period d* and a *spatial frequency* $1/d$. Spatial frequency is measured in cycles per unit length and has dimensions of inverse length. For instance, the grating used in the experiment described in the previous section has a spatial frequency of 1200 lines/mm. From circuit theory we know that a square pulse in time contains the fundamental frequency as well as higher harmonics. Similarly the square grating contains not only the fundamental spatial frequency $1/d$, but also its harmonics n/d. This is seen when light incident on the grating is diffracted at the angles θ, with

$$\sin\theta_n = n\,\frac{\lambda}{d}.$$

If the grating profile was sinusoidal, as in Fig. 5.15b diffraction would occur only for $n = 0$ and $n = 1$.

We can place a lens after the grating to relocate the far field into the (back) focal plane of the lens as shown in Fig. 5.16. We will then see the diffraction maxima, namely the Fourier transform of the grating: we refer to this plane as the *transform plane*. If the distance s_1 from the grating to the lens exceeds the focal length f, an image of the grating will be formed in the *image plane* located at s_2, where

$$\frac{1}{s_1} + \frac{1}{s_2} = \frac{1}{f}.$$

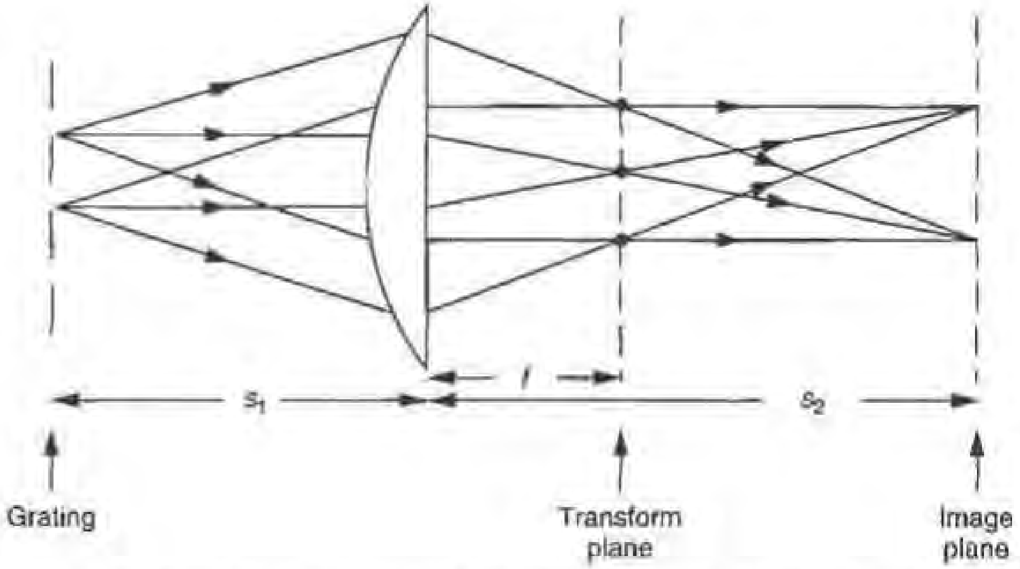


FIGURE 5.16 Location of the source plane, the transform plane (the back focal plane of the lens), and the image plane.

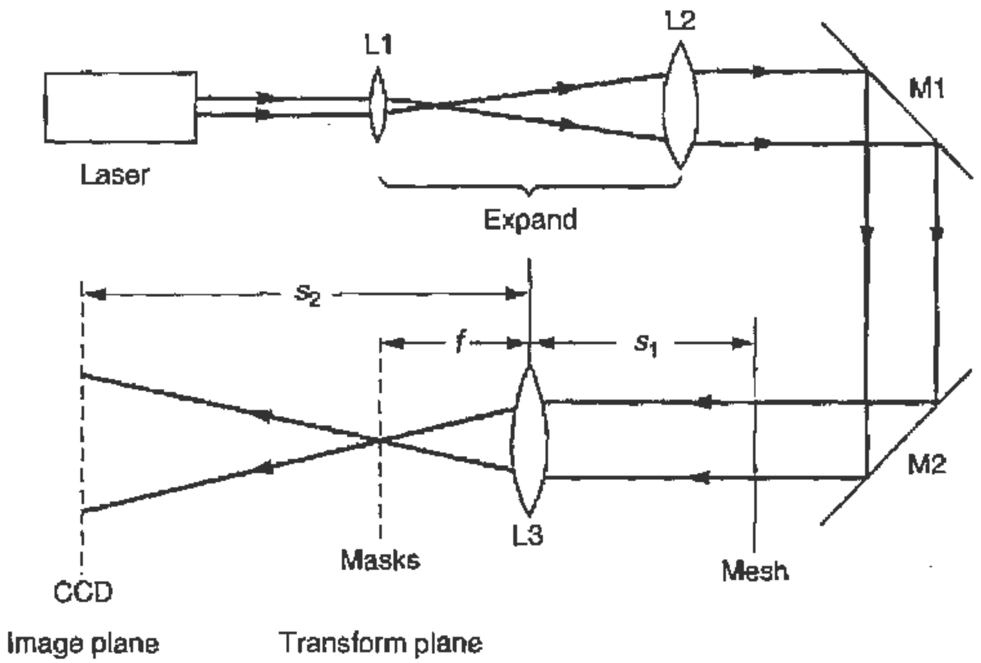


FIGURE 5.17 Experimental layout for demonstrating Fourier optics.

This image is the Fourier transform of the amplitude in the transform plane. Therefore, by altering the pattern in the transform plane, we can *modify* the image being formed. There are several applications of this principle, as for instance in smoothing out images that contain noise or in pattern recognition.

A simple demonstration of Fourier optics can be carried out in the laboratory with the setup shown in Fig. 5.17. The laser beam is expanded and allowed to illuminate a mesh with 270 lines/in. and transmission factor ~50%. Lens L3 is used to image the grating onto a CCD camera. Various masks are then inserted in the focal plane of the lens, the transform plane, to modify the image.

The results are shown in Fig. 5.18. In Fig. 5.18a, no obstacle is in the transform plane, and the pattern represents the image of the mesh. Next, a *vertical* slit 1.5 mm wide is placed in the transform plane, and the pattern in the image plane contains horizontal stripes as shown in Fig. 5.18b. The effect of the mask is to allow passage only of components of the wavefront that were dispersed vertically in the transform plane. These components carry the information about the horizontal structure of the object (the mesh) and thus show horizontal lines in the image plane. Figure 5.18c was obtained with a horizontal slit as the mask in the transform plane.

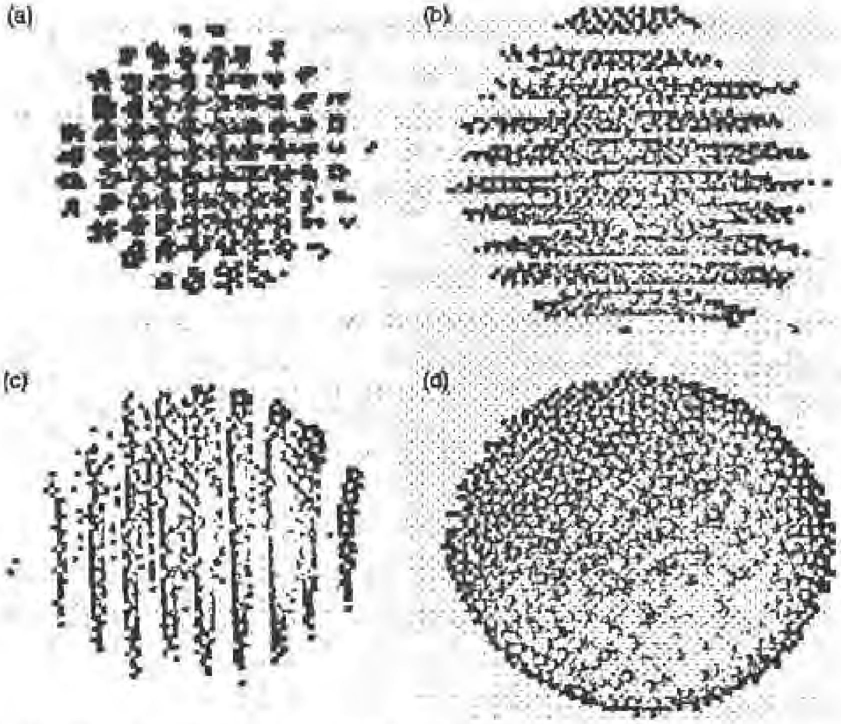


FIGURE 5.18 Results from placing masks in the transform plane: (a) Image of a square mesh in the absence of a mask, (b) placing a vertical slit in the transform plane, (c) placing a horizontal slit in the transform plane, and (d) placing a pinhole in the transform plane.

Finally Fig. 5.18d shows the result of placing a 1-mm-diameter pinhole in the transform plane. Now all high spatial frequencies are filtered, and the pattern in the image plane is significantly smoothed out.

Spatial filtering by using a pinhole is often used to "clean" laser beams that have acquired structure due to imperfect optics, dust on components, and other aberrations. This is analogous to using a capacitor to filter out high-frequency noise in an electric circuit.

5.7. THE FARADAY EFFECT

5.7.1. Discussion

As already mentioned, the Faraday effect refers to the rotation of the plane of polarization when light propagates through certain media subject to an axial magnetic field. It was discovered in 1845 by Faraday long before the nature of light or matter was understood. We now know that the electric field of light is transversely polarized with respect to its direction of propagation, z, and we can express it, in exponential notation, as

$$\mathbf{E}(z,t) = \Re e\left\{E_0 e^{-i(\omega t - kz)}\mathbf{e}\right\}. \tag{5.42}$$

Here $\Re e$ means to take the real part of the expression; for simplicity of notation we will omit this designation in what follows but it is always implied. As usual $\omega = 2\pi\nu$ and $k = 2\pi/\lambda$. $\mathbf{e}$ is the polarization vector, which can be expressed in terms of *two* unit vectors (since $\mathbf{e}$ is restricted to the x, y plane). We can choose linearly polarized unit vectors

$$\mathbf{e}_1 = \mathbf{u}_x, \qquad \mathbf{e}_2 = \mathbf{u}_y \tag{5.43}$$

or circularly polarized unit vectors

$$\mathbf{e}_\mathrm{R} = \mathbf{u}_x + i\mathbf{u}_y, \qquad \mathbf{e}_\mathrm{L} = \mathbf{u}_x - i\mathbf{u}_y. \tag{5.44}$$

If we now examine the electric field at a fixed position z, in the case of circular polarization we will have the two components

$$\begin{aligned} E_\mathrm{R} &= E_0[\cos\omega t\,\mathbf{u}_x + \sin\omega t\,\mathbf{u}_y] \\ E_\mathrm{L} &= E_0[\cos\omega t\,\mathbf{u}_x - \sin\omega t\,\mathbf{u}_y]. \end{aligned} \tag{5.45}$$

These were obtained by introducing Eqs. (5.44) into Eq. (5.42). The fields rotate in the transverse plane, in the first case according to the right-hand rule (with the thumb along the direction of propagation), in the second case according to the left hand. This is shown in Fig. 5.19 where we use a

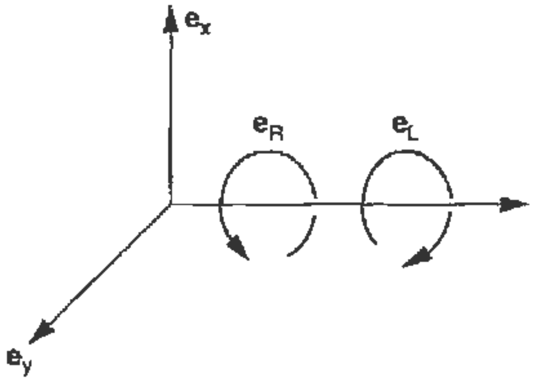


FIGURE 5.19 The right-handed coordinate system used to define right- and left-circular polarization.

right-handed coordinate system. Note that we can write Eqs. (5.45) as

$$E_\mathrm{R} = E_x + iE_y \qquad E_\mathrm{L} = E_x - iE_y \tag{5.46}$$

and by solving

$$E_x = \frac{1}{2}(E_\mathrm{R} + E_\mathrm{L}) \qquad E_y = \frac{1}{2i}(E_\mathrm{R} - E_\mathrm{L}). \tag{5.47}$$

The Faraday effect arises because in certain materials the application of a magnetic field results in *different* refractive indices for the right and left circularly polarized light propagating along the direction of the field. Materials that have a different refractive index for two given polarization orientations are called birefringent. The birefringence is natural in certain crystals or can be induced by the application of an electric field (Pockels effect).

The physical interpretation of the Faraday effect is related to the shift of the atomic energy levels when an external magnetic field is applied. This is the Zeeman effect, which is discussed in some detail in the following chapter. When the light propagates along the axis of the field the right polarized light can excite only a particular set of sublevels ($\Delta m = +1$, where m is the magnetic quantum number) and conversely for the left polarized light ($\Delta m = -1$). These levels have different excitation energy and this results in different refractive indices, n_R and n_L. For more details the reader should consult the references cited at the end of the chapter.

We know that the velocity of propagation of the wave, the phase velocity, is given by $c' = c/n$; thus the phase advance in a length L of material is

$$\theta = kL = \frac{2\pi}{\lambda}L = \frac{2\pi\nu}{c'}L = \frac{2\pi\nu}{c}nL, \tag{5.48}$$

where the frequency ν of the light is fixed and n is the refractive index of the material. Thus the right and left polarized light will acquire different phases. If the incident light was linearly polarized when entering the material, say along x, E_R and E_L would have the same phase (see Eq. (5.47)). However, upon exiting the material their relative phase would be shifted and the light, while still linearly polarized, would also contain a small E_y component. Namely it will have rotated by an angle

$$\phi = \frac{1}{2}(\theta_\mathrm{R} - \theta_\mathrm{L}) = \frac{\pi\nu}{c}L(n_\mathrm{R} - n_\mathrm{L}). \tag{5.49}$$

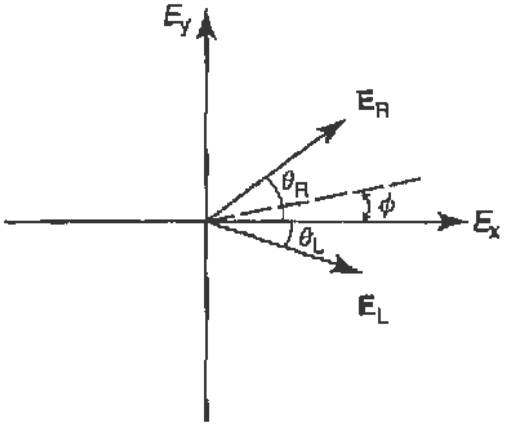


FIGURE 5.20 The rotation of the plane of linear polarization after propagation in a region where the right- and left-handed circular components have different phase velocity. Because $\mathbf{E}_R$ and $\mathbf{E}_L$ rotate by different amounts, the plane of linear polarization for $\mathbf{E} = \mathbf{E}_R + \mathbf{E}_L$ rotates away from the x axis by an angle $\phi = (\theta_R - \theta_L)/2$.

TABLE 5.1 Verdet Constant for Distilled Water

λ (nm)	C_V (rad/T-m)	
590	3.81	(Na *D*-lines)
600	3.66	
800	2.04	
1000	1.28	
1250	0.84	

This is shown in Fig. 5.20. The change in the refractive index is proportional to the external magnetic field, B, so that we can write

$$\phi = C_V B L, \tag{5.50}$$

where C_V is called the *Verdet constant*. We expect C_V to be a function of wavelength, as well as of the medium. Values for distilled water at various wavelengths[2] are listed in Table 5.1.

5.7.2. Procedure and Analysis

It is difficult to generate axial magnetic fields in the kilogauss range. Instead, a small but oscillating magnetic field will be used. The size of

[2] Data from E. U. Condon and H. Odishaw (Eds.), *Handbook of Physics*, second ed., McGraw-Hill, New York, 1967.

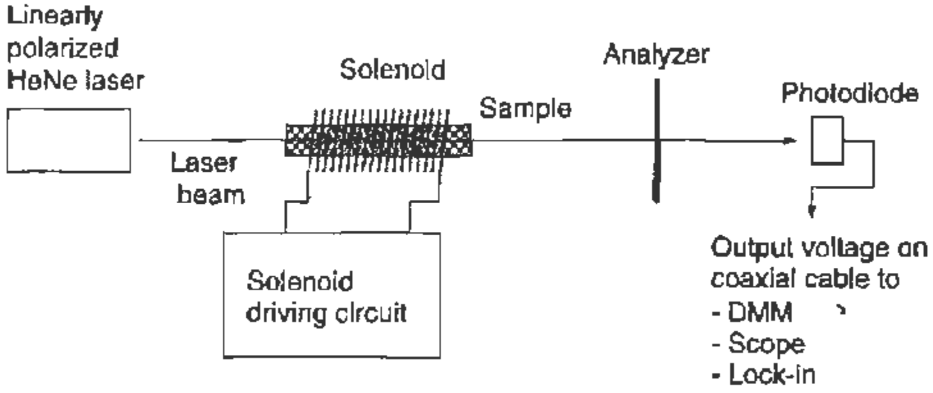


FIGURE 5.21 Experimental setup used for the Faraday effect. The photodiode output goes to the DMM for the polarization calibration, and to the oscilloscope or lock-in to measure the Verdet constant.

the effect will be small, but the oscillations make it possible to pick it out of the noise, by using lock-in detection.

The experimental setup is shown in Fig. 5.21. The source of polarized light is a HeNe laser. The magnetic field is supplied by a 1026-turn solenoid driven by the amplified signal of a waveform generator, in series with a monitor resistor. After passing through the sample and polarization analyzing filter, the light is detected in a photodiode. The signal is measured by the output voltage of the photodiode, and is given by

$$I(\phi) = I_0 \cos^2 \phi, \tag{5.51}$$

where ϕ is the angle of the linear polarization with respect to the analyzer axis. We are interested in $d\phi/dt$ and in this case the sensitivity is maximized by "biasing" the polarizer at $\phi_0 = 45°$, where $\phi = \phi_0 + \phi(t)$. Note that

$$\frac{dI(\phi)}{dt} = \frac{d\phi}{dt}\,\frac{dI}{d\phi} = -\frac{d\phi}{dt}\,I_0 \sin 2\phi \simeq -\frac{d\phi}{dt}\,I_0 \sin 2\phi_0. \tag{5.52}$$

We can calibrate the polarization analyzer by recording the photodiode voltage as a function of the analyzer angle. The result is shown in Fig. 5.22 and exhibits the $\cos^2 \phi$ dependence of Eq. (5.51). The maximum sensitivity $dV_{\mathrm{D}}/d\phi$ is found near $\phi = 180°$ and as predicted by Eq. (5.52) equals $V_{\mathrm{D}}^{\mathrm{Max}}$, namely $dV_{\mathrm{D}}/d\phi \approx 0.4$ V/rad.

The magnetic field is provided by the 1026-turn solenoid coil around the sample, driven by a sinusoidally varying current. The current is provided by an HP3311A waveform generator (sine wave, 600 Ω output) amplified by a Bogen MU10 monaural audio amplifier. The driver setup

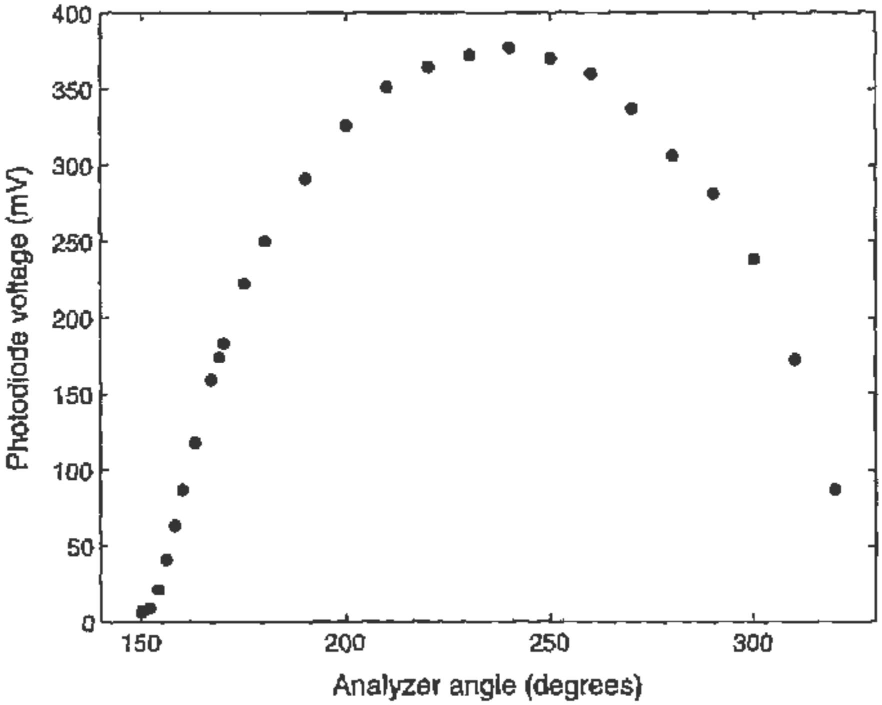


FIGURE 5.22 Sample polarization calibration data. The plot shows the full range of angles.

is shown in Fig. 5.23. The wave generator provides the input to the audio amplifier, and the output loops through the solenoid coil with a high-power resistor R_{coil} in series. The current and thus the magnetic field are determined by measuring the voltage drop across this resistor. *Do not ground either side of the amplifier output signal.* Using clip leads on a coaxial cable measure the voltage V_{coil} across R_{coil} on an oscilloscope. The shape should be a good sine wave with no DC offset and amplitude on the order of 10 V peak to peak. This is achieved by adjusting the amplitude of the HP3311A and the amplification (i.e., "volume") of the audio amplifier appropriately. It may be necessary to adjust the distortion on the amplifier so that the shape is alright.

The photodiode output is now connected to the other channel of the oscilloscope. The scope trigger is set to fire on coil voltage, and both channels are viewed simultaneously. If the channel on which V_D is measured is DC-coupled, one sees a large DC level, corresponding to the mean light intensity on the photodiode. (This DC level should agree with what was measured with the DMM.) The Faraday effect, on the other hand, shows up as a small oscillation on top of this DC level, in time with the V_{coil}. One is just able to see this small oscillation if the channel sensitivity is set to

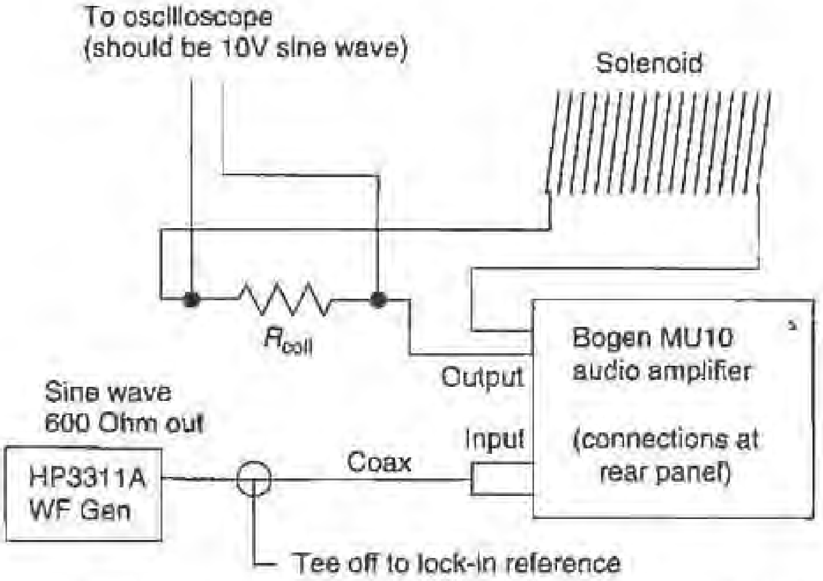


FIGURE 5.23 The driver circuit used to generate the oscillating magnetic field for measurement of the Faraday effect.

its lowest scale and AC-coupled to the input so that the large DC level is removed. Confirm that the amplitude of these small oscillations move up or down with the amplitude of V_{coil}, which is best adjusted by changing the amplifier gain. Confirm also that the oscillations disappear if the photodiode is blocked from the laser. In fact, the amplitude of the oscillations should change (and the phase reverse) as the analyzer is rotated.

We can now check that we are getting about the right Verdet constant, although it is hard to do a careful job with the small signal on the oscilloscope. From Eq. (5.50), we know that the small changes in polarization angle $\Delta\phi$ are related to the changes in magnetic field ΔB through

$$\Delta\phi = C_{\text{V}} \cdot \Delta B \cdot L_{\text{sample}}, \tag{5.53}$$

and from the calibration, we can convert $\Delta\phi$ to a change in photodiode voltage ΔV_{D} through

$$\Delta\phi\,\frac{dV_{\text{D}}}{d\phi} = \Delta V_{\text{D}}. \tag{5.54}$$

The magnetic field in a solenoid of length L_{solenoid} and $N = 1026$ turns is given by

$$B = \mu_0 i_{\text{coil}} N / L_{\text{solenoid}} \tag{5.55}$$

when a current i_{coil} passes through the coil. By combining Eqs. (5.53)–(5.55), one obtains an expression for the Verdet constant C_V in terms of V_D, V_{coil}, and other quantities that you know or can measure separately.

Consistent definitions should be used for V_{coil} and for V_D. That is, if V_{coil} is the amplitude of the sine wave, we make sure to do the same for V_D.

5.7.3. Results Using the Lock-In

The lock-in amplifier allows us to measure oscillations in V_D more precisely than with the oscilloscope. Furthermore, the lock-in will remove any noise that is out of phase or is at the wrong frequency. Refer to Section 3.8 for an explanation of lock-in detection.

The lock-in is a PARC Model 120 with a fixed reference frequency of ~100 Hz. It is best used by defining the reference wave externally, but it needs to be close to 100 Hz so that the internal circuit responds correctly. The lock-in mode dial is set to "SEL.EXT." and the HP3311A to a frequency near 100 Hz; using a BNC Tee connector the reference input is applied to the lock-in, while the signal is on the way to the audio amplifier. This assures us that we are using a reference signal with precisely the same frequency as the Faraday effect signal in V_D. The photodiode output should be connected to the lock-in input.

One still needs to tune the phase of the lock-in amplifier so as to have maximum sensitivity to the oscillating V_D signal. There are a few ways to do this, but the most instructive is to use the oscilloscope.

1. With the oscilloscope still triggered on the V_{coil} signal, use the other channel to view the "monitor out" port of the lock-in, with the switch set to "OUT × 1," which is the basic output signal of the lock-in. If the time constant is set to a value much smaller than $(100 \text{ Hz})^{-1}$ (1 ms will do), then you should just get the sine wave folded with the reference signal oscillating between ± 1. That is, it should look pretty much like Fig. 3.37 or Fig. 3.38, or something in between, depending on the phase setting.
2. Adjust the phase knob so that it looks like Fig. 3.37, that is, symmetric about the cusps, and with the cusp points at ground level. If you flip the relative phase quadrant knob so that the phase is 90° lesser or greater, the trace should look like Fig. 3.38. On the other hand, it should change sign if you flip by 180°.
3. With the phase adjusted so the output looks like Fig. 3.37, turn the time constant up to 1 s or so. You can read the monitor out on the

DMM, or use the meter on the lock-in. It is probably a good idea to block the light to the photodiode, and adjust the zero-trim so that the lock-in output is 0.

Vary V_{coil} by adjusting the audio amplifier gain. (You should not touch the wavcform generator settings anymore, since it is now serving a dual role as both amplifier input and lock-in reference.) Make a table of V_D as measured with the lock-in and V_{coil}. *Realize that the value of V_D provided by the lock-in is the RMS value, i.e., $1/\sqrt{2}$ times the amplitude.* Plot V_D versus V_{coil} and make sure you get a straight line through 0. Either fit to find the slope or average your values of V_D/V_{coil} to determine the Verdet constant with an uncertainty estimate.

Results obtained by a student are shown in Fig. 5.24 for a water sample. The parameters used to obtain these data were

$$\begin{aligned} R_{\text{coil}} &= 5.3\ \Omega \\ N &= 1026 \\ L_{\text{solenoid}} &= 0.265\ \text{m} \\ L_{\text{sample}} &= 0.265\ \text{m}. \end{aligned}$$

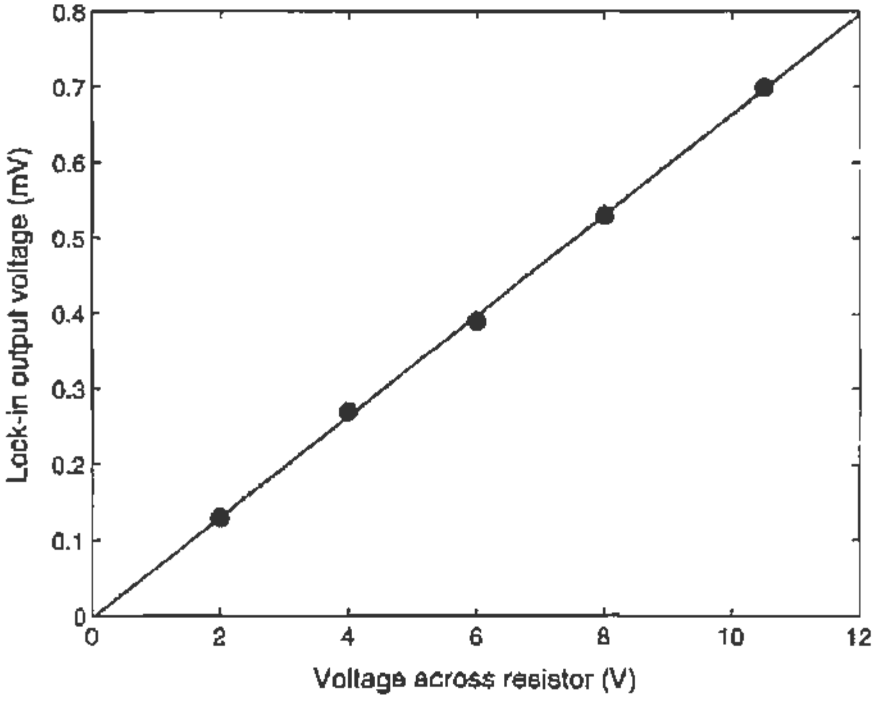


FIGURE 5.24 Results on the Faraday rotation angle as a function of magnetic field, obtained by a student.

We first calculate the magnetic field as a function of V_{coil}

$$B = \mu_0 \frac{V_{\text{coil}}}{R_{\text{coil}}} \frac{N}{L_{\text{solenoid}}} = V_{\text{coil}} \times (9.18 \times 10^{-4})\ \text{T}.$$

Next we use the relation of the optical rotation to V_{D}, which in this case was

$$\phi = V_{\text{D}}/(0.098)\ \text{rads}.$$

The measured values (see Fig. 5.24) are

$$V_{\text{D}}/V_{\text{coil}} = (6.7 \pm 0.52) \times 10^{-5}.$$

Thus we find for the Verdet constant

$$\begin{aligned} C_{\text{V}} &= \frac{\phi}{BL_{\text{sample}}} = \frac{1}{L_{\text{sample}}} \left(\frac{V_{\text{D}}}{V_{\text{coil}}} \right) \frac{1}{(9.8 \times 10^{-2})(9.18 \times 10^{-4})} \\ &= 2.80 \pm 0.2\ \text{rad/T-m}. \end{aligned}$$

From Table 5.1, extrapolating to $\lambda = 633$ nm, we would expect $C_{\text{V}} \approx 3.2$ rad/T-m. The difference could be accounted for in part by the short length of the solenoid, which results in a weaker field than what we calculate.

5.8. BERRY'S PHASE

We will demonstrate this effect by the rotation of the polarization vector of a beam of light, as in the Faraday effect, but in the present case the light propagates in a vacuum. The reason for the rotation of the polarization is that the propagation vector of the light, the k vector, performs a closed circuit around its direction of propagation. This is shown in Fig. 5.25 where light propagates from point A to point B. In part (a) of the figure the k vector describes a helix on its way, namely a closed loop in the transverse plane; therefore the polarization rotates. In examples (b) and (c) the initial and final values of **k** are the same as in example (a) but there is no looping around the direction of propagation; therefore the polarization does not rotate. We speak of a "topological" change in phase because the effect depends on the path followed while the initial and final points (in phase space) are the same.

This effect was first predicted by M. V. Berry in his 1984 paper (see Section 5.9). He analyzed the behavior of a quantum mechanical wave

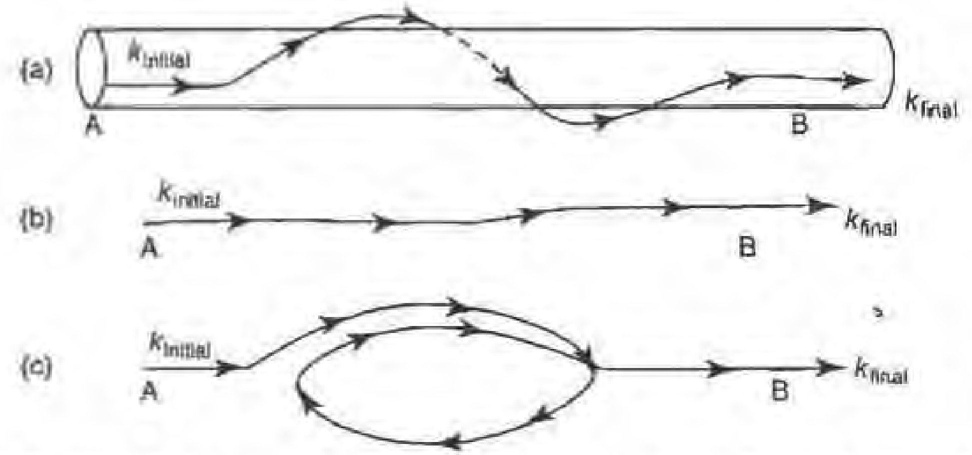


FIGURE 5.25 Topology of the optical fiber between A and B with $k_{\text{final}} = k_{\text{initial}}$: (a) helical winding, (b) direct (straight line path), and (c) circular path on a flat surface.

function when a parameter on which the wave function depends is slowly varied over a closed circuit. He showed that the wave function can acquire an extra phase factor even though the final state is identical to the initial state. It was soon realized that the same results should also hold for the electric field (the wave function) of a beam of light. Thus, the extra phase appears in classical as well as quantum-mechanical systems. In fact the precession of the Foucault pendulum or the Bohm–Aharonov effect can be interpreted as manifestations of Berry's phase.

When the k vector of light is transported through a closed circuit subtending a solid angle $\Delta\Omega$ at the origin, the right polarized light acquires a phase factor

$$E_{\mathrm{R}f} = e^{i\Delta\Omega} E_{\mathrm{R}i}, \tag{5.56}$$

whereas the left polarized light acquires a phase factor

$$E_{\mathrm{L}f} = e^{-i\Delta\Omega} E_{\mathrm{L}i} \tag{5.57}$$

where i and f refer to the initial and final state. This is a consequence of Maxwell's equations, which require that the k vector and the two polarization vectors always form an *orthogonal triad*.

To become convinced about this statement we show in Fig. 5.26 the unit sphere on which we can indicate the *directions* of $\mathbf{k}$, $\mathbf{e}_1$, and $\mathbf{e}_2$. Suppose we start from point A on the sphere and parallel transport the triad to point B along the equator. We then parallel transport it to point C along a great circle and return to point A by the corresponding great circle. At each point

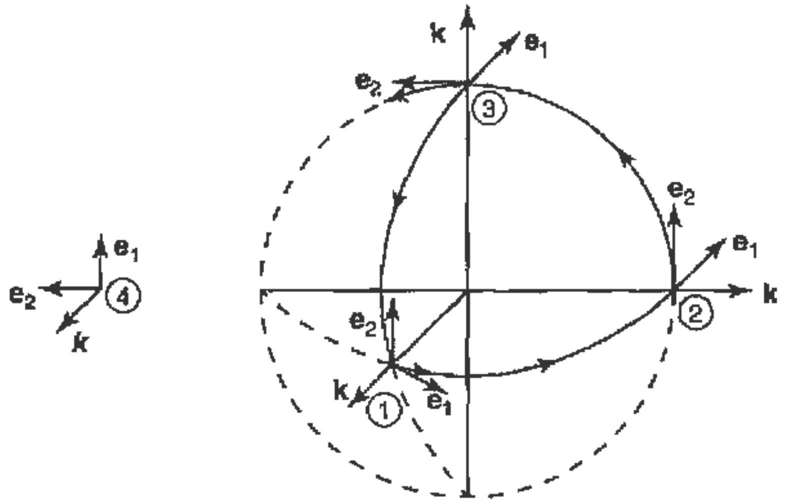


FIGURE 5.26 Parallel transport of the triad of orthogonal vectors $\mathbf{k}, \mathbf{e}_1, \mathbf{e}_2$ along the equator and two great circles. Note that $\mathbf{k}$ returns to its initial position but $\mathbf{e}_1$ and $\mathbf{e}_2$ are rotated by 90°. The solid angle enclosed by the path is 90°.

we have shown the orientation of the triad, and it is evident that upon return to A, the $\mathbf{k}$ vector has not changed but the $\mathbf{e}_1$ and $\mathbf{e}_2$ polarization vectors have been rotated by 90°. The solid angle subtended by the path that we followed is 1/8 of 4π or $\pi/2 = 90°$ equal to the observed rotation of $\mathbf{e}_1$, $\mathbf{e}_2$.

Let us now assume that the incident light is linearly polarized along the x axis. From Eq. (5.47) we can write

$$E_{\text{in}} = E_x = \tfrac{1}{2}(E_{\text{R}} + E_{\text{L}}).$$

After completing the circuit, we will have according to Eqs. (5.56) and (5.57)

$$E_f = \tfrac{1}{2}\left(E_{\text{R}}e^{i\Delta\Omega} + E_{\text{L}}e^{-i\Delta\Omega}\right).$$

However, this corresponds to linearly polarized light at an angle

$$\phi = \tfrac{1}{2}\,[\Delta\Omega - (-\Delta\Omega)] = \Delta\Omega \tag{5.58}$$

with respect to the x axis. The argument is exactly the same as that used in Fig. 5.20.

To carry out the experiment we must find a way to adiabatically change the orientation of the k vector. This can be done most conveniently by injecting the light into an optical fiber and then laying out the fiber on the desired path. One must use a *single-mode* fiber in order to preserve the polarization of the light and the path must be continuous (i.e., no kinks in

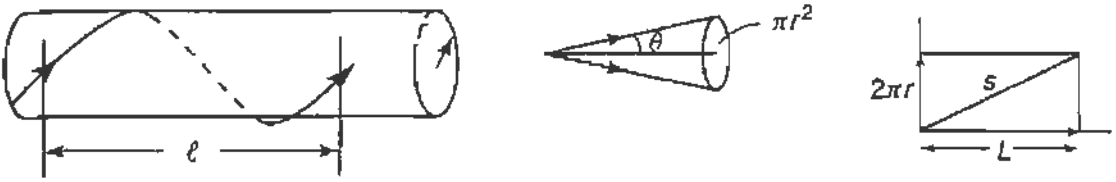


FIGURE 5.27 Layout of the fiber winding on a cylinder. Here the fiber length is s and the radius of the cylinder r.

the fiber). For instance we can wind the fiber on a cardboard tube as shown in Fig. 5.27a. If the radius of the tube is r and the length for one revolution (the pitch) is ℓ, the winding angle θ is given by

$$\cos\theta = \ell/s \qquad s = \sqrt{\ell^2 + (2\pi r)^2}. \tag{5.59}$$

The solid angle described by the fiber is then

$$\Delta\Omega = 2\pi(1 - \cos\theta) = 2\pi(1 - \ell/s). \tag{5.60}$$

The experimental setup is relatively simple. A HeNe laser beam is polarized and injected through a fiber coupler into the (single-mode) fiber. At the

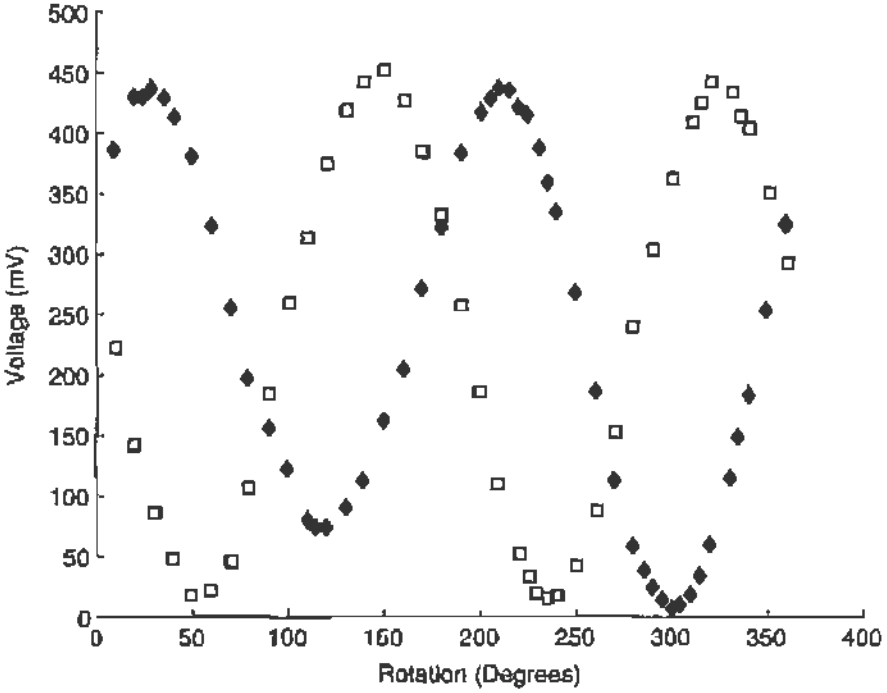


FIGURE 5.28 Results from a measurement of Berry's phase. The transmitted intensity is shown as a function of the angle of the analyzing polarizer. Open squares are for the flat topology, filled squares for helical winding. The polarization has rotated by 245° between the two measurements.

end of the fiber the light exits through another fiber coupler and is analyzed by a rotatable polarizer and a photodiode. We use two configurations, one in which the fiber is wound along the cylinder and the other when the fiber is laid out flat on the table. The detected intensity as a function of the angle of the analyzing polarizer is shown in Fig. 5.28. The open squares were obtained with the flat fiber, the solid squares with the helical winding. We see that the polarization has rotated by $\theta = 245°$ (or it could be 115° in the opposite direction!).

In this case the radius of the cylinder was $r = 14$ cm and the pitch $\ell = 28$ cm, for one complete turn. Thus $s = 92$ cm and

$$\Delta\Omega = 2\pi(1 - \ell/s) = 4.37 \text{ sr}.$$

Thus we expect a rotation angle $\phi = \Delta\Omega = 251°$ in excellent agreement with observation. One should repeat the measurement by making more than one turn on the cylinder (using a longer fiber) to fully confirm Eq. (5.53). More details on the first demonstration of Berry's phase with an optical fiber are given by Tomita and Chiao (1986).

5.9. REFERENCES

M. V. Berry, *Proc. R. Soc. London Ser. A* **392**, 45 (1984).
M. V. Berry, *Phys. Today* 36 (Dec. 1990).
A. Tomita and R. Y. Chiao, *Phys. Rev. Lett.* **57**, 927 (1986).

CHAPTER 6

High-Resolution Spectroscopy

6.1. INTRODUCTION

In 1896, P. Zeeman observed that when a sodium source was placed in a strong magnetic field, the yellow D lines were split into several components. Faraday had performed the same experiment some thirty years earlier but had failed to observe an effect because of the low resolution of his spectrograph. We also know from Chapter 1 that even in the absence of a magnetic field the atomic spectral lines have a fine structure that was easily observed with the small grating spectrometer; with a high-resolution instrument, however, it becomes possible to observe that each of these fine structure lines may again be resolved into closely spaced components, which form the so-called *hyperfine structure* (hfs) of atomic lines.[1]

[1] To set the reader at ease, no further splitting beyond the hyperfine structure has been observed, nor can it be expected for free atoms; in the hyperfine structure we include both the splitting due to *nuclear spin* and that due to the *isotope shift*.

The splitting of a spectral line is a consequence of a splitting of the energy of the initial state, of the final state, or of both states between which the transition takes place. The energy-level splittings produced by the application of an external magnetic field **B** (Zeeman effect) are on the order of

$$\Delta E = \boldsymbol{\mu} \cdot \mathbf{B} = \frac{e}{2m_e} \mathbf{L} \cdot \mathbf{B} \sim \frac{e\hbar}{2m_e} B, \tag{6.1}$$

where μ is the magnetic moment of the state (see Section 2 of this chapter). The constant $\mu_B = e\hbar/2m = 5.79 \times 10^{-11}$ MeV/T is called the Bohr magneton, so that in units of wave numbers the displacement for *one* Bohr magneton is

$$\Delta\bar{\nu} = \frac{\Delta\nu}{c} = \frac{\Delta E}{hc} = \frac{e}{4\pi m_e c} B = 46.69\ B \quad \mathrm{m}^{-1} \tag{6.2}$$

or

$$\Delta\nu = 14.01 B \text{ GHz}$$

with B in Tesla.

The hyperfine structure splitting is due to the interaction of the magnetic-dipole, electric-quadrupole, etc., moment of the nucleus, with the electromagnetic field produced by the electrons at the nucleus. The interaction energy for the magnetic-dipole terms is of the order of

$$\Delta E = \mu_N \cdot \langle \mathbf{B}_J(0) \rangle \ \sim \mu_N \mu_B \left\langle \frac{1}{r^3} \right\rangle = \frac{\mu_B^2}{1837} \left\langle \frac{1}{r^3} \right\rangle, \tag{6.3}$$

where μ_N is the nuclear magneton

$$\mu_N = \frac{e\hbar}{2m_N} = \frac{\mu_B}{1837},$$

and $\langle \mathbf{B}_J(0) \rangle$ is the expectation value for the magnetic field of the electrons at the origin; it is equal to $\mu_B \langle 1/r^3 \rangle$ (except for configurations with $\ell = 0$). Instead of evaluating $\langle 1/r^3 \rangle$ we recall that the fine structure splitting is due to an $\mathbf{L} \cdot \mathbf{S}$ coupling of the electrons, and therefore is of the order of $\mu_B^2 \langle 1/r^3 \rangle$ so that we expect

$$\Delta E(\text{hfs}) \sim \frac{\Delta E(\text{fs})}{1837}. \tag{6.4}$$

Let us substitute reasonable numbers in Eqs. (6.2) and (6.4); for example, if $B \simeq 1\mathrm{T}$

$$\Delta\overline{\nu}\ (\text{Zeeman}) \sim 46.0\ \mathrm{m}^{-1},$$

and since[2] $\Delta\overline{\nu}$ (fine structure) $\sim 10^4\ \mathrm{m}^{-1}$, we find that

$$\Delta\overline{\nu}(\text{hfs}) \sim 5.0\ \mathrm{m}^{-1} = 1.5\ \text{GHz}.$$

Thus the splitting of the lines is very small and can be observed only with a high-resolution instrument. Assuming $\lambda \approx 500$ nm and $\Delta\overline{\nu} \approx 5.0\ \mathrm{m}^{-1}$ we find that the required resolving power is

$$\frac{\lambda}{\Delta\lambda} = \frac{\overline{\nu}}{\Delta\overline{\nu}} = 4 \times 10^5.$$

Such a resolution may be achieved in two ways:

(a) With a large grating used in a high order; the resolving power of a grating is given by

$$\frac{\lambda}{\Delta\lambda} = Nn,$$

where n, the diffraction order, can be as large as 20, and for a 10-in. grating with 7000 rulings to the inch, the number of rulings is $N = 7 \times 10^4$, so that

$$\frac{\lambda}{\Delta\lambda} \sim 10^6.$$

Such gratings, are, however, very difficult to construct, but can now be obtained commercially.

(b) With a "multiple-beam" interferometer, the most common one today and easiest to use being the Fabry–Perot, which was discussed in Section 4.6. One can directly observe the "rings" of the interference pattern for a diverging beam. An optical filter or a dispersive element is needed to select the line of interest. Alternately one can use the Fabry–Perot in the "scanning mode" by moving one of the end-mirrors, through half a wavelength, and observing the transmission of a collimated beam. For instance a Fabry–Perot with 5 cm spacing has an FSR (free spectral range) of 3 GHz;

[2] See Section 1.6.3 and recall that $\overline{\nu} = \nu/c = 1/\lambda$.

even with modest finesse $F = 100$, the resolution (see Eq. (4.62)) is $\Delta\nu = 30$ MHz. Thus for $\lambda = 500$ nm, namely $\nu = 6 \times 10^{14}$ Hz

$$\frac{\nu}{\Delta\nu} = 2 \times 10^{7}.$$

In the following two sections the Zeeman effect and the theory of hyperfine structure are discussed in some detail. We also discuss the isotope shift and present data on the shift between the spectral lines of hydrogen and deuterium. We then describe a measurement of the Zeeman splitting of the 546.1-nm green line of Hg, using a Fabry–Perot etalon. The final section is devoted to a measurement of the hyperfine structure of rubidium using Doppler-free saturation spectroscopy.

The bibliography on atomic spectroscopy is vast and because of the "reach" of laser experiments it is kept up-to-date. A list of suggested references is given at the end of the chapter.

6.2. THE ZEEMAN EFFECT

6.2.1. The Normal Zeeman Effect

As already discussed in Section 1.4, the solution of the Schrödinger equation[3] yields "stationary states" labeled by three integer indices, n, l, and m, where $l < n$ and $m = -l, -l+1, \ldots, l-1, l$. For the screened Coulomb potential, the energy of these states depends on n and l but not on m; we therefore say that the $(2l+1)$ states with the same n and l index are "degenerate" in the m quantum number. Classically we can attribute this degeneracy to the fact that the plane of the "orbit" of the electron may be oriented in any direction without affecting the energy of the state, since the potential is spherically symmetric.

If a magnetic field B is switched on in the region of the atom, we should expect that the electrons (and the nucleus[4]) will interact with it. We need only consider the electrons *outside* closed shells, and assume there is one such electron; indeed the interaction of the magnetic field with this electron

[3] "Quantum Mechanics" A. Das and A. Melissinos, Gordon and Breach (1986), New York. Or any other text on quantum mechanics.

[4] For our present discussion this interaction of the nucleus with the external field is so small that we will neglect it.

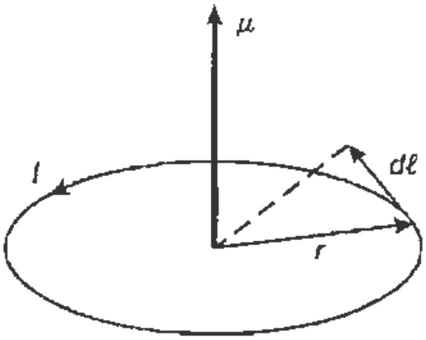


FIGURE 6.1 Magnetic moment due to a current circulating in a closed loop.

yields for each state an additional energy ΔE, given by

$$\Delta E = m\mu_B B. \tag{6.5}$$

Thus, the total energy of a state depends now on n, l, and m, and the degeneracy has been removed.

To see how this additional energy arises we consider the classical analogy. See Fig. 6.1. The orbiting electron is equivalent to a current density[5]

$$\mathbf{J}(\mathbf{x}) = -e\mathbf{v}\delta(\mathbf{x}-\mathbf{r}),$$

where $\mathbf{r}$ is the equation of the orbit and $\mathbf{x}$ gives the position of the electron; the negative sign arises from the negative charge of the electron. Such a current density gives rise to a magnetic-dipole moment

$$\boldsymbol{\mu} = \frac{1}{2}\int \mathbf{x} \times \mathbf{J}(\mathbf{x})\, d^3x = -\frac{1}{2}\, e(\mathbf{r} \times \mathbf{v}).$$

[5] For a circular orbit, the electron is equivalent to a current $I = \Delta Q/\Delta T = e/T = e\omega/2\pi$, where ω is the angular frequency $\omega = v/a$; a is the radius of the orbit. However, a plane closed loop of current gives rise to a magnetic moment $\mu = IA$, where A is the area enclosed by the loop; in our case $A = \pi a^2$, hence

$$\mu = \frac{ev}{2\pi a}\,\pi a^2 = \frac{eva}{2}.$$

The angular momentum for the circular orbit is $L = m_e va$, hence

$$\mu = \frac{e}{2m_e}L$$

as in Eq. (6.1).

However, the angular momentum of the orbit is given by

$$\mathbf{L} = \mathbf{r} \times \mathbf{p} = m_e(\mathbf{r} \times \mathbf{v}),$$

so that

$$\mu = -\frac{e}{2m_e}\,\mathbf{L} = -\frac{e\hbar}{2m_e}\,l\mathbf{u}_L, \tag{6.6}$$

where we expressed the angular momentum of the electron in terms of its quantized value $\mathbf{L} = l(h/2\pi)\mathbf{u}_L$ and $\mathbf{u}_L$ is a unit vector along the direction of **L**. The energy of a magnetic dipole in a homogeneous field is

$$E = -\mu \cdot \mathbf{B} = \frac{e}{2m_e}\,\mathbf{L} \cdot \mathbf{B}, \tag{6.7}$$

but the angle between **L** and the external field **B** cannot take all possible values.[6] We know that it is quantized, so that the projection of **L** on the z axis (which we can take to coincide with the direction of **B** since no other preferred direction exists) can only take the values $m = -l, -l+1, \ldots, l-1, l$. Thus the energy of a particular state n, l, m in the presence of a magnetic field will be given by[7]

$$E_{n,l,m} = -E_{n,l} + mB\mu_B, \tag{6.8}$$

where[8]

$$\mu_B = \frac{e\hbar}{2m_e}.$$

In Fig. 6.2 is shown the energy-level diagram for the five states with given n and $l = 2$, before and after the application of a magnetic field **B**. We note that all the levels are equidistantly spaced, the energy difference between them being

$$\Delta E = \mu_B B.$$

Let us next consider the transition between a state with n_i, l_i, m_i and one with n_f, l_f, m_f. As an example we choose $l_i = 2$ and $l_f = 1$, so that the

[6] This was first clearly shown in the Stern–Gerlach experiment. W. Gerlach and O. Stern, Z. Physik 9, 349 (192).

[7] The energy in the field is positive because the electron charge is taken as negative.

[8] m_e in this expression is the mass of the electron, not to be confused with the magnetic quantum number m.

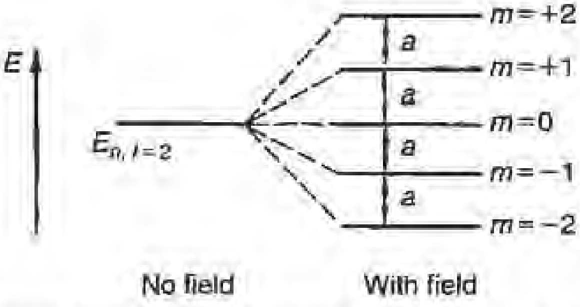


FIGURE 6.2 Splitting of an energy level under the influence of an external magnetic field. The level is assumed to have $l = 2$ and therefore is split into five equidistant sublevels.

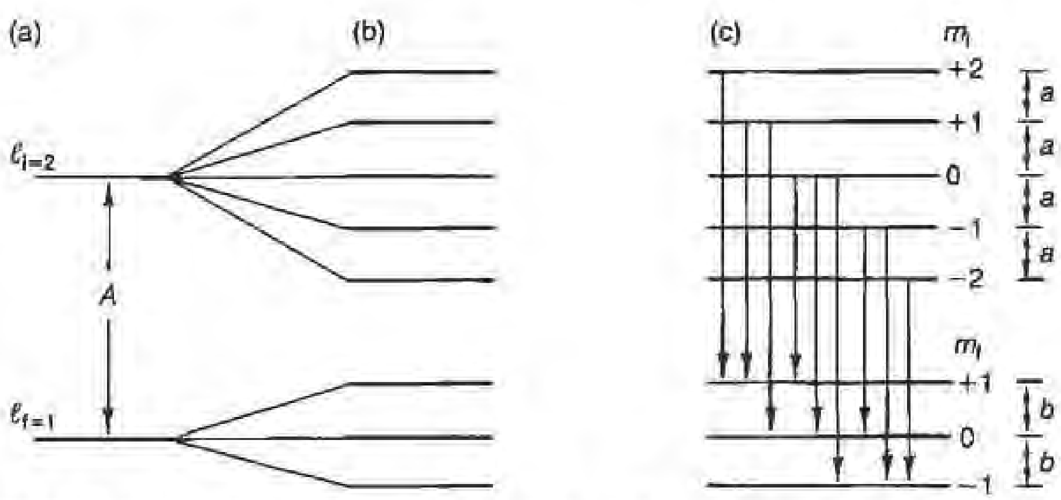


FIGURE 6.3 Splitting of a spectral line under the influence of an external magnetic field. (a) The initial level ($l = 2$) and the final level ($l = 1$) with no magnetic field are shown. A transition between these levels gives rise to the spectral lines. (b) The two levels after the magnetic field has been applied. (c) The nine allowed transitions between the eight sublevels of the initial and final states.

energy-level diagram is as shown in Fig. 6.3: without a magnetic field in Fig. 6.3a, and when the magnetic field is present in Fig. 6.3b.

However, for an *electric-dipole* transition to take place between two levels, certain selection rules must be fulfilled; in particular,

$$\Delta l = \pm 1. \tag{6.9}$$

Thus, when the field is turned on, we cannot expect transitions between the m sublevels with the same l, since they do not satisfy Eq. (6.9). Further, the transitions between the sublevels with $l_i = 2$ to the sublevels with

$l_f = 1$ that do satisfy Eq. (6.9) are now governed by the *additional* selection rule[9]

$$\Delta m = 0, \ \pm 1, \tag{6.10}$$

and thus only the transitions shown in Fig. 6.3c are allowed.

Let the energy splitting in the initial level be a, and in the final level be b, and let A be the energy difference between the two levels when no magnetic field is applied. Then the energy released in a transition $i \rightarrow f$ is given by

$$E_i - E_f = A_{if} + m_i a - m_f b. \tag{6.11}$$

These energy differences for the nine possible transitions shown in Fig. 6.3c are given in matrix form in Table 6.1; × indicates that the transition is forbidden and will not take place.

At this point the reader must be concerned about the use of a and b; according to our previous argument (Eq. (6.8)), as long as all levels are subject to the same magnetic field B, their splitting must also be the same, and

$$a = b = \mu_B B.$$

Thus, we see from Eq. (6.11) (or Table 6.1) that only *three* energy differences are possible

$$E_i - E_f = A + a(m_f - m_i) = A + a\Delta m,$$

where Δm is limited by the selection rule, Eq. (6.10), to the *three* values $+1, 0, -1$. Consequently, in the presence of a magnetic field B, the single

TABLE 6.1 Allowed Transitions from $l_i = 2$ to $l_f = 1$ and the Corresponding Energies

m of	m of initial state				
final state	$+2$	$+1$	0	-1	-2
$+1$	$A + 2a - b$	$A + a - b$	$A - b$	×	×
0	×	$A + a$	A	$A - a$	×
-1	×	×	$A + b$	$A - a + b$	$A - 2a + b$

[9] The selection rules of atomic spectroscopy are a consequence of the addition of angular momenta. In this specific case the selection rules indicate that we consider only *electric-dipole* radiation.

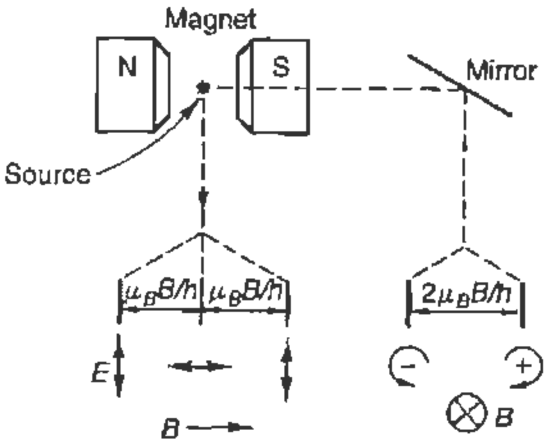


FIGURE 6.4 The polarization and separation of the components of a normal Zeeman multiplet when viewed in a direction normal to, and in a direction parallel to, the magnetic field.

spectral line of frequency $\nu = A/h$ is split into three components with frequencies

$$\nu_- = (A - \mu_B B)/h, \quad \nu_0 = A/h, \quad \text{and} \quad \nu_+ = (A + \mu_B B)/h$$

irrespective of the values of l_i and l_f. Furthermore, these spectral lines are polarized, as shown in Fig. 6.4. When the Zeeman effect is viewed in a direction *normal* to the axis of the magnetic field, the central component is polarized parallel to the axis, whereas the two outer ones are polarized normal to the axis of the field. When the Zeeman effect is observed *along* the axis of the field (by making a hole in the pole face, or using a mirror), only the two outer components appear, circularly polarized. The lines from $\Delta m = +1$ transitions appear with right-hand circular polarization, and from $\Delta m = -1$ transitions with left-hand circular polarization. The central line does not appear, since the electromagnetic field must always have the field vectors (**E** and **B**) normal to the direction of propagation.

The splitting of a spectral line into a triplet under the influence of a magnetic field is called the "*normal*" *Zeeman effect*, and is occasionally observed experimentally, as, for example, in the 579.0-nm line of mercury arising in a transition[10] from ${}^1D_2$ to ${}^1P_1$. However, in most cases the lines are split into more components, and even where a triplet appears it does not always show the spacing predicted by Eq. (6.8). This is due to the

[10] Note that both the initial and final states have $S = 0$.

intrinsic magnetic moment of the electron (associated with its spin) and will be discussed in the following sections.

6.2.2. The Influence of the Magnetic Moment of the Electron

In Section 1.6 it was discussed how the intrinsic angular momentum (spin) of the electrons **S** couples with the orbital angular momentum of the electrons **L** to give a resultant **J**; this coupling gave rise to the "fine structure" of the spectra.[11] The projections of **J** on the z axis are given by m_J, and we could expect (on the basis of our previous discussion) that the *total* magnetic moment of the electron will be given by

$$\mu = \frac{\mu_B}{\hbar}\,\mathbf{J}. \tag{6.12}$$

Consequently, the energy-level splitting in a magnetic field B would be in analogy to Eq. (6.8):

$$\Delta E = -m_J \mu_B B. \tag{6.13}$$

These conclusions, however, are not correct because the *intrinsic* magnetic moment of the electron is related to the *intrinsic* angular momentum of the electron (the spin) through

$$\mu_s = 2\,\frac{e}{2m_e}\,\mathbf{S} = 2\,\frac{e\hbar}{2m_e}\,s\mathbf{u}_s \tag{6.14}$$

and *not* according[12] to Eq. (6.6). Consequently, the *total* magnetic moment of the electron is given by the operator

$$\mu = (\mu_B/\hbar)[\mathbf{L} + 2\mathbf{S}]. \tag{6.15}$$

[11] We will use the following notation: **L**, **S**, **J** represent angular momentum vectors that have magnitude $\hbar\sqrt{l(l+1)}$, $\hbar\sqrt{s(s+1)}$, $\hbar\sqrt{j(j+1)}$. The symbols l, j, etc. (s is always $s = \frac{1}{2}$), are the *quantum numbers* that label a one-electron state and appear in the above square root expressions. The symbols L, S, J, etc., are *quantum numbers* that label a state with more than one electron and are then used instead of l, s, j.

[12] The result of Eq. (6.14) is obtained in a natural way from the solution of the Dirac equation; it also emerges from the classical relativistic calculation of the "Thomas precession."

We can think of μ as a vector oriented along **J** but of magnitude

$$\mu = \mu_B g\, J. \tag{6.16}$$

The numerical factor g is called the Landé g factor and a correct quantum-mechanical calculation gives[13]

$$g = 1 + \frac{j(j+1) + s(s+1) - l(l+1)}{2j(j+1)}. \tag{6.17}$$

The interesting consequence of Eqs. (6.16) and (6.17) is that now the splitting of a level due to an external field B is

$$E_{n,j,l,m_j} = -E_{n,j,l} + g\mu_B B m_j \tag{6.18}$$

and in contrast to Eq. (6.8) is *not* the same for all levels; it depends on the values of j and l of the level ($s = \frac{1}{2}$ always when one electron is considered). The sublevels are still equidistantly spaced but by an amount

$$\Delta E = g\mu_B B.$$

Consider then again the transitions between sublevels belonging to two states with different l (in order to satisfy Eq. (6.9)). However, since we are taking into account the electron spin, l is not a good quantum number, and instead the j values of the initial and final levels must be specified. If we

[13] This result can also be obtained from the vector model for the atomic electron. In Fig. 6.5 the three vectors **J**, **L**, and **S** are shown, and **L** and **S** couple into the resultant **J**, so that

$$\mathbf{J} = \mathbf{L} + \mathbf{S}.$$

By taking the squares of the vectors, we obtain the following values for the cosines

$$\cos(\mathbf{L}, \mathbf{J}) = \frac{j^2 + l^2 - s^2}{2lj} \qquad \cos(\mathbf{S}, \mathbf{J}) = \frac{j^2 + s^2 - l^2}{2sj}.$$

From Eq. (6.15) we see that

$$\mu/\mu_B = l\cos(\mathbf{L}, \mathbf{J}) + 2s\cos(\mathbf{S}, \mathbf{J}).$$

Thus

$$g = \frac{\mu}{\mu_B j} = \frac{j^2 + l^2 - s^2}{2j^2} + \frac{2j^2 + 2s^2 - 2l^2}{2j^2} = 1 + \frac{j^2 + s^2 - l^2}{2j^2}.$$

Finally we must replace j^2, s^2, and l^2 by their quantum-mechanical expectation values $j(j+1)$, etc., and we obtain Eq. (6.17).

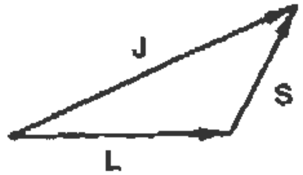


FIGURE 6.5 Addition of the orbital angular momentum **L** and of the spin angular momentum **S** into the total angular momentum **J**, according to the "vector model."

choose for this example $l_i = 1$ and $l_f = 0$, we have the choice of $j_i = \frac{3}{2}$ or $j_i = \frac{1}{2}$, whereas $j_f = \frac{1}{2}$. Transitions may occur only if they satisfy, in addition to Eq. (6.9), *also* the selection rules for j

$$\Delta j = 0\,,\ \pm 1 \qquad \text{not} \qquad j = 0 \rightarrow j = 0. \tag{6.9a}$$

Furthermore the selection rules for m_j must also be satisfied; they are the same as given by Eq. (6.10)

$$\Delta m_j = 0\,,\ \pm 1. \tag{6.10a}$$

In Fig. 6.6 the energy-level diagram is given without and with a magnetic field for the doublet initial state with $l = 1$, and the singlet final state, $l = 0$. Six possible transitions between the initial states with $j = \frac{3}{2}$ to the final state with $j = \frac{1}{2}$ are shown (as well as the four possible transitions from $j = \frac{1}{2}$ to $j = \frac{1}{2}$). By using Eq. (6.17) we obtain the following g factors

$$\begin{array}{llll} l = 1 & j = \frac{3}{2} & s = \frac{1}{2} & g = \frac{4}{3} \\ l = 1 & j = \frac{1}{2} & s = \frac{1}{2} & g = \frac{2}{3} \\ l = 0 & j = \frac{1}{2} & s = \frac{1}{2} & g = 2. \end{array}$$

The sublevels in Fig. 6.6 have been spaced accordingly.

In Table 6.2 are listed the six transitions from $j = \frac{3}{2}$ to $j = \frac{1}{2}$ in analogy with Table 6.1. However, *since now* $a \neq b$, the spectral line is split into a six-component (symmetric) pattern. This structure of the spectral line is indicated in the lower part of Fig. 6.6; following adopted convention, the components with polarization parallel to the field are indicated above the base line, and with polarization normal to the field, below.[14] As before the parallel components have $\Delta m = 0$, the normal ones $\Delta m \pm 1$.

[14] It is also conventional to label the parallel components with π, and the normal ones by σ (from the German "Senkrecht").

TABLE 6.2 Allowed Transitions from $j_i = \frac{3}{2}$ to $j_f = \frac{1}{2}$ and the Corresponding Energies

m_j of final state	m_j of initial state			
	$+\frac{3}{2}$	$+\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{3}{2}$
$+\frac{1}{2}$	$A + \frac{3a}{2} - \frac{b}{2}$	$A + \frac{a}{2} - \frac{b}{2}$	$A - \frac{a}{2} - \frac{b}{2}$	×
$-\frac{1}{2}$	×	$A + \frac{a}{2} + \frac{b}{2}$	$A - \frac{a}{2} + \frac{b}{2}$	$A - \frac{3a}{2} + \frac{b}{2}$

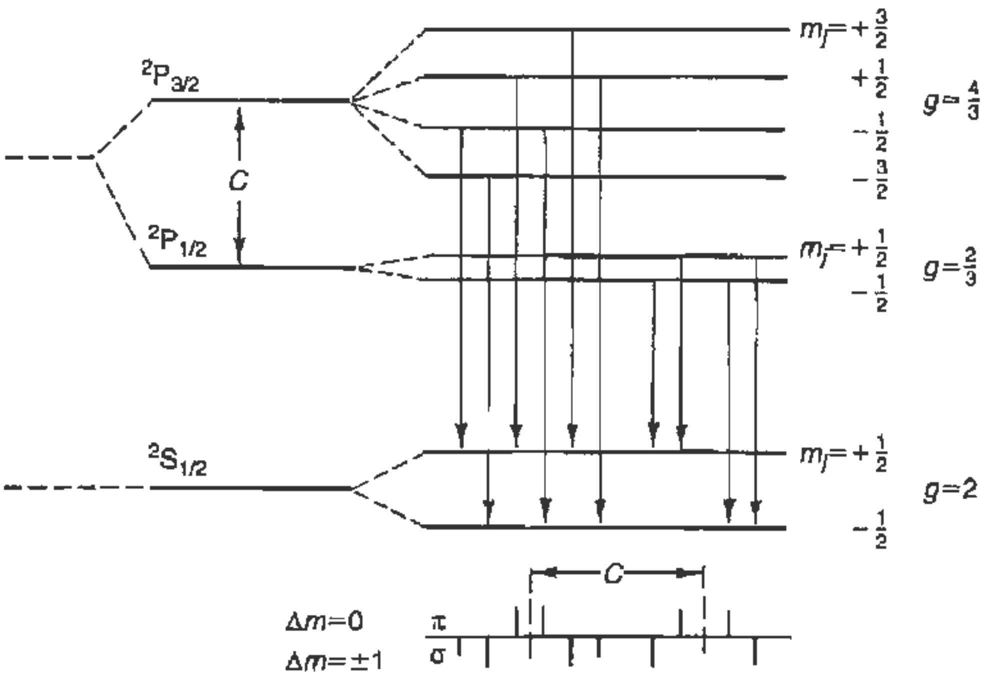


FIGURE 6.6 Energy levels of a single valence electron atom showing a P state and an S state. Due to the fine structure, the P state is split into a doublet with $j = \frac{3}{2}$ and $j = \frac{1}{2}$. Further, under the influence of an external magnetic field each of the three levels is split into sublevels as shown in the figure where account has been taken of the magnetic moment of the electron. The magnetic quantum number m_j for each sublevel is also shown as is the g factor for each level. The arrows indicate the allowed transitions between the initial and final states, and the structure of the line is shown in the lower part of the figure.

The horizontal spacing between the components is proportional to the differences in the energy of the transition, and the vertical height is proportional to the intensity of the components; the relative intensity can be predicted exactly since it involves only the comparison of matrix elements between the angular parts of the wave function.

As the magnetic field is raised, the separation of the components continues to increase linearly with the field until the separation between Zeeman components becomes on the order of the fine-structure separation (spacing C in Fig. 6.6). At this point the Zeeman components from the $j = \frac{3}{2} \to \frac{1}{2}$ and $j = \frac{1}{2} \to \frac{1}{2}$ transition begin to overlap; clearly the perturbation caused by the external magnetic field is on the order of the $\mathbf{L} \cdot \mathbf{S}$ energy and affects the coupling of $\mathbf{L}$ and $\mathbf{S}$ into $\mathbf{J}$; $\mathbf{J}$ ceases to be a "good quantum number."

For very strong fields, $\mathbf{L}$ and $\mathbf{S}$ become completely uncoupled, so that the orbital and intrinsic magnetic moments of the electron interact with the field independently, giving rise to an energy shift

$$\begin{aligned}\Delta E &= -\frac{\mu_B}{\hbar}\,\mathbf{L} \cdot B - 2\,\frac{\mu_B}{\hbar}\,\mathbf{S} \cdot \mathbf{B} - a\mathbf{L} \cdot \mathbf{S} \\ &= -\mu_B B(m_l + 2m_s) - am_l m_s. \qquad (6.19)\end{aligned}$$

In this region one speaks of the Pashen–Back effect. The reader can find more details in the references, in particular in the classic text by Condon and Shortley.

So far we have discussed the case where the atom has only a single valence electron. In Section 1.6 we considered also atoms with two valence electrons and saw that for Hg the total angular momentum $\mathbf{J} = \mathbf{L}+\mathbf{S}$, where $\mathbf{L}$ results from the coupling of l_1 and l_2 and $\mathbf{S}$ from the coupling of s_1 and s_2. In this case the g factor is still given by Eq. (6.17), but by using L, S, and J, the quantum numbers for the coupled angular momenta.

An interesting case arises in the 579.07-nm yellow line of Hg, which is due to the transition from the 6 $^1D_2$ state to the 6 $^1P_1$ state. (See Fig. 1.24 for the energy level diagram of Hg.) As the reader should verify, by using Eq. (6.17), the g factors of the initial and final state are both equal to 1. Thus we have exactly the situation shown in Fig. 6.3, and the line splits into three components (normal Zeeman effect).

6.3. HYPERFINE STRUCTURE

Spectral lines, when examined under high resolution, do show structure even in the absence of an external magnetic field. As already mentioned this *hyperfine* structure arises from the interaction of the atomic electrons with the nucleus. The largest effect arises from the magnetic-dipole moment of the nucleus, but the effect of higher order moments are also observed. A related effect is the isotope shift, which shifts the spectral lines between isotopes, i.e., atoms of the same element but with nuclei of different mass.

6.3.1. The Effects of Nuclear Spin

Nuclei can have an intrinsic angular momentum (spin) different from 0. We use **I** to designate the nuclear spin which can take the values (i.e., the quantum number) 0, $\frac{1}{2}$, 1, $\frac{3}{2}$, ... that can reach very high values for excited nuclear states. When $I \geq \frac{1}{2}$ we can expect that the "spinning" charge of the nucleus will give rise to a magnetic moment (see Eq. (6.6)) oriented along the spin axis

$$\mu = -\frac{e}{2M}\mathbf{I},$$

where M is the mass of the nucleus. In addition, nuclei exhibit an intrinsic magnetization,[15] so that in general we have

$$\mu = -g_I \frac{e}{2m_p}\mathbf{I} = g_I \mu_N I \mathbf{u}_I,$$

where $\mathbf{u}_I$ is a unit vector along the spin direction, and

$$\mu_N = \frac{e\hbar}{2m_p}$$

is the *nuclear magneton*; m_p is the proton mass. The numerical factor g_I includes all the effects of intrinsic and orbital magnetization of the nucleus and can be obtained only from a theory of nuclear structure.

The magnetic moment of the nucleus, μ, will interact with the magnetic field $\mathbf{B}_e(0)$ produced by the atomic electrons (at the nucleus; Fig. 6.7). This interaction then results in a shift of the energy levels of the atom by the amount

$$\Delta E = -\mu \cdot \mathbf{B}_e(0). \tag{6.20}$$

The direction of $\mathbf{B}_e(0)$ is that given by the total angular momentum of the atomic electrons, namely,[16] **J**, so that

$$\Delta E = +\left(\frac{\mu}{|I|}\right)\left(\frac{B_e(0)}{|J|}\right)\mathbf{I}\cdot\mathbf{J}. \tag{6.21}$$

[15] This gives rise to the so-called "anomalous" magnetic moment of the nucleon; for example, the neutron (an uncharged particle) has a magnetic moment of $-1.91\ \mu_N$.

[16] The direction of $\mathbf{B}_e(0)$ is really opposite to **J** because the electron has negative charge.

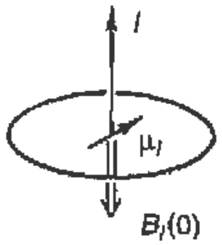


FIGURE 6.7 Interaction of the nuclear magnetic moment with the magnetic field produced by the electrons at the nucleus.

Thus, we expect the splitting of a level of given J according to the possible values of $(\mathbf{I} \cdot \mathbf{J})$, which, as we know, are quantized. The situation is analogous to that of the fine structure, where the interaction was proportional to the $(\mathbf{L} \cdot \mathbf{S})$ term. In that instance the two angular momenta coupled into a resultant $\mathbf{J} = (\mathbf{L} + \mathbf{S})$ according to the quantum-mechanical laws of addition of angular momentum. In the present situation, $\mathbf{J}$ and $\mathbf{I}$ couple into a total angular momentum of the atom designated by $\mathbf{F}$:

$$\mathbf{F} = (\mathbf{I} + \mathbf{J}). \tag{6.22}$$

An energy level of given $\mathbf{J}$ is then split into sublevels having all possible values of F, namely, the integers (or half-integers)

$$|J - I| \leq F \leq |J + I|.$$

Thus if $I = \frac{1}{2}$, the level is split into two components, with $F_1 = J + \frac{1}{2}$ and $F_2 = J - \frac{1}{2}$ (provided $J \geq \frac{1}{2}$); if $I = 1$, the level is split into three components with $F_1 = J - 1$, $F_2 = J$, and $F_3 = J + 1$ (provided $J \geq 1$); etc. This situation is shown in Fig. 6.8, and we see that if J is known, the number of hyperfine structure components of a spectral line provides direct information on the spin of the nucleus.

If either $I = 0$ or $J = 0$, no splitting of the energy levels can occur since the interaction energy specified by Eq. (6.21) vanishes. This is to be expected because if $I = 0$, the nucleus cannot have a dipole moment, and if $J = 0$, then by symmetry, the magnetic field at the origin $B_e(0) = 0$.

Using Eq. (6.22), we can now obtain the expectation value of the operator $(\mathbf{I} \cdot \mathbf{J})$ that appears in Eq. (6.21); referring to the vector model

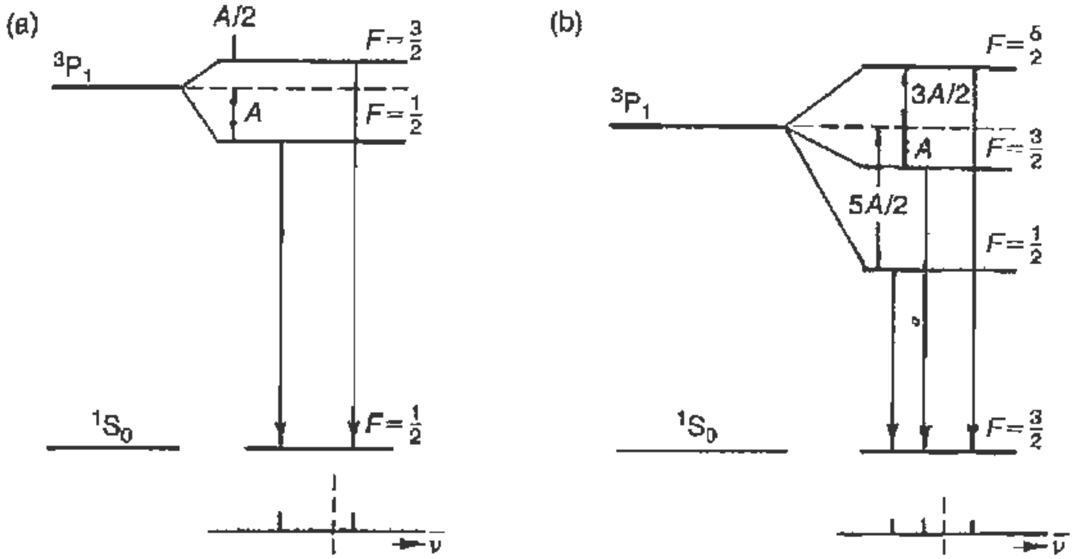


FIGURE 6.8 Hyperfine structure splitting of a $^3P_1$ atomic energy level, and the allowed transitions between the hyperfine structure components of this level and a $^1S_0$ final state when the spin of the nucleus is (a) $I = \frac{1}{2}$ and (b) $I = \frac{3}{2}$.

we write "classically"

$$\cos(\mathbf{I}, \mathbf{J}) = \frac{F^2 - I^2 - J^2}{2IJ}$$

and replacing F^2, etc., by the quantum-mechanical expectation values $F(F+1)$ we obtain

$$\Delta E = \frac{A}{2}\,[F(F+1) - I(I+1) - J(J+1)], \tag{6.23}$$

where the constant A is given by

$$A = \frac{\mu}{|I|}\,\frac{\langle B_e(0)\rangle}{|J|}. \tag{6.24}$$

Note that the energy splitting between sublevels, as given by Eq. (6.23) (and shown in Fig. 6.8), is not symmetric. Further, if we succeed in extracting from the experimental data the constant A, we can obtain the nuclear magnetic moment if $\langle B_e(0)\rangle$ is known.

The calculation of the average value of the magnetic field of the electrons at the nucleus $\langle B_e(0)\rangle$, however, is not easy to perform, and depends on the orbital angular momentum of the valence electron or electrons. Expressions for the "constant" A in terms of the atomic wave function can be found in the references (see Kopferman).

6.3.2. Isotope Shift

Figure 6.9 shows the hyperfine structure of the 253.7-nm line of *natural* mercury when examined under high resolution. When the lines are correctly identified we note that the different isotopes have different energies. Indeed natural mercury consists of several isotopes with the abundances shown in Table 6.3; the nuclear spin, nuclear-dipole magnetic moment, and electric-quadrupole moment are also indicated.

The isotope shift arises from two effects: (a) The finite mass of the nucleus: The nucleus is much heavier than the electron, but we can consider its mass as infinite only to a first approximation. (b) The finite size of the nucleus: The nuclear radius is much smaller than the orbit of the electron, but we can consider the nucleus as a point only to a first-order approximation. For light elements the isotope shift is mainly due to the effect of the finite mass, whereas for the heavy elements it is mainly due to the finite size effect. It should also be evident that we cannot measure the shift in the energy level of a single isotope, but only the difference in the shift between two or more isotopes. This is shown in Fig. 6.10a.

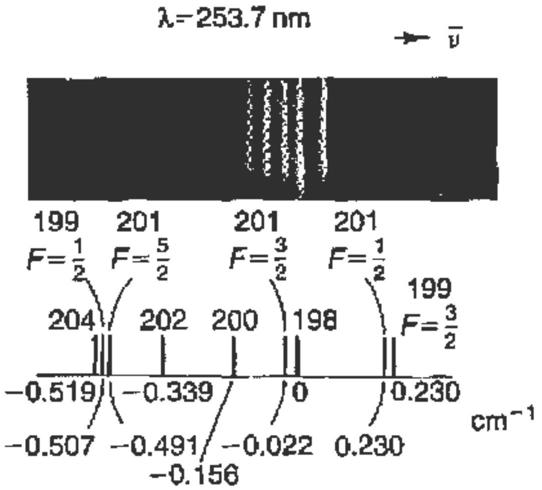


FIGURE 6.9 High-resolution spectrogram of the 253.7-nm line of natural mercury. In the lower part of the figure the various components are identified and their separation from the position of the ^{198}Hg component is also indicated. (Note that the ^{198}Hg component appears in the spectrogram as the longer line.)

TABLE 6.3 Properties of the Isotopes of Natural Hg ($Z = 80$)

Isotope	Abundance (percent)	N (neutrons)	I (nuclear spin)	μ (units of μ_N)	Q ($\text{cm}^2 \times 10^{-24}$)
198	10.1	118	0	0	
199	17.0	119	$\frac{1}{2}$	0.876	
200	23.2	120	0	0	
201	13.2	121	$\frac{3}{2}$	−0.723	0.38
202	29.6	122	0	0	
204	6.7	124	0	0	

In terms of the solutions of the Schrödinger equation we must consider both the electron and nucleus as revolving about the *center of mass* of the electron–nucleus system. This leads back to the Schrödinger equation for a stationary attractive center (nucleus) if the mass of the electron is replaced by its *reduced mass*

$$m' = m_e \frac{M}{M + m_e}, \tag{6.25}$$

where M is the mass of the nucleus. Then the energy of a hydrogen-like level is given by

$$E_n = -\frac{hcR_\infty Z^2}{n^2}\left(\frac{M}{M + m_e}\right) \simeq -\frac{hcR_\infty Z^2}{n^2}\left(1 - \frac{m_e}{M}\right) \tag{6.26}$$

where Z is the nuclear charge. For instance, the value of the Rydberg as obtained from the spectra of hydrogen and deuterium will differ by

$$\frac{R_\text{H}}{R_\text{D}} \simeq \left(1 - \frac{m_e}{2m_p}\right), \tag{6.27}$$

where we set the mass of the deuteron $m_d \sim 2m_p$. This will shift the spectral lines by 3×10^{-4}, which we can observe in the laboratory.

For the heavier elements the isotope shift due to finite mass becomes very small. Instead it is the finite size of the nucleus that is the dominant reason for a shift of the energy levels. Consider Fig. 6.10b where curve (a) represents the Coulomb potential of a point charge. If it is assumed that the electric charge of the nucleus is distributed on a spherical surface of radius r_0, then the potential will not diverge at $r = 0$, but will be constant for all $r \leq r_0$. Thus the potential seen by an electron will be of the form shown

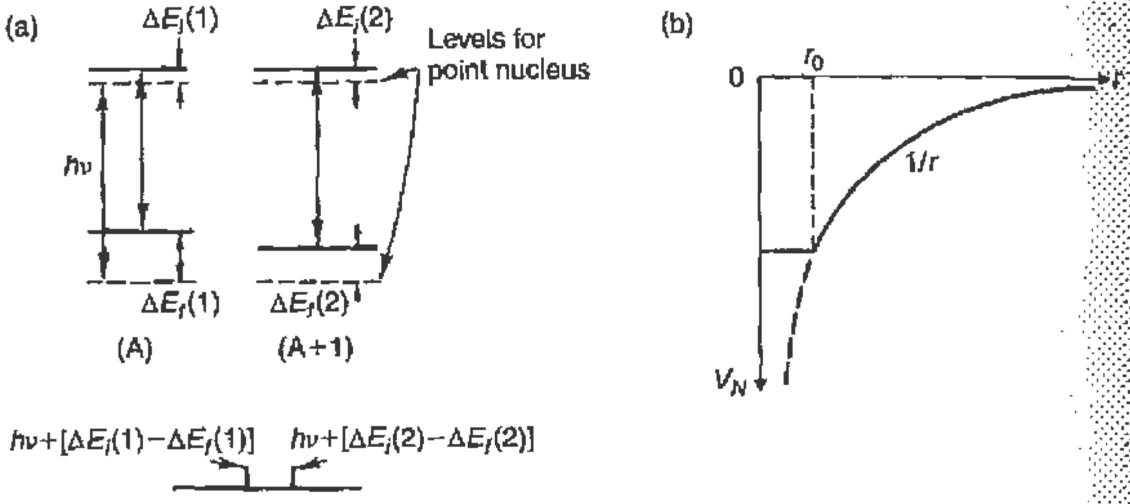


FIGURE 6.10 The isotope shift of atomic spectral lines. (a) The energy levels of the initial and final states of two different isotopes with mass numbers A and $A + 1$ are shown. The dashed lines show the position the levels would have if the nucleus was an infinitely heavy point; the solid lines show the actual position of the levels, which are shifted by a different amount for each isotope, and for each level. (b) Modification of the Coulomb potential of the nucleus due to its finite size.

by the solid curve of Fig. 6.10b. This leads to a significant energy shift as a function of r_0. Since the nuclear radius can be expressed as

$$r_0 = A^{1/3} \times 1.2 \times 10^{-13}\ \text{cm},$$

where A is the number of nucleons (protons and neutrons) in the nucleus, we see that $\Delta r_0/r_0 = \Delta A/3A$, which can be significant.

6.3.3. Measurement of the H–D Isotope Shift

The hydrogen–deuterium shift is quite large and can be measured with an instrument of modest resolution. The results presented here were obtained with a Jarrell–Ash grating spectrometer. A schematic of the spectrometer is shown in Fig. 6.11 and conforms with the generic spectrometer design introduced in Fig. 5.13. Instead of lenses, focusing mirrors are used to image the entrance slit onto the photomultiplier tube (PMT). The advantage of using a PMT is that very low levels of light can be detected so that the entrance and detector slits can be set to very narrow width. The grating had 630 rulings per millimeter, and the focal length of the lens was $f = 0.5$ m. The spectrum was viewed in second order with a resolution $\Delta\lambda/\lambda \sim 2 \times 10^{-5}$. The angle of the grating was computer controlled so that the speed at which

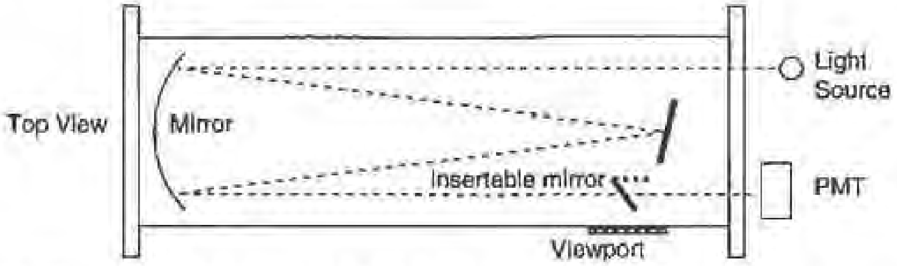


FIGURE 6.11 Schematic layout of the high-resolution Jarrell–Ash grating spectrometer.

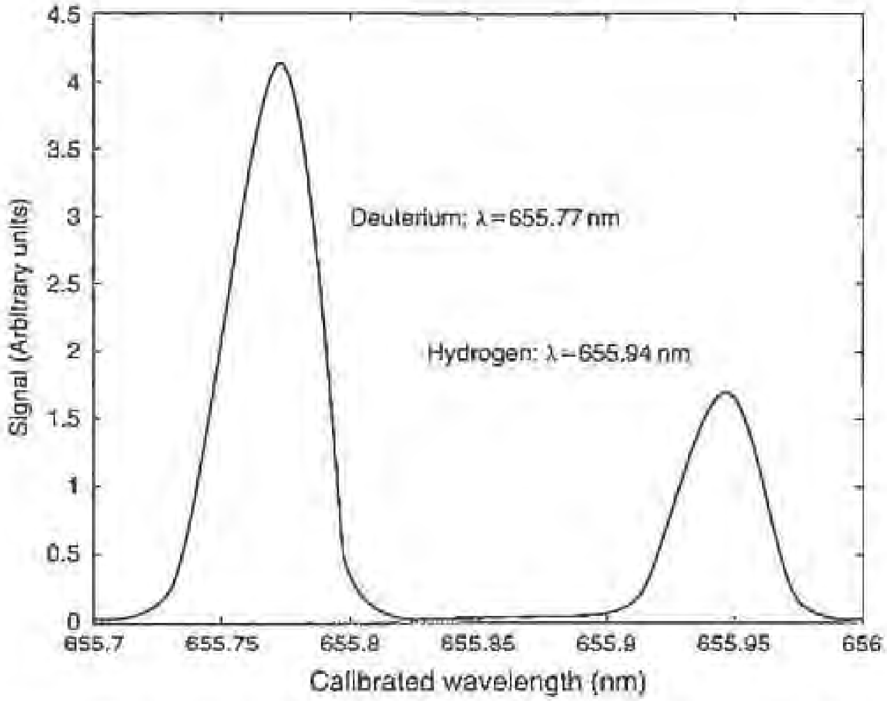


FIGURE 6.12 The red line of the Balmer series for a source containing hydrogen and deuterium observed in high resolution. The absolute wavelength calibration is not exact but this has insignificant effect on the wavelength difference between the two lines.

the spectrum was swept could be adjusted; slow speed for high resolution and vice versa. Furthermore, the grating angle was calibrated to indicate wavelength in nanometers.

For this experiment the source was a discharge tube containing deuterium and an admixture of hydrogen. The entrance slit was closed to a few hundred μm, and the first (red) line of the Balmer series ($n_i = 3$, $n_f = 2$), $\lambda = 656.28$ nm, was examined. The resulting spectrum is shown in Fig. 6.12 where the hydrogen line (longer wavelength) is well separated from the deuterium line. Note that the absolute calibration of the wavelength scale is off by almost 0.3 nm; this is not important in the present case where we are interested in the wavelength difference.

In terms of the calibration we find that

$$\lambda_{\mathrm{H}} = 655.94 \text{ nm}$$
$$\lambda_{\mathrm{D}} = 655.77 \text{ nm}.$$

Convert the wavelength difference into frequency difference

$$\nu_{\mathrm{H}} - \nu_{\mathrm{D}} = c\,\frac{\lambda_{\mathrm{D}} - \lambda_{\mathrm{H}}}{\lambda_{\mathrm{D}}\lambda_{\mathrm{H}}} = -11.85 \text{ GHz},$$

namely, a fractional frequency change

$$\frac{\Delta\nu_{\mathrm{H-D}}}{\nu_{\mathrm{D}}} = -\frac{\Delta\lambda}{\lambda} = -2.59 \times 10^{-4}.$$

From Eq. (6.27) we expect that

$$\frac{\Delta\nu_{\mathrm{H-D}}}{\nu_{\mathrm{D}}} = \frac{R_{\mathrm{H}} - R_{\mathrm{D}}}{R_{\mathrm{D}}} \simeq -\frac{m_e}{2m_p} = -2.72 \times 10^{-4}$$

in close agreement (within 5%) with the measured value.

6.4. THE LINE WIDTH

Since we are trying to resolve very small differences between the components of a spectral line, it is evident that the width of these components must be narrower than the separation between them. Before the advent of the laser, this was a very difficult task, but today laser lines can be stabilized to a remarkably narrow width, and used for spectroscopic studies.

Spectral lines have a natural width given by

$$\Delta\nu = \frac{\Delta E}{h} \simeq \frac{1}{2\pi\,\Delta\tau}, \tag{6.28}$$

where $\Delta\tau$ is the lifetime of the state; this is usually negligible, since atomic lifetimes are on the order of $\tau \gtrsim 10^{-8}$ s. Thus

$$\Delta\nu \lesssim \frac{1}{2\pi \times 10^{-8}} \sim 15 \text{ MHz}.$$

In wave numbers we find $\Delta\bar{\nu} < 0.05 \text{ m}^{-1}$. However, external influences do broaden spectral lines considerably; the main causes are as follows:

(a) *Doppler Broadening.* Due to their thermal energy, the atoms in the source move in random directions with a velocity given by the

Maxwell-Boltzmann distribution. Consequently, the wavelength emitted in a transition of the atom is Doppler-shifted; this results in a broadening of the line, which can be shown to have a half-width

$$\frac{\Delta\nu}{\nu} = 10^{-6}\sqrt{\frac{T}{A}}, \tag{6.29}$$

where T is the absolute temperature in Kelvins, and A is the atomic number of the element. Doppler broadening is most serious for the light elements and in sources that operate at high temperatures. For example, in an arc discharge operating at $T = 3600\,\text{K}$, a hydrogen line of $\lambda = 500$ nm will have a Doppler width of 36 GHz, which will mask any hyperfine structure. For heavy elements, as in Hg ($A \sim 200$), $\Delta\nu = 3$ GHz, which is still quite broad.

(b) *Pressure (or Collision) Broadening.* When the pressure in the source vapor is too high, the atoms are subject to frequent collisions, which in a way can be thought of as reducing the time interval $\Delta\tau$ entering into Eq. (6.28).

(c) *External Fields.* Magnetic or electric fields produce Zeeman or Stark splitting of the components, resulting in effective broadening of the line. Electric fields of 1000 V/cm can cause a broadening of tens of gigahertz.

(d) *Self-Absorption and Reversal.* This phenomenon is most pronounced with resonance lines. As the radiation emitted from the atoms in the middle of the source travels through the vapor, it has a probability of being absorbed that is proportional to the path length it traverses and to the absorption cross section; this will be strongest in the center of the line and weaker in the wings. The result shown in Fig. 6.13a is that the line becomes "squashed" in the center; that is, it is broadened.

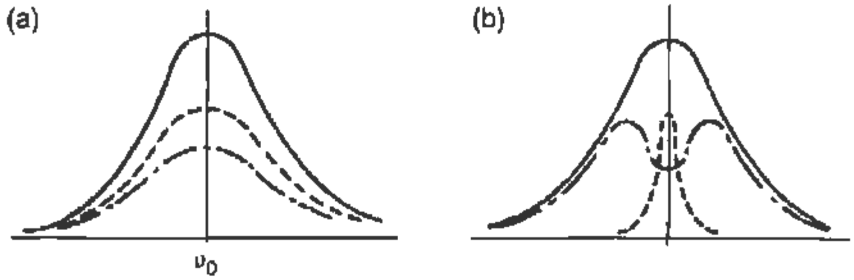


FIGURE 6.13 Broadening of a spectral line due to self-absorption in the source. The solid curve is the emitted line, the dashed curve represents the part of the radiation that is absorbed, and the dash–dot curve shows the transmitted line, which is the difference of the two former curves. (a) Normal absorption, and (b) strong absorption especially in the central region leading to self-reversal.

If the outer layers of the source are much cooler than the middle ones, the width of the particular energy level (due to the Doppler effect) is smaller in the outer layers and absorption takes place only at the central frequency with almost none in the wings. The result is a "self-reversed" line as shown in Fig. 6.13b. This effect is very pronounced in the sodium D lines, and when it is viewed with a high-resolution instrument, the line exhibits a doublet structure that is frequently mistaken for hyperfine structure. One can verify the origin of the effect because it varies with the voltage used to excite the source.

6.5. THE ZEEMAN EFFECT OF THE GREEN LINE OF ^{198}Hg

6.5.1. Equipment and Alignment

We now discuss the observation of the Zeeman effect on the $\lambda = 546.1$-nm line of ^{198}Hg. The choice of the green line is due to its predominance in the mercury spectrum, and the ease with which it can be observed. In an external magnetic field, it is split into nine components, as discussed in detail in Section 6.2.2. In the present observations, a polarizer parallel to the magnetic field was used, so that only three of the nine components (the π light) appeared. Furthermore, natural mercury exhibits in the green line a large number of hyperfine structure components, and each of them forms a Zeeman pattern. To avoid a multiplicity of components in one spectral line, a separated isotope of mercury was used as the source. ^{198}Hg is well suited for this purpose since $I = 0$, and therefore it exhibits no hyperfine structure.

The optical system used for this investigation is shown in Fig. 6.14. The Fabry–Perot was crossed in the parallel-beam method with a small constant-deviation spectrograph (see Chapter 1). The etalon and lenses are all mounted on an optical bench to which the spectrograph is rigidly attached. The pair of lenses L_1 forms the light from the source into a parallel beam, while the pair L_2 focuses the Fabry–Perot ring pattern onto the spectrograph slit; the effective focal length of L_2 is 8 cm, and a further magnification of 2 takes place in the spectrograph.

The discharge tube is mounted vertically, as is the spectrograph slit; the slit width was 1 mm. It is clear that in this arrangement not only is the ring pattern focused onto the spectrometer slit but also the image of the source. A sheet of Polaroid film that could be rotated at will was used as a polarizer.

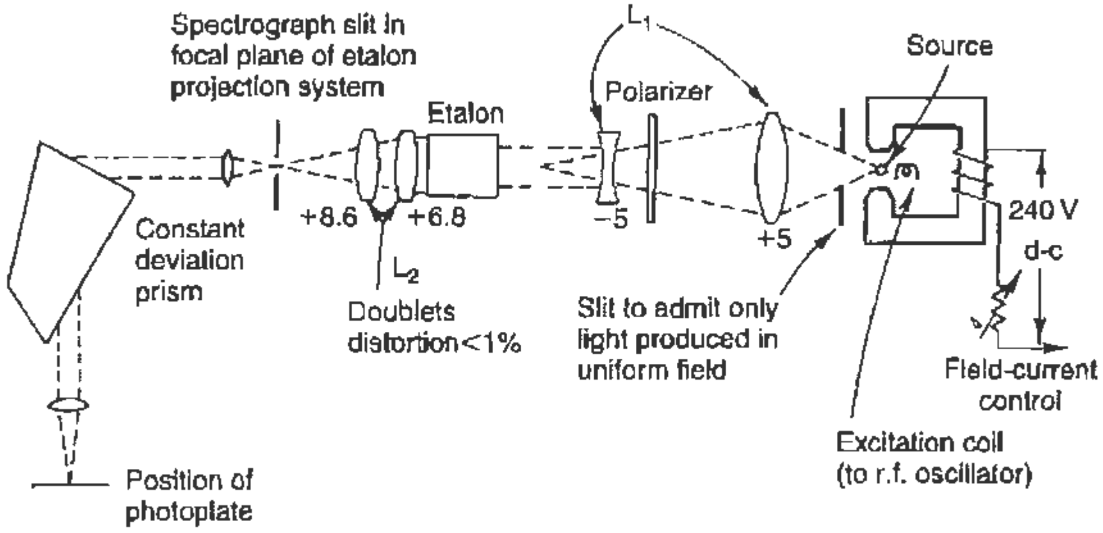

Spectrograph slit in focal plane of etalon projection system
Etalon
L_1
Polarizer
Source
Constant deviation prism
+8.6
+6.8
−5
+5
240 V d-c
L_2
Doublets distortion<1%
Slit to admit only light produced in uniform field
Field-current control
Excitation coil (to r.f. oscillator)
Position of photoplate

FIGURE 6.14 Experimental arrangement used for observing the Zeeman effect with a Fabry–Perot etalon, crossed by a constant-deviation prism spectrograph.

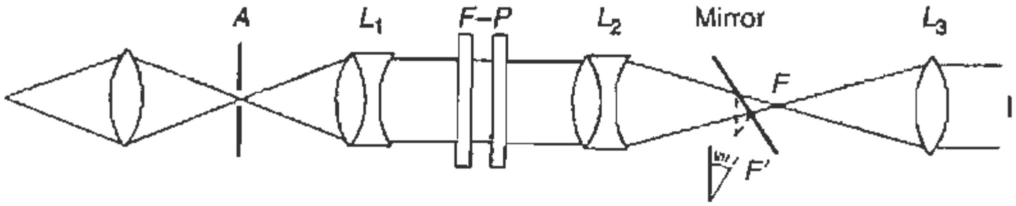


FIGURE 6.15 Optical arrangement for aligning a Fabry–Perot etalon. Rough adjustment is made by viewing the image formed by L_2. Final adjustment is made by viewing the etalon from the point F (or F').

The spacing of the Fabry–Perot etalon is $t = 0.5002$ cm; namely, the free spectral range is FSR = 30 GHz. It is important to adjust the plates carefully for parallelism. This can be done either by viewing through the spectrograph with a frosted glass in the focal plane, and adjusting for the best quality of the pattern, or by a much more sensitive arrangement as shown in Fig. 6.15. A very small aperture (less than 1 mm in diameter) is placed at the position of the source and illuminated with an intense sodium lamp. The Fabry–Perot plates are adjusted to be normal to the optical axis by bringing the image of A reflected by the etalon back onto A. Next, L_3 is adjusted until a series of multiple images of A appears when the observer is located at I; the plates of the etalon can then be roughly adjusted for parallelism by bringing all the images into coincidence. The final adjustment is made by removing L_3 so that the observer locates his eye at F (or a mirror can be used); then fringes of equal width do appear

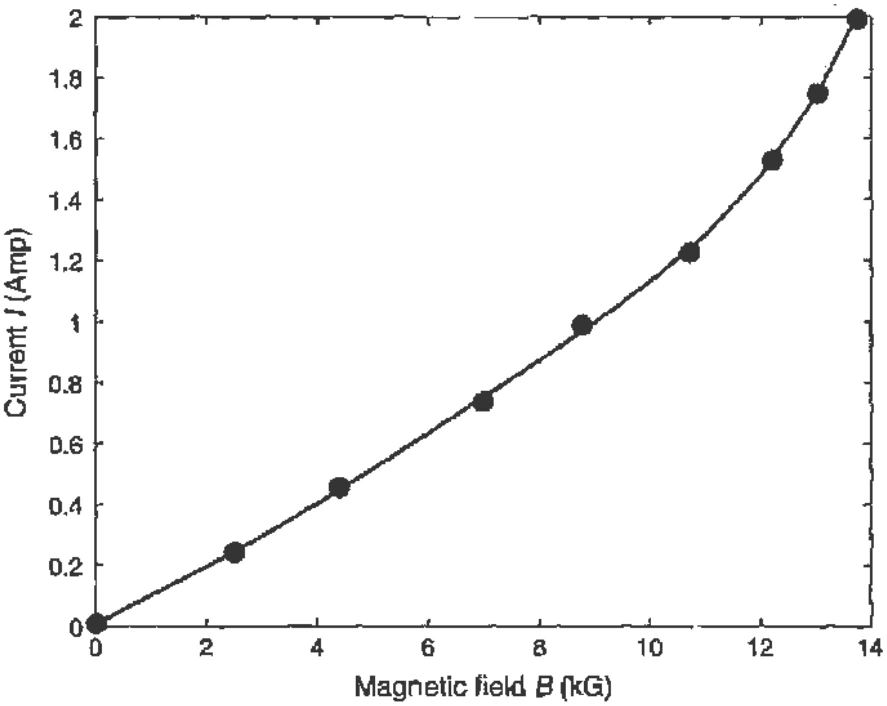


FIGURE 6.16 Calibration of the electromagnet used in the Zeeman effect experiment. The magnetic field is plotted against current; note the saturation at high fields.

parallel to the base of the wedge formed by the two plates. As the plates are moved into parallelism, the fringes become broader and finally the whole image of the aperture A seems to have a uniform illumination (bright or dark depending on the exact value of $n_0 = 2t/\lambda$). It is equally important that the ring pattern be in sharp focus at the plane of the photographic plate. For this experiment Kodak Royal-Pan film was used.

The electrodeless discharge tube was placed in a magnetic field. A small iron core electromagnet powered by a 220-V DC supply was used to produce the field. The diameter of the pole faces was only $1\frac{1}{2}$ in., and a small gap ($\frac{1}{2}$ in.) was used. By tapering the pole faces, higher magnetic fields can be achieved but this reduces the effective area of the field as well as the homogeneity. The magnetic field was measured with a "flip coil" and the calibration of field against current is given in Fig. 6.16. It is seen that field strengths of 1.2 T could be reached.

6.5.2. Data on the Zeeman Effect

The data presented below were obtained by students. Figure 6.17 shows the 546.1-nm Hg line photographed at various magnet settings. As explained

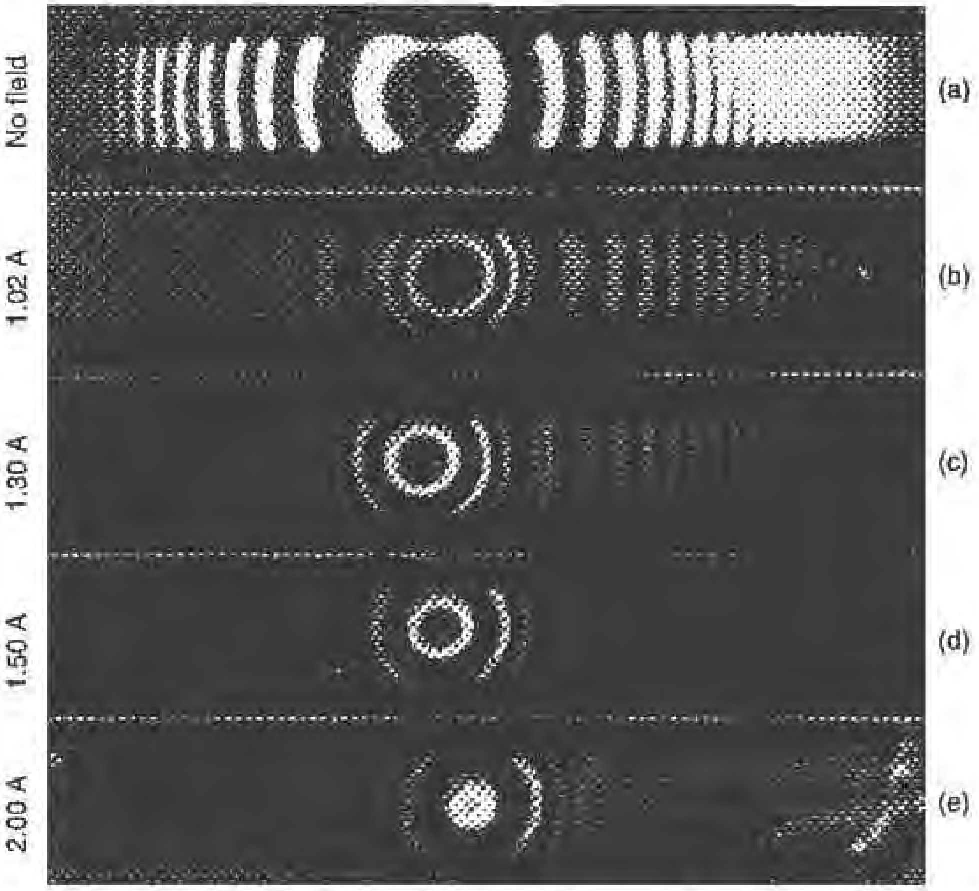


FIGURE 6.17 Fabry–Perot patterns showing the Zeeman effect of the green line of mercury. (See the text for additional details.) (a) No magnetic field applied. (b–e) A magnetic field of progressively greater strength is applied. Note the splitting of the original line into a triplet of increasing separation.

earlier, the source contains a single isotope, and the polarizer allows only the observation of π light. We note that the fringes are rather broad, but it can clearly be seen that when the field is applied the single–line pattern breaks up into a triplet, the separation between the components of the triplet becoming larger with increasing field.

The initial step in the reduction of the data is the measurement of the diameters (or radii) of the rings. To this effect a traveling microscope was used, and readings were taken directly off the plate; care must be taken to ensure that the travel of the microscope is indeed along the diameter of the rings and that the crosshairs are properly oriented. When the fringes in the pattern are as broad as those in Fig. 6.17, it is much more accurate to measure the two edges and take the average rather than try to set the crosshairs in the center of the fringe. The ring radii squared in the absence of the field provide the calibration of the data.

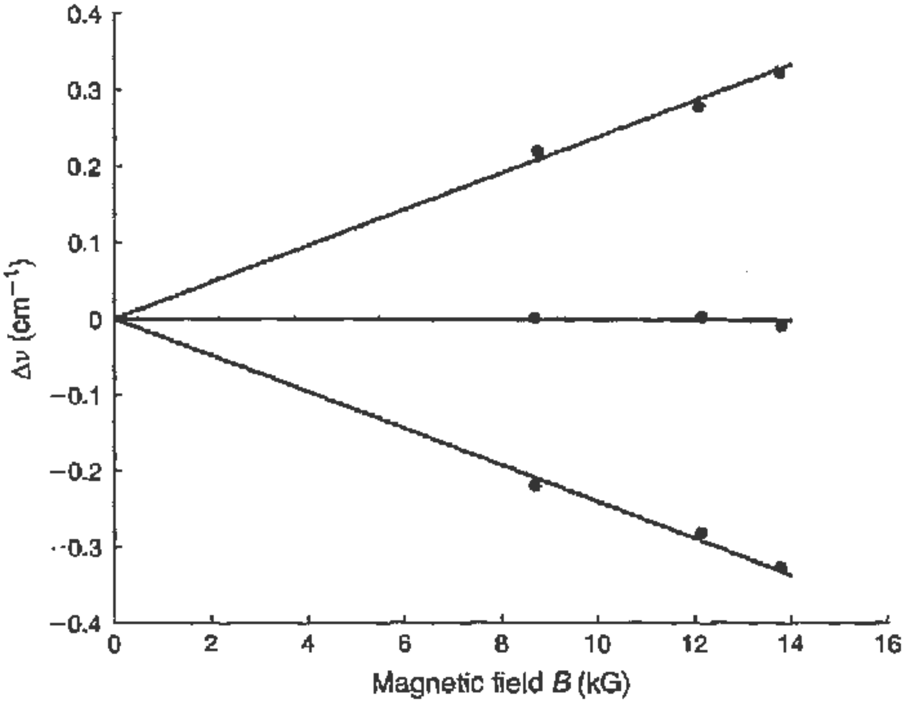


FIGURE 6.18 Results obtained on the Zeeman effect of the green line of mercury (see text). The observed displacement of the three components from the zero field value (of the single line) is plotted against the magnetic field.

Next the radii of the rings for the exposures taken at 1.0, 1.5, and 2.0 A were analyzed, and it was found that the central line is not shifted. However, the following shifts are observed for the outer rings for the 1.0-A data:

$$\Delta\nu_+ = 6.81 \text{ GHz} \qquad \Delta\nu_- = 6.60 \text{ GHz}.$$

The complete set of data is plotted in Fig. 6.18, and we see that as predicted the spacing varies linearly with the field, yielding

$$\Delta\nu = (7.2 \text{ GHz}) \times B(\text{T}). \tag{6.30}$$

The green line of Hg (546.1 nm) connects the ${}^3S_1$ state to the ${}^3P_2$. Its Zeeman splitting is shown in Fig. 6.19 where the g factors have been calculated according to Eq. (6.17). Since the polarizer was set to select only components arising in transitions with $\Delta m = 0$, we expect to observe only the three central components, which will be separated by

$$\Delta\nu = \frac{\mu_B}{h}\,(g_i - g_f)B = \frac{1}{2}\frac{\mu_B}{h}B. \tag{6.31}$$

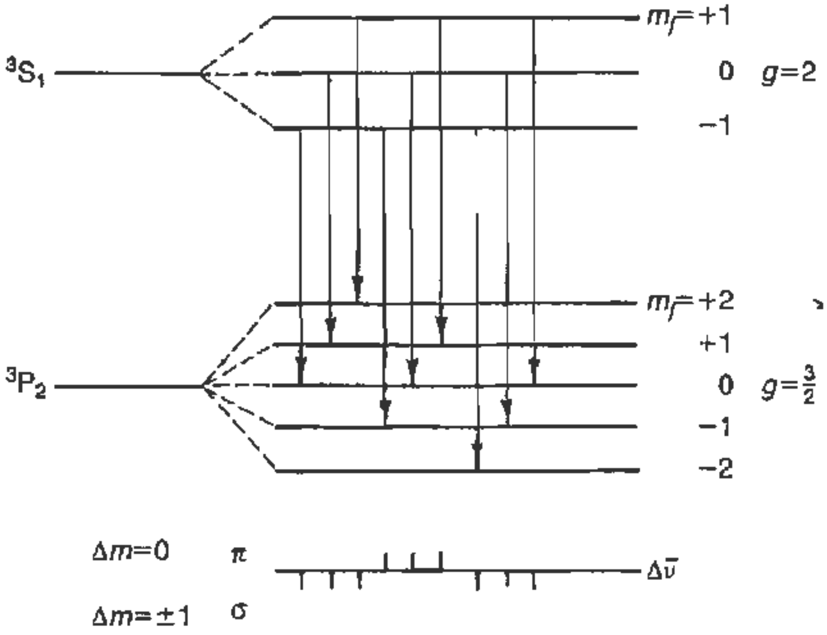


FIGURE 6.19 The Zeeman multiplet splitting of the 546.1-nm green line of Hg. It arises from a $^3S_1$ to $^3P_1$ transition.

By comparing with the experimental result of Eq. (6.30), we obtain

$$\mu_B = 5.95 \times 10^{-11}\ \mathrm{MeV/T}$$

in good agreement with the accepted value of

$$\mu_B = 5.79 \times 10^{-11}\ \mathrm{MeV/T}.$$

From these data we conclude that indeed spectral lines are split into components when the source is placed in a magnetic field. Further, the splitting observed was in excellent agreement with the theory of the anomalous Zeeman effect; the normal Zeeman effect can be excluded, since the energy difference between the components of the line was not $\mu_B B$ but $\frac{1}{2}\mu_B B$; compare to Eq. (6.1).

6.6. SATURATION ABSORPTION SPECTROSCOPY OF RUBIDIUM

6.6.1. Introduction

We mentioned in Section 6.4 that if an intense spectral line is passed through a region of dense atomic vapor of the same element it may become absorbed at the center of the Doppler pattern and develop self-reversal (see Fig. 6.11). We can take advantage of this effect to make measurements that are free from the Doppler effect.

Consider a monochromatic (laser) source of which we can sweep the frequency. This is easily achieved with many lasers and in particular with diode lasers as discussed in the following section. The light (the pump beam) is incident on a vapor cell and tuned in the region of a strong line from the ground to an excited state. If one monitors the transmitted light as a function of frequency, a Doppler-broadened absorption spectrum, such as shown schematically in Fig. 6.20a, will be observed.

Next we take light from the same source and direct it through the cell in the opposite direction and monitor the transmission at D_2. This is the probe beam, the experimental arrangement being as shown in Fig. 6.21. The signal at D_2 will exhibit the same general behavior as D_1 except that there will be a sharp spike at the center of the profile: see Fig. 6.20b.

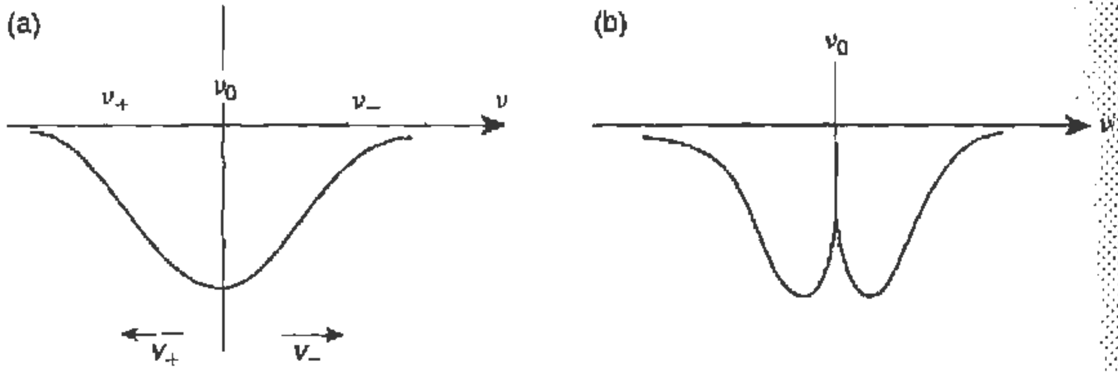


FIGURE 6.20 Absorption profiles of a resonance line: (a) The pump beam and (b) the probe beam.

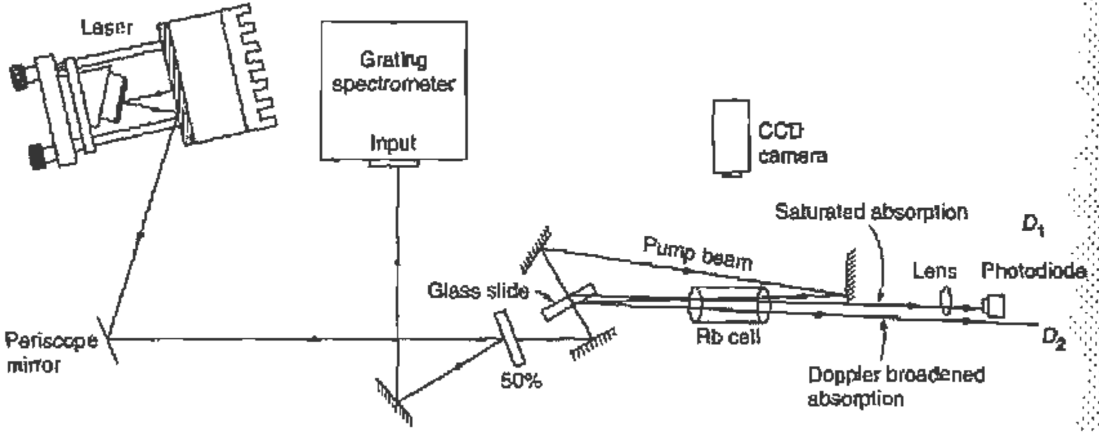


FIGURE 6.21 Schematic layout of the saturation absorption experiment.

Let us examine what happens when the pump beam of frequency ν_+ (refer to Fig. 6.20a) is incident on the cell: it excites atoms with a particular velocity v_+ moving toward the wave vector of the laser beam. When the pump has frequency ν_- it excites atoms that move in the same direction as the wave vector $\mathbf{k}_P$ with velocity v_-. At ν_0 the excited atoms have no velocity component along $\mathbf{k}_P$. The probe beam has the *same* frequency as the pump at all times but its k vector is opposite to $\mathbf{k}_P$. Thus when $\nu_L = \nu_+$, the atoms excited by the pump cannot absorb photons from the probe since they are moving in the v_+ direction, namely along the probe wave vector; similarly when $\nu_L = \nu_-$. However, when $\nu_L = \nu_0$ the atoms that could absorb the probe beam are already in the excited state due to the presence of the pump beam. As a result there is less absorption and a spike appears in the profile when ν sweeps through ν_0. The spike is very narrow as compared to the Doppler profile.

The situation becomes more complicated when there are several lines (that is, hyperfine structure) under the Doppler profile. For a single line of frequency ν_0 we found that the spike appears at ν_0. For two lines present at ν_1 and ν_2, one will see spikes not only when the laser frequency reaches $\nu_L = \nu_1, \nu_2$ but also when[17]

$$\nu_L = (\nu_1 + \nu_2)/2. \tag{6.32}$$

Such spikes are "crossover" lines and are often stronger than the direct lines.

Saturation spectroscopy can be easily observed in rubidium, cesium, and sodium and is used to lock lasers to a narrow frequency. For a practical

[17] Note that if for the laser frequency ν_L the Doppler shift (for the pump beam) by a class of atoms with velocity v_α is ν_α, then the state that is excited has frequency ν_1 where

$$\nu_L + \nu_\alpha = \nu_1.$$

For the probe beam the effective frequency (for this *same* class of atoms) is

$$\nu_L - \nu_\alpha.$$

If this frequency happens to correspond to another atomic transition, say at frequency ν_2, then the absorption will again be saturated. Therefore the condition is

$$\nu_L - \nu_\alpha = \nu_2$$

or

$$\nu_L = (\nu_1 + \nu_2)/2$$

as given by Eq. (6.32).

apparatus that can be used in a teaching laboratory, Thorlabs markets a complete setup to demonstrate the effect. An excellent description of the experimental details can be found in a classic paper by K. B. MacAdam, A. Steinbach, and C. Wieman, *Am. J. Phys.* **60**, 1098 (1992).

6.6.2. The Rubidium hfs Spectrum

Rubidium is an alkali ($Z = 37$) with a single $5s$ valence electron outside the closed shell of krypton ($n = 1, 2$, and 3 fully filled, $4s^2 4p^6$). Natural rubidium has two isotopes

$$
\begin{array}{lll}
{}^{85}\text{Rb} & \text{with nuclear spin} & I = \frac{5}{2} \\
{}^{87}\text{Rb} & & I = \frac{3}{2}.
\end{array}
$$

In the absence of nuclear spin the ground state is a ${}^1S_{1/2}$ state and the first excited states are ${}^2P_{1/2}$ and ${}^2P_{3/2}$. When the nuclear spin is included the energy level diagram is as shown in Fig. 6.22.

We will work with a single isotope, ^{85}Rb, and consider the transitions from the ground state to the $P_{3/2}$ excited state. In this case the ground state

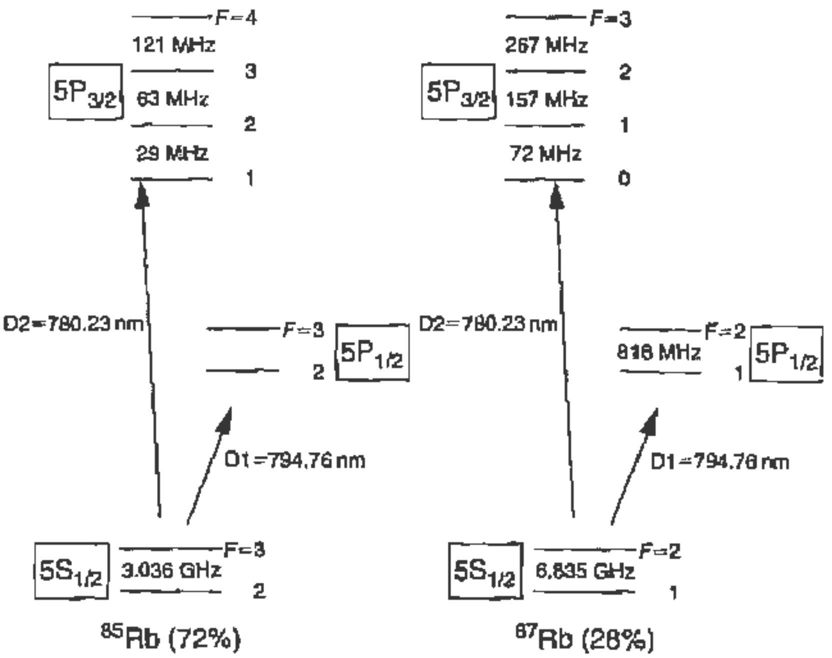


FIGURE 6.22 Energy level diagram of the low-lying atomic states of rubidium: (a) ^{85}Rb and (b) ^{87}Rb.

has two F levels

$$F = 3 \qquad \text{and} \qquad F = 2,$$

whereas the excited state has four F levels

$$F = 4,\ 3,\ 2,\ \text{and}\ 1.$$

As can be seen from Fig. 6.22 the hfs in the ground state is quite large, of the order of 3 GHz, so that one can tune the laser to select transitions from either the $F = 2$ or $F = 3$ state. Obviously the $P_{1/2}$ state is too far away to cause confusion. However, the Doppler profile, which is of the order of 1.0 GHz, covers all four hfs levels of the excited state. Recall that only transitions with $\Delta F = 0, \pm 1$ are allowed for electric dipole.

The laser frequency must be at 780.23 nm, which is in the infrared. It is conveniently obtainable from a diode laser. The diode laser is mounted in an external cavity, which is used to select the desired wavelength and can deliver up to 10 mW of power. Usually it suffices to send 3 mW to the pump beam and only a tenth of that to the probe beam.

6.6.3. Saturation Absorption Experiment

The overall schematic of the experiment is shown in Fig. 6.21. The diode laser is mounted in the heat sink on a thermoelectric cooler to adjust its temperature. The cavity is completed by a grating that returns the first-order diffraction peak into the laser. Thus, the frequency is tuned by adjusting the grating angle with piezo controls.

The diode laser output is a very strong function of laser temperature. Figure 6.23 shows such a calibration curve, and one selects the appropriate temperature with the help of a medium resolution spectrometer. Then the piezo is set to sweep the frequency, and one adjusts the laser current to shift the central frequency while the pump beam is going through the cell. At some point one will observe fluorescence, with an IR viewer or a CCD camera, or by monitoring the transmitted beam.

At this point one can reduce the sweep and setup for saturation absorption measurements. It is convenient to display the probe beam on a scope with the sweep on the horizontal axis. A picture of the observed fluorescence and of the saturated absorption of the probe beam are shown in Fig. 6.24. It is always possible to run a second low-intensity beam through

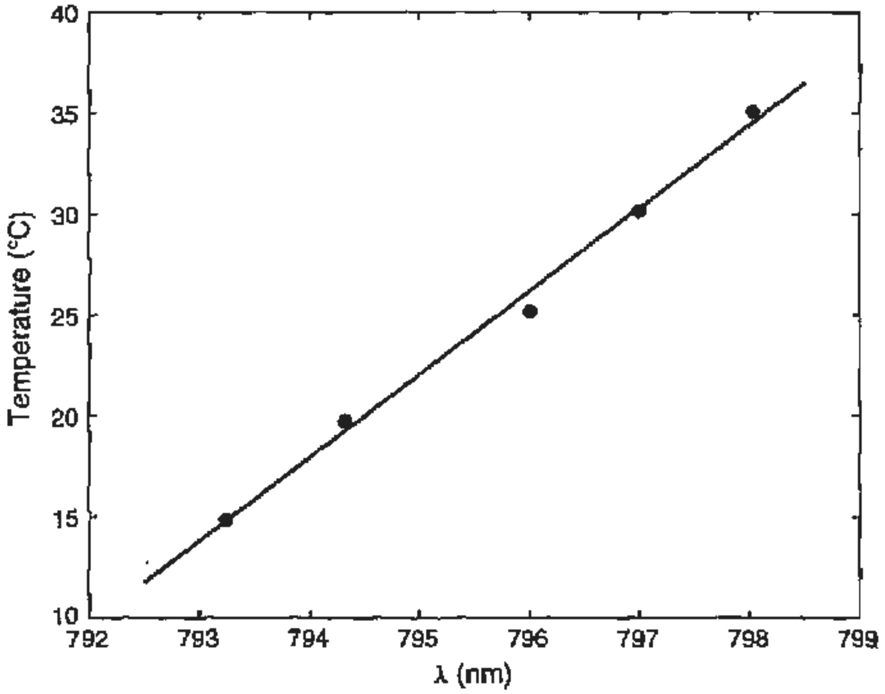


FIGURE 6.23 Wavelength as a function of temperature for the diode laser used in the experiment.

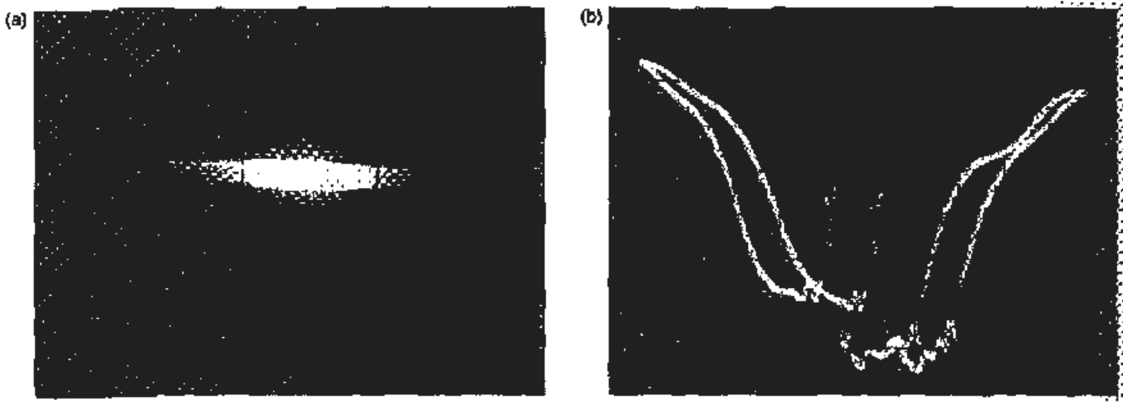


FIGURE 6.24 (a) Fluorescence emitted by the pump beam when properly tuned onto the Rb resonance line. (b) The probe beam signal when the frequency is swept over the entire Doppler peak. The displaced curves are due to hysteresis in the piezo electric driver.

the nonsaturated part of the cell to obtain the Doppler absorption profile and subtract it from the saturated absorption.

Data obtained by students on ^{85}Rb pumping from the $F = 3$ ground state are shown in Fig. 6.25. The two prominent lines are the crossover lines $[\nu(F'=2)+\nu(F'=4)]/2$, and $[\nu = (F'=3)+\nu(F'=4)]/2$, and the $\nu(F'=4)$ line can also be distinguished. On the assumption that the sweep is linear, the position of the other expected lines is indicated.

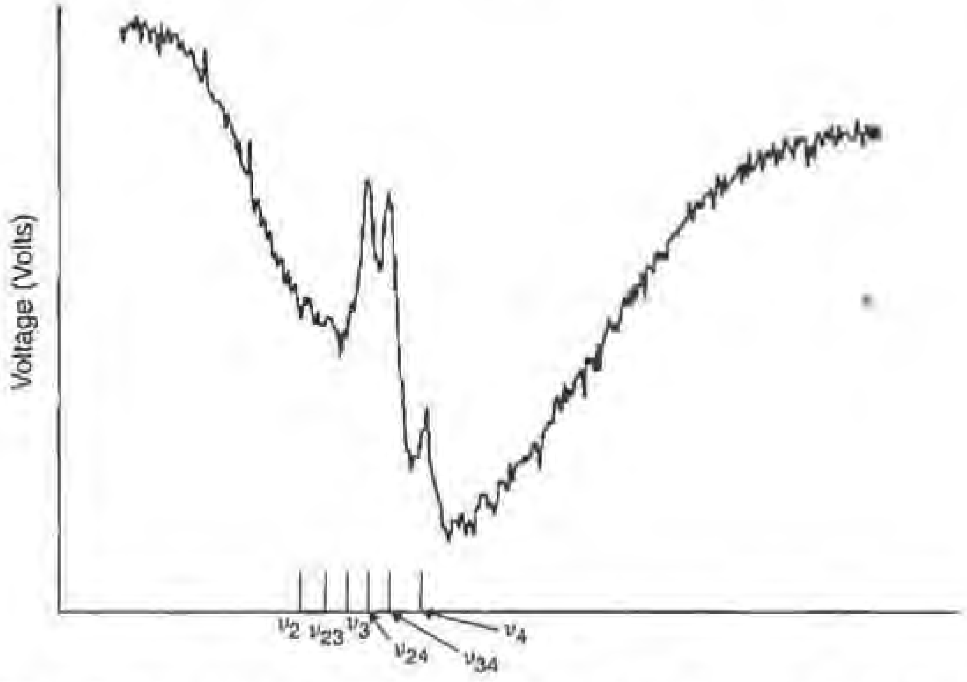


FIGURE 6.25 Saturation absorption spectrum obtained by students for ^{85}Rb ($F = 3 \rightarrow F'$). The position of all expected lines is indicated.

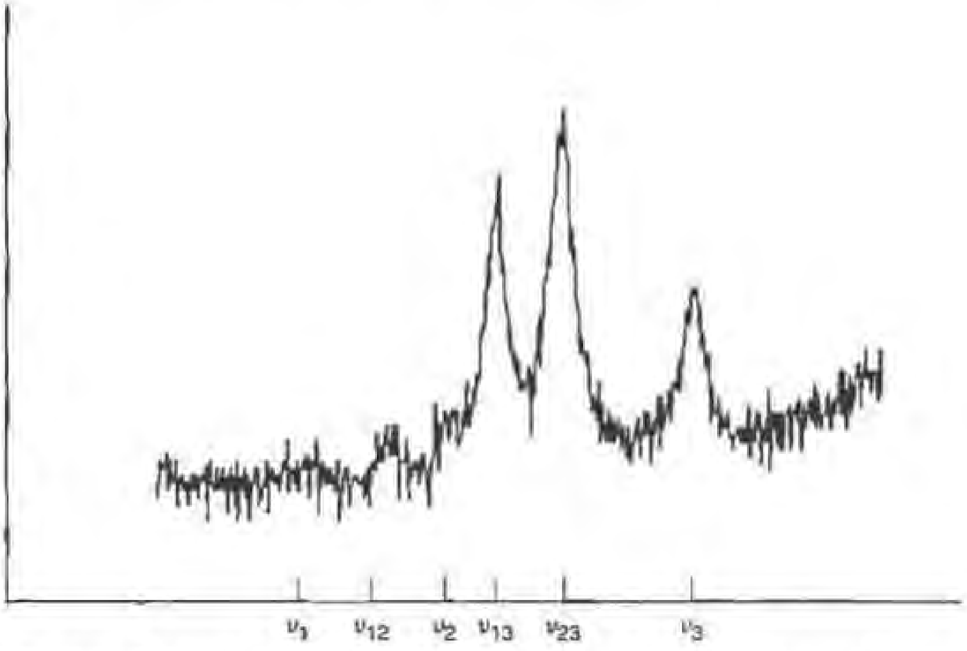


FIGURE 6.26 Subtracted saturation absorption spectrum obtained by students for ^{87}Rb ($F = 2 \rightarrow F'$). The position of all expected lines is indicated.

Finally Fig. 6.26 gives the subtracted saturated absorption spectrum for ^{87}Rb starting from the $F = 2$ ground state. Again the prominent lines are the crossover lines $[\nu(F' = 1)+\nu(F' = 3)]/2$ and $[\nu(F' = 2)+\nu(F' = 3)]/2$; the $\nu(F' = 3)$ line is also evident. The location of the other expected lines is indicated.

As is evident from the data the saturated absorption lines are very sharp. Thus instead of sweeping the laser frequency one can use a servo circuit to keep the laser frequency fixed on one of the lines (actually on its slope), reaching a stability of ± few megahertz, in absolute terms.

6.7. REFERENCES

E. U. Condon and G. H. Shortley, *The Theory of Atomic Spectra*, Cambridge Univ. Press, Cambridge, UK, 1951. This is one of the most complete theoretical treatments on atomic spectroscopy, but at an advanced level.

H. E. White, *Introduction to Atomic Spectra*, McGraw-Hill, New York, 1934. This book contains extensive data on atomic spectra, and the treatment of the theory is based on the semiclassical approach of the vector model.

H. Kuhn, *Atomic Spectra*, Longman's, London, 1962. A good book on a slightly more advanced level than White's book referred to above.

S. Tolansky, *High Resolution Spectroscopy*, Methuen, London, 1947. A very comprehensive and clear treatise on the instruments and techniques of high-resolution spectroscopy.

H. Kopferman, *Nuclear Moments*, Academic Press, New York, 1958. This book contains a very complete discussion of atomic hyperfine structure, of analysis methods, and of the conclusions obtained from it.

W. Demtröeder, *Laser Spectroscopy*, 2nd ed., Springer-Verlag, Berlin, 1996. A very comprehensive and up-to-date coverage of the field.

CHAPTER 7

Magnetic Resonance Experiments

7.1. INTRODUCTION

We saw in the previous chapter that when an atom (or a nucleus), with angular momentum **L** (or **I**), different from 0, is placed in a magnetic field **B** the states that correspond to different values of the quantum number m acquire an additional energy

$$\Delta E = \frac{\mu}{L} Bm. \tag{7.1}$$

Here μ is the "magnetic moment" of the atom or nucleus. When electrons are involved, μ is on the order of the Bohr magneton μ_B while for nuclei μ is on the order of the nuclear magneton, μ_N. In convenient units

$$\begin{aligned} \mu_B/h &= 14.01 \quad \text{GHz/T} \\ \mu_N/h = (\mu_B/h)/1836 &= 7.62 \quad \text{MHz/T.} \end{aligned} \tag{7.2}$$

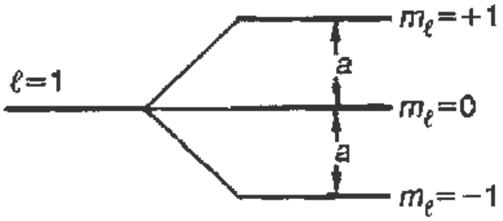


FIGURE 7.1 Splitting of an energy level with $l = 1$ into three components when placed in a magnetic field.

In Fig. 7.1 is shown the splitting of an energy state with $l = 1$ into its three sublevels. As discussed in Chapter 6, in optical spectroscopy we do not observe the spontaneous transitions (labeled a in the figure) between sublevels with different m, because they do not satisfy the selection rule $\Delta l = \pm 1$. Instead the splitting of a level is observed through the small difference in the frequency of the radiation emitted in the transitions between widely distant levels (with $\Delta l = \pm 1$). It is clear that if we could directly measure the frequency corresponding to a transition between the m sublevels of the same state, a much more precise knowledge of the energy splitting would be obtained.

The selection rule $\Delta l = \pm 1$ is applicable to electric-dipole radiation; however, transitions with $\Delta l = 0$, $\Delta m = \pm 1$ do occur when *magnetic-dipole* radiation is emitted, but the probability for such a transition is reduced by a factor[1] $(v/c)^2$ from the case of an electric dipole transition. We therefore conclude that spontaneous transitions with $\Delta l = 0$, $\Delta m = \pm 1$ will be very rare, especially if the system can preferentially return to its ground state (lowest energy state) by a $\Delta l = \pm 1$ transition. On the other hand, in the presence of an electromagnetic field, *induced* transitions have a probability of occurring if the frequency of the field is equal (or at least fairly close) to the energy difference between the two levels; induced transitions toward higher or lower energy states are equally probable. Further, the transition probability is proportional to the square of the strength of the electromagnetic field (that is, the total number of quanta) so that if a sufficiently strong radiofrequency magnetic field (of frequency ν_0) is available, magnetic-dipole transitions should take place.

This fact is, of course, central to the operation of the laser discussed in Section 4.1. In that case the atomic state has an electric-dipole moment and

[1] For atomic systems v is on the order of the velocity in a Bohr orbit, namely, $(v/c)^2 \approx 5 \times 10^{-6}$.

electric-dipole transitions are induced by the external electric field (at the optical frequency) of the laser beam.

By referring to Eq. (7.2) we see that for a 1-T magnetic field the energy splitting of either nuclei or electrons falls in the range of frequencies that can be easily generated. It is also of interest to estimate the magnitude of the *radiofrequency* (or microwave) magnetizing field, which we will designate by H, to distinguish it from the static magnetic (induction) field B; in vacuum $B = \mu_0 H$. An H field of magnitude $10^3/4\pi$ A/m (equivalent to a B field of 10^{-4} T $= 1$ G) corresponds to an energy flow of

$$\langle S \rangle = \frac{1}{2}\sqrt{\frac{\mu_0}{\epsilon_0}}\, H^2 = \frac{1}{2}\sqrt{\frac{4\pi \times 10^{-7}}{8.85 \times 10^{-12}}} \times \left(\frac{10^3}{4\pi}\right)^2 \approx 2.35 \times 10^2\, \frac{\text{W}}{\text{cm}^2}, \tag{7.3}$$

which can be easily generated. Calculation shows that this field strength is adequate for inducing transitions. Finally we must be able to detect the fact that a transition took place; this may be done in several ways and is one of the distinguishing factors between the various types of magnetic resonance experiments.

For example, in the first magnetic resonance experiment, performed by I. I. Rabi and collaborators in 1939, a beam of atoms having $J = \frac{1}{2}$ was passed in succession through two very inhomogeneous magnets A and B shown in Fig. 7.2. A homogeneous magnetic field existed in the intermediate region C where a radiofrequency (RF) field was applied. If a transition took place in region C from a state $m = +\frac{1}{2}$ to $m = -\frac{1}{2}$, that particular atom was deflected in an opposite direction in field B and thus missed the detector. Hence, resonance was detected by a decrease in beam current when the frequency of the RF field was the appropriate one for the magnetic field strength in C.

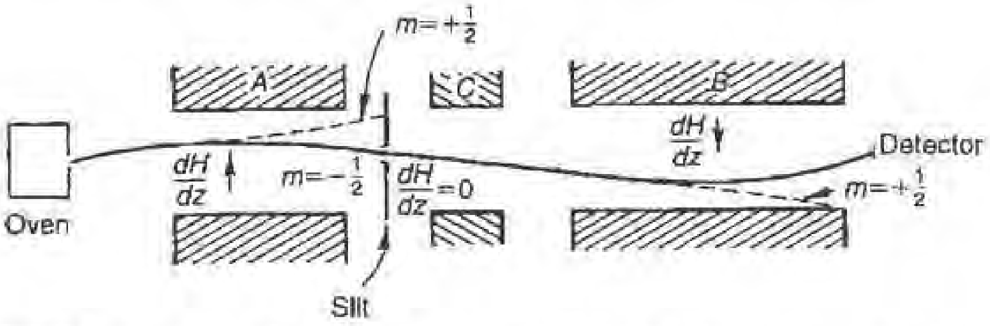


FIGURE 7.2 The atomic beam arrangement of I. I. Rabi and collaborators used to detect magnetic resonance transitions in atomic energy levels.

Another method for detecting the occurrence of resonance is to observe the absorption of energy from the radiofrequency field when transitions toward higher energy levels take place. This technique is used in most nuclear magnetic resonance (NMR) experiments and in electron magnetic resonance (called "electron spin resonance," ESR) experiments. In experiments with atomic vapors or transparent materials it is possible to detect the magnetic resonance effect by changes in the polarization of the atomic radiation ($\Delta m \neq 0$) or by selective absorption effects.

Apart from its intrinsic interest as a way of inducing transitions between the energy sublevels of atoms or nuclei, magnetic resonance has become an important tool of physics. The atomic beam experiments of Rabi and his coworkers led to very precise measurements of the hyperfine structure of atomic systems and thus to accurate values of the nuclear moments. In a nuclear magnetic resonance experiment transitions are induced between the sublevels of a nucleus placed in an external magnetic field. However, the atom to which the nucleus belongs must have $J = 0$ (diamagnetic material), since otherwise the nuclear spin would be coupled to J and the large electronic magnetic moment would mask the effect. By means of such experiments, nuclear magnetic moments are measured directly and to a high accuracy.

The NMR signal depends not only on the nucleus under study but also on the environment in which the nucleus finds itself. In fact the observation of nuclear magnetic resonance in solids and liquids depends on the relaxation of the nuclear spins through their interaction with the lattice. Thus, nuclear magnetic resonance studies have yielded a very large amount of information on the properties of many materials in the solid or liquid state.

Soon after the first successful nuclear magnetic resonance experiments, it was realized that the width of the observed resonance line for protons was mostly due to inhomogeneities in the constant magnetic field used to split the energy sublevels. When a very homogeneous field was applied, the proton resonance line was shown to exhibit a fine structure on the order of 0.01 G (10^{-6} T). This structure depends on the organic compound to which the hydrogens of the sample belong. With even more homogeneous fields a hyperfine structure on the order of 0.001 G (10^{-7} T) is observed. It is this fine structure that has made NMR such an important tool for analytical chemistry.

The term electron spin (or paramagnetic) resonance is used for transitions between the Zeeman levels of quasi-free electrons in liquids and solids. In principle, we should always measure a g factor of 2.00 (if we deal with free electrons); instead a great variety of g factors and structure appears in the resonance lines due to the different effective coupling of the electron with the crystalline field. These effects depend on the relative orientation of the magnetic field B_0 and the crystal axis. Thus, electron spin resonance is a very important tool in the study of crystalline structures as well as in the identification of free radicals in chemistry, medicine, and biophysics.

This chapter is organized as follows. In Section 7.2 the conditions for inducing magnetic-dipole transitions are discussed from both the quantum and classical point of view. In Section 7.3 we introduce the mechanisms essential for the observation of energy absorption in nuclear magnetic resonance and electron spin resonance experiments, namely relaxation and saturation. We also discuss the idea of free induction decay and pulsed NMR. The techniques and results of nuclear magnetic resonance experiments with protons are presented in Section 7.4. We conclude with a discussion of an electron spin resonance experiment that operates at microwave frequencies.

As was the case in the previous chapter the discussion is limited, and the reader may wish to refer to some of the many excellent monographs and texts on this subject. A list of suggested references is given at the end of the chapter.

7.2. THE RATE FOR MAGNETIC-DIPOLE TRANSITIONS

7.2.1. Quantum Calculation

The experimental signals in NMR involve the participation of many nuclei. In this section, however, we will consider the effects associated with a single nucleus: we use the term *a single spin*. We will return to an ensemble of nuclei in Section 7.3.

Let us consider, for example, a nucleus with angular momentum **I** (magnitude $\hbar\sqrt{I(I+1)}$) and magnetic moment μ oriented along the spin axis. For nuclei it is customary to express the proportionality between the spin **I** and magnetic moment μ by

$$\mu = \gamma\hbar\mathbf{I}, \tag{7.4}$$

where γ is called the gyromagnetic ratio; as can be seen from Eq. (7.6) below, γ has dimensions of radians per second–tesla. The gyromagnetic

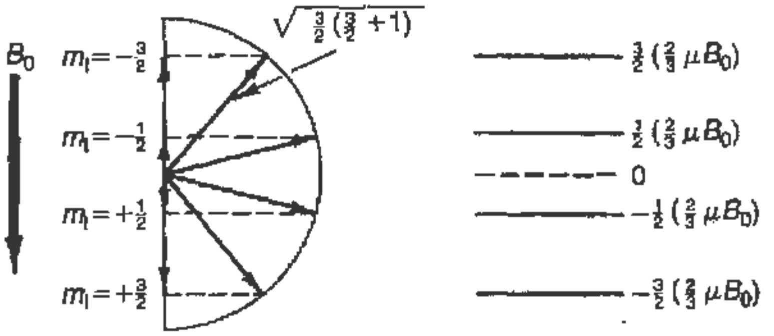

$\sqrt{\frac{3}{2}(\frac{3}{2}+1)}$
B_0
$m_I=-\frac{3}{2}$
$m_I=-\frac{1}{2}$
$m_I=+\frac{1}{2}$
$m_I=+\frac{3}{2}$
$\frac{3}{2}(\frac{2}{3}\mu B_0)$
$\frac{1}{2}(\frac{2}{3}\mu B_0)$
0
$-\frac{1}{2}(\frac{2}{3}\mu B_0)$
$-\frac{3}{2}(\frac{2}{3}\mu B_0)$

FIGURE 7.3 The energy of the four sublevels of a nucleus with spin $I = \frac{3}{2}$ when placed in a magnetic field B_0. Note that the energy depends on the "orientation" of the spin with respect to B_0; the magnitude of the spin vector is $|\mathbf{I}| = \sqrt{\frac{3}{2}(\frac{3}{2}+1)}$.

ratio γ cannot be calculated from a simple expression such as found for the g factor of atomic electrons in Eq. (6.17). (For instance, for the proton $\gamma = 5.586\,\mu_N$, where μ_N is the nuclear magneton.)

In the presence of an external magnetic field B_0, the nucleus can be in any of the $(2I + 1)$ sublevels labeled by m_I as shown also in Fig. 7.3. We can then write for the energy[2] of these sublevels (see Eq. (7.1))

$$\frac{E}{\hbar} = -\frac{1}{\hbar}\frac{\mu}{I}\,B_0 m = -\gamma B_0 m, \tag{7.5}$$

so that the energy difference between any adjacent sublevels ($\Delta m = \pm 1$) is simply

$$\frac{\Delta E}{\hbar} = \gamma B_0 = \omega_0. \tag{7.6}$$

Thus for protons in a field of 1 T the resonance frequency will be

$$\nu_0 = 5.586\mu_N B_0 = 42.581 \text{ MHz} \qquad (B_0 = 1 \text{ T}).$$

Consider then the simplest case, namely, $I = \frac{1}{2}$, for which only two sublevels exist, $m = -\frac{1}{2}$ and $m = +\frac{1}{2}$. In addition to B_0, let a weak field H_1, rotating in a plane normal to B_0 with an angular frequency ω be introduced. Taking the z axis along B_0 we write the two components of H_1 as

$$(H_1)_x = H_x = H_1 \cos\omega t \qquad (H_1)_y = H_y = H_1 \sin\omega t,$$

[2] Instead of energy, we use for convenience angular frequency; the transition frequency is $\Delta\nu = (\Delta E/\hbar)/(2\pi) = \omega_0/2\pi$.

and we assume that

$$\mu_0 H_1 \ll B_0.$$

The additional energy of the nucleus, due to the field H_1, is

$$\mathcal{H}_1 = \mu \cdot \mathbf{H}_1 = \gamma\hbar\left(H_x I_x + H_y I_y\right) = \frac{\gamma\hbar H_1}{2}\left(I_+ e^{-i\omega t} + I_- e^{+i\omega t}\right), \tag{7.7}$$

where[3]

$$I_+ = I_x + i I_y \qquad \text{and} \qquad I_- = I_x - i I_y. \tag{7.8}$$

Since the energy specified by Eq. (7.7) is very small as compared to that given by Eq. (7.5), it can be treated as a time-dependent perturbation[4]; thus, to first order, the transition probability is proportional to the absolute square of the matrix element

$$\mathcal{M} = \frac{\gamma\hbar H_1}{2}\ \langle f | I_+ e^{-i\omega t} + I_- e^{i\omega t} | i \rangle, \tag{7.9}$$

where i and f stand for the initial and final state. As usual the matrix element is evaluated by performing the integral

$$\mathcal{M} = \int \psi_f^* \mathcal{H}_1 \psi_i d^3x\, dt, \tag{7.10}$$

where $\mathcal{H}_1$ is the perturbing energy of Eq. (7.7). We must include the time dependence of the wave functions

$$\begin{aligned}\psi_f &= u(I, m') \exp\left(-i\tfrac{E'}{\hbar}t\right)\\ \psi_i &= u(I, m) \exp\left(-i\tfrac{E}{\hbar}t\right).\end{aligned} \tag{7.11}$$

Here primes refer to the final state, and $u(I, m)$ stands for the time-independent part of the wave function. Evaluating Eq. (7.9) with the help

[3] We expand the exponentials and obtain

$$\begin{aligned}&(I_x \cos\omega t\ +\ i I_y(-i)\sin\omega t) + (I_x \cos\omega t - i I_y(+i)\sin\omega t)\\ &\quad = 2(I_x \cos\omega t + I_y \sin\omega t).\end{aligned}$$

[4] See, for example, E. Fermi, *Notes on Quantum Mechanics*, Lecture 23, Univ. of Chicago Press, Chicago, 1961.

of Eqs. (7.10) and (7.11) we find that

$$\mathcal{M} = \frac{\gamma \hbar H_1}{2} \left\{ \langle I, m' | I_+ | I, m \rangle \int \exp\left[-i \left(\frac{E - E'}{\hbar} + \omega \right) t \right] dt \right.$$
$$\left. + \langle I, m' | I_- | I, m \rangle \int \exp\left[-i \left(\frac{E - E'}{\hbar} - \omega \right) t \right] dt \right\}. \tag{7.12}$$

The matrix elements of the operators I_+ and I_- are[5]

$$\langle m' | I_+ | m \rangle = \sqrt{I(I+1) - m(m+1)}\ \delta_{m',m+1}$$
$$\langle m' | I_- | m \rangle = \sqrt{I(I+1) - m(m-1)}\ \delta_{m',m-1},$$

and thus I_+ connects only states with $m' - m = 1$ while I_- connects states $m' - m = -1$. For $I = \frac{1}{2}$ the above matrix elements reduce to 1 for either I_+ or I_-. The integrals over time in Eq. (7.12) are essentially δ functions (but see below) expressing the conservation of energy and showing that the transition probability is different from zero only if

$$E' - E = \hbar\omega \qquad \text{for } m' = m + 1$$

and

$$E - E' = \hbar\omega \qquad \text{for } m' = m - 1, \tag{7.13}$$

that is, when the angular frequency of the rotating field is equal to the energy difference between adjacent m sublevels. Using Eq. (7.6), the conditions of Eqs. (7.13) become simply

$$\hbar\omega = \hbar\gamma B_0 = \hbar\omega_0.$$

To complete the calculation of the transition rate we must integrate (the absolute square of Eq. (7.12)) over the density of final states. This leads to Fermi's golden rule[6]

$$R_{if} = \frac{2\pi}{\hbar} |\mathcal{M}|^2 \rho(E), \tag{7.14}$$

[5] See E. Fermi (1961), Lecture 28.

[6] See E. Fermi (1961), or L. Shiff, *Quantum Mechanics*, Chapter 8, McGraw-Hill, New York, 1968.

where R_{if} is the *transition probability per unit time* (or transition rate) from the initial state i to the final state f. In Eq. (7.14), $\mathcal{M}$ is the time-independent part of the matrix element given by Eq. (7.12) (that is, without the integrals). $\rho(E)$ is the "density of final states" and gives the number of states f per unit energy interval that have energy close to E'. For example, if the final state f has an extremely well-defined energy E_0, then $\rho(E) \to \delta(E - E_0)$; if the final state has a certain width due for instance to a finite lifetime or other broadening effects, then $\rho(E)$ expresses this fact mathematically. We require the function $\rho(E)$ to be normalized and can also express it in terms of frequency

$$\rho(E) = \rho(h\nu) = \frac{1}{h}\, g(\nu)$$

with

$$\int \rho(E)\, dE = \int g(\nu)\, d\nu = 1. \tag{7.15}$$

Combining Eqs. (7.12), (7.14), and (7.15) we obtain for the transition rate in the case $I = \frac{1}{2}$ the elegant result

$$R_{-1/2\to+1/2} = R_{+1/2\to-1/2} = \frac{\gamma^2 H_1^2}{4}\, g(\nu). \tag{7.16}$$

In the above equation ν is the frequency of the perturbing field (RF or microwave), and $g(\nu)$ gives the shape of the resonance line; note that $g(\nu)$ will be significantly different from zero only for $\nu \approx \nu_0$. Note also that in Eq. (7.16) and in the equations leading up to it, H_1 must be expressed in tesla, namely its value in amperes per meter must be multiplied by the permeability of free space μ_0. We have deliberately not included this factor in the equations to avoid confusion with the symbol for magnetic moments.

There are two important comments we want to make at this point. First as can be seen from Eq. (7.12) or (7.16) the rotating field H_1 will induce transitions from $m_i = -\frac{1}{2}$ to $m_f = +\frac{1}{2}$ with exactly the same probability as from $m_f = +\frac{1}{2}$ to $m_i = -\frac{1}{2}$. As a result, in the presence of the field H_1 both levels will, on average, be *equally populated*. This argument remains valid for any value of the nuclear spin. Secondly, while we used a perturbative calculation the two-level system can be solved exactly in terms of simple functions as described, for instance, in the *Feynman Lectures*, Vol. III, Lecture 30.[7] We will make use of the exact solution in Section 7.3.4 when we discuss pulsed NMR and free induction decay.

7.2.2. Classical Interpretation

Below we show how the effect of a rotating radiofrequency field can be understood also on the basis of a classical model. Consider again a nucleus with spin **I** and magnetic moment $\mu = \gamma\hbar\mathbf{I}$. Let J be the magnitude of the angular momentum, which classically[8] will be just $J = \hbar I$, and let it make an angle θ with the z axis as shown in Fig. 7.4a. If a constant magnetic field B_0 is applied along the z axis, the field will exert a torque on the magnetic moment, given by

$$\boldsymbol{\tau} = \boldsymbol{\mu} \times \boldsymbol{B_0} = \gamma(\boldsymbol{J} \times \boldsymbol{B_0}). \tag{7.17}$$

This must equal the time derivative of the angular momentum

$$\frac{d\mathbf{J}}{dt} = \boldsymbol{\tau} = \gamma(\boldsymbol{J} \times \boldsymbol{B_0}). \tag{7.18}$$

The solution of Eq. (7.18) leads to a precession of the angular momentum vector **J** about the z axis, preserving the angle θ, and at an angular frequency ω_0 independent of θ,

$$\boldsymbol{\omega_0} = -\frac{|d\mathbf{J}/dt|}{|\boldsymbol{J} \times \boldsymbol{n}_z|}\,\mathbf{n}_z = -\gamma B_0 \mathbf{n}_z, \tag{7.19}$$

where $\mathbf{n}_z$ is the unit vector in the z direction.

This phenomenon is called the *Larmor precession* and the angular frequency given by Eq. (7.19) is the "Larmor" frequency. It is fascinating even though not surprising that the Larmor frequency has the same value as given by Eq. (7.6) for the transition frequency between any adjacent levels ($\Delta m = \pm 1$). Further, since the angle θ is preserved, the energy of the nucleus in the magnetic field remains a constant

$$E = -\boldsymbol{\mu} \cdot \boldsymbol{B_0} = -\gamma\hbar I B_0 \cos\theta. \tag{7.20}$$

We now introduce an additional weak magnetic field $\mathbf{H}_1$ oriented in the x–y plane and rotating about the z axis (in the same direction as the

[7] See also A. Das and A. C. Melissinos, *Quantum Mechanics*, Section 5.1, Gordon and Breach, New York, 1986.

[8] Instead of its quantum-mechanical (QM) value $I = \hbar\sqrt{I(I+1)}$.

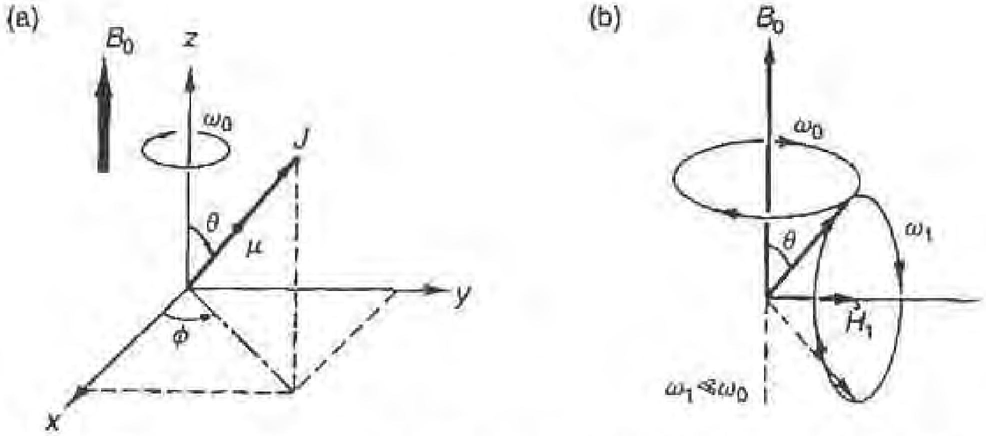


FIGURE 7.4 Precession of a magnetic moment μ when placed in a magnetic field B_0. (a) The spin precesses with angular frequency $\omega_0 = \gamma B_0$; the angle θ is a constant of the motion. (b) In addition to B_0 a weak magnetic field H_1 is now also applied. H_1 is rotating about the z axis with angular frequency ω_0 and therefore μ precesses about $\mathbf{H}_1$ with angular frequency $\omega_1 = \gamma H_1$; θ is no longer conserved.

"Larmor precessing" spin I) with an angular frequency ω. If the frequency ω is different from ω_0, the angle between the field $\mathbf{H}_1$ and the magnetic moment μ will continuously change so that their interaction will average out to 0. If, however, $\omega \approx \omega_0$, the angle between μ and $\mathbf{H}_1$ is maintained and a net interaction is effective (Fig. 7.4b). If we look at the system in a reference frame *rotating about the z axis with the angular velocity* ω_0, then the spin will appear to make an angle $\psi = 90° - \theta$ with H_1 and according to the previous argument will start to precess (in the rotating frame) about H_1. This corresponds to a "nutation" and a consequent change of the angle θ, which implies a change in the potential energy of the nucleus in the magnetic field (Eq. (7.20)). The change in θ is the classical analogy to a *transition* between sublevels with different m. We see that (a) such transitions may take place only if the rotating field has an angular frequency $\omega = \omega_0 = \gamma B_0$, and (b) that the angle θ will continuously change with an angular frequency $\omega_1 = \gamma H_1$. The effect of the radiofrequency is to populate, on the average, all values of θ, that is, *all levels, equally*.

However, if the field H_1 is applied only for a short time t, such that $\omega_1 t = \pi$, then a spin that was originally at an angle θ (w.r.t the z axis) will find itself at an angle $\pi - \theta$ (or at an angle θ from the $-z$ axis). This is the equivalent of the QM transition from $m = -\frac{1}{2}$ to $m = +\frac{1}{2}$. If the field is applied for a time t such that $\omega_1 t = 2\pi$, then the spin will end up at the same angle w.r.t the z axis (in the same state) and so on. By applying

RF pulses of selected duration we can thus manipulate the spin state. We will make use of this idea in Section 7.3.4.

7.3. ABSORPTION OF ENERGY BY THE NUCLEAR MOMENTS

7.3.1. Relaxation and Saturation

We saw in the previous section that a radiofrequency magnetic field may induce transitions between the magnetic sublevels of a nucleus, electron, or atom. In the case of atomic-beam experiments the atoms are free, while in nuclear magnetic resonance or electron spin resonance experiments the nuclei or electrons are in constant interaction with their surroundings. There are mainly two types of interactions: (a) *spin–lattice*, by which we mean the interaction with the thermal bath that tends to restore the Boltzmann distribution, where the spin can relax by transferring energy to the lattice; and (b) *spin–spin*, in which the nuclear spin interacts with a neighboring nuclear spin, but the total energy of the spin system remains constant. As a matter of fact, it is the spin–lattice interaction that makes possible the observation of energy absorption from the radiofrequency field when the resonance frequency is reached.

To understand this last statement, consider again the simple case of a nucleus with spin $I = \frac{1}{2}$. In the presence of a magnetic field B_0 it is split into the two energy sublevels with $m = +\frac{1}{2}$ and $m = -\frac{1}{2}$. As remarked before, the rate (Eq. (7.16)) for transitions

$$\left(m = +\frac{1}{2}\right) \to \left(m = -\frac{1}{2}\right) \tag{7.21a}$$

is equal to the rate for transitions

$$\left(m = -\frac{1}{2}\right) \to \left(m = +\frac{1}{2}\right). \tag{7.21b}$$

The number of transitions per unit time is given in either case by

$$R_{if} \times N_i, \tag{7.22}$$

where N_i is the number of nuclei in the initial state. Further, transitions of the type in Eq. (7.21a) absorb energy from the radiofrequency field, whereas transitions of the type in Eq. (7.21b) give energy to the radiofrequency field (recall Eq. (7.5)). Thus the net power absorbed from the radiofrequency field is (we also multiply by the energy necessary for one transition)

$$\begin{aligned} P &= \left[N_{+1/2} \times R\left(+\frac{1}{2} \to -\frac{1}{2}\right)\right]\hbar\omega_0 \\ &\quad - \left[N_{-1/2} \times R\left(-\frac{1}{2} \to +\frac{1}{2}\right)\right]\hbar\omega_0 \quad , \\ &= \left(N_{+1/2} - N_{-1/2}\right) R\hbar\omega_0 . \end{aligned} \tag{7.23}$$

Thus if $N_{-1/2} = N_{+1/2}$, no net power can be absorbed from the field. However, if we consider a system consisting of a large number of spins in equilibrium with its surroundings, it is known from a very general theorem of statistical mechanics that every state of energy E will be populated according to the Boltzmann distribution

$$N(E) = N_0 e^{-E/kT} \tag{7.24}$$

with k the Boltzmann constant and T the absolute temperature in Kelvins. It follows that for a system of N particles with spins I in the presence of a magnetic field B_0, each m sublevel will be populated according to

$$N(m) = \frac{N}{2I+1}\exp\left(+\frac{m\gamma\hbar B_0}{kT}\right). \tag{7.25}$$

The normalizing factor was approximated by $N/(2I+1)$, which holds[9] for $\gamma\hbar B_0 \ll kT$; $-m\gamma\hbar B_0$ is the energy of the m sublevel. Note that T in Eq. (7.25) is the temperature of the spin system and equals the lattice temperature, if no external perturbations (such as the radiofrequency field) are present.

It follows from Eq. (7.25) that the populations $N_{+1/2}$ and $N_{-1/2}$ entering Eq. (7.23) of our previous discussion ($I = \frac{1}{2}$) will not be equal. There will be a number of excess nuclei N_s, in the lower energy state given by

$$N_s = N_{+1/2} - N_{-1/2} = \frac{N}{2}\left[\exp\left(+\frac{\hbar\omega_0}{2kT}\right) - \exp\left(-\frac{\hbar\omega_0}{2kT}\right)\right],$$

[9] Expand the exponential through first order, to obtain correctly

$$\sum_{m=-I}^{m=+I} N(m) = N.$$

and since $\hbar\omega_0$ is always much smaller than kT, we may write for the above equation

$$N_s \approx \frac{N}{2}\frac{\hbar\omega_0}{kT}. \tag{7.26}$$

It is only these N_s nuclei that can contribute toward a net absorption of energy, and the power absorbed from the RF field is given by

$$P = N_s \times R \times \hbar\omega_0 = \frac{N}{2}\frac{(\hbar\omega_0)}{kT} \times (\hbar\omega_0) \times R. \tag{7.27}$$

Before proceeding further, we introduce some numerical values: for protons $\gamma = 2.673 \times 10^8$ rad/s-T, so that for $B_0 = 1$ T and $T = 300$ K we obtain

$$\frac{N_s}{N} = \frac{\omega_0\hbar}{2kT} = \frac{(2.67 \times 10^8) \times (6.6 \times 10^{-16})\ \text{eV}}{2(1/40)\ \text{eV}} \approx 4 \times 10^{-6},$$

which justifies the approximation used to obtain Eq. (7.26). If we further consider a sample of 1 cm^3 of water, the number of protons contained in it is

$$N = N_0 \times (2/18) = 6 \times 10^{23} \times (2/18) = (2/3) \times 10^{23}.$$

If we use for $R = 1$/s (as can be seen from Eq. (7.16), this is a conservative value; R, however, can be as large as 10^3/s as discussed below), we obtain from Eq. (7.27)

$$P = (\hbar\omega_0) \times \left(N \times \frac{\hbar\omega_0}{2kT_s}\right) \times R \simeq 5 \times 10^{10}\ \text{eV/s} = 8 \times 10^{-9}\ \text{W}. \tag{7.28}$$

This is a very small amount of power, especially since the applied radiofrequency field may be on the order of milliwatts. Therefore, a sensitive null method greatly facilitates the observation of nuclear resonance absorption.

In writing Eq. (7.27), we assumed that the power absorbed is proportional to the number of excess nuclei which we now designate by n_s; however, as transitions are induced to the upper state, the number n_s will continuously decrease. The decrease will be exponential at the rate R

$$n_s = N_s e^{-Rt}.$$

Soon the populations of the two levels will be practically equalized, $N_{+1/2} \simeq N_{-1/2}$, and no more absorption will be observed.

However, while the radiofrequency field tends to equalize the populations, the "spin–lattice" interaction tends to restore the Boltzmann distribution at a rate characterized by $1/T_1$. We say that the nuclei are "relaxing" through their interaction with the lattice, and the characteristic time T_1 for this process is called the *spin–lattice relaxation time.* Therefore, in the presence of a radiofrequency field tuned to the resonance frequency, the number of excess nuclei at equilibrium n_s depends on T_1 and on R; if $R \ll 1/T_1$, then $n_s \to N_s$, while if $R \gg 1/T_1$, $n_s \to 0$. The value of n_s can be easily obtained[10]

$$n_s = \frac{N_s}{1 + 2RT_1}, \tag{7.29}$$

where N_s (Eq. (7.27)) is the equilibrium excess of population in the absence of the radiofrequency field.

By using Eq. (7.16) for R, we obtain

$$n_s = \frac{N_s}{1 + \frac{1}{2}\gamma^2 H_1^2 T_1 g(\nu)}. \tag{7.30}$$

From the above result we see that when too much radiofrequency power is used, the number of excess nuclei n_s decreases, and so does the resonance signal. We say that the sample has been saturated, and the ratio n_s/N_s is frequently referred to as the saturation factor Z:

$$\frac{n_s}{N_s} = \frac{1}{1 + \frac{1}{2}\gamma^2 H_1^2 T_1 g(\nu)} \equiv Z. \tag{7.31}$$

[10] Let $n = n_{+1/2} - n_{-1/2}$ be the instantaneous excess of nuclei in the presence of both radiofrequency and relaxation. The effect of the radiofrequency is to make $n \to 0$

$$\left(\frac{dn}{dt}\right)_{\text{RF}} = -2Rn.$$

(The factor of 2 arises because each transition up decreases $n_{+1/2}$ by 1, and also increases $n_{-1/2}$ by 1.) The effect of relaxation is to return $n \to N_s$

$$\frac{d(N_s - n)}{dt} = -(N_s - n)\,\frac{1}{T_1} = -\left(\frac{dn}{dt}\right)_{\text{relax}}.$$

Equilibrium is reached when the sum of the two rates is zero; that is,

$$-2Rn + \frac{N_s - n}{T_1} = 0$$

which yields Eq. (7.29).

The maximum useful value of the radiofrequency power therefore depends on the relaxation time T_1. For solids, T_1 is large (it takes a long time for the spins to reorient themselves in the equilibrium position), and therefore only weak radiofrequency fields may be applied. For example, for protons in ice $T_1 = 10^4$ s. In contrast, in liquids, especially in solutions containing paramagnetic ions, the relaxation time for protons may be as short as $T_1 = 10^{-4}$ s.

7.3.2. Line Width and T_2

Just as optical spectral lines can be broadened by external factors (see Section 6.4) the NMR signal is not perfectly sharp but has a certain width. Excluding inhomogeneities of the magnetic field B_0 over the size of the sample, the principal cause for the line width is the interaction between neighboring spins. In the classical analogy of Section 7.2.2 we say that the spin–spin interaction is destroying the phase coherence between the precessing spins and the rotating radiofrequency field. Another way of thinking of the spin–spin interaction is that one nuclear spin produces a local magnetic field B_{local} at the position of another spin, which then finds itself in a field

$$B_0' = B_0 + B_{\text{local}}$$

and consequently has a resonance frequency $\omega_0' = \gamma B_0'$ slightly different from ω_0. To estimate this effect, we calculate the magnetic field produced by a magnetic dipole one nuclear magneton strong, at a typical distance of 0.1 nm.

$$B_{\text{local}} \approx \left(\frac{\mu_0}{4\pi}\right) \frac{\mu_N}{r^3} = \left(\frac{\mu_0}{4\pi}\right) \times \frac{e\hbar}{2M_p} \times \frac{1}{r^3},$$

where μ_N is the nuclear magneton $e\hbar/2M_p$ and $\mu_0 = 4\pi \times 10^{-7}$ V-s/A-m is the permeability of free space. Numerically we find that

$$B_{\text{local}} \simeq 5 \times 10^{-4}\ \text{T},$$

which is a significant broadening of the line. In liquids and gases, however, the reorientation of the molecules is so fast that the average local field is very close to zero, and therefore very narrow lines can be obtained.

In Eqs. (7.15) and (7.16) we introduced the function $g(\nu)$ to describe the width of the NMR line. We now see that this width is mainly due to the spin–spin interaction. Since $g(\nu)$ has dimensions of inverse frequency, namely, of time, we *define* one-half of its maximum value by T_2

$$\frac{1}{2}\, g(\nu_0) = T_2, \tag{7.32}$$

where ν_0 is the resonance frequency in the absence of any broadening effects. T_2 is called the *transverse relaxation time*. In view of the normalization condition (Eq. (7.15)),

$$\int g(\nu)\, d\nu = 1,$$

(which also fixes the dimensions of $g(\nu)$), we see that a short T_2 implies broad lines, whereas when T_2 is long, the line is narrow.

Using the definition of Eq. (7.32), we can then write for the saturation factor Z (Eq. (7.31)) at resonance

$$Z_0 = Z(\nu_0) = \frac{1}{\left[1 + \gamma^2 H_1^2 T_1 T_2\right]}. \tag{7.33}$$

It is of interest to estimate T_2 for protons when $B_{\text{local}} = 5 \times 10^{-4}$ T as found previously. From the uncertainty principle $\Delta E \Delta t \sim \hbar$ and the line width $\Delta E = \gamma B_{\text{local}}$ so that

$$T_2 \sim \Delta t \sim \frac{1}{B_{\text{local}}}\, \frac{1}{(5.58\mu_N/\hbar)} \sim 7 \times 10^{-6}\ \text{s},$$

where we used $\gamma_p = 5.58$ and $\mu_N/\hbar = 2\pi \times 7.62$ MHz/T (see Eq. (7.2)).

Finally, as already mentioned, inhomogeneities in the magnetic field introduce spurious broadening effects that not only mask the fine structure of the line but also decrease the signal amplitude; hence the use of very homogeneous magnets and of the "spinning sample" technique.

7.3.3. The Bloch Magnetic Susceptibilities[11]

F. Bloch, who shared with E. M. Purcell the Nobel prize for the discovery of NMR, gave a macroscopic description of nuclear magnetic resonance,

[11] This section may be omitted without a loss of continuity and the reader can proceed directly to the discussion of the experimental technique and results in Section 7.4. However, the discussion should be quite helpful for understanding the meaning of the "dispersion" curve as well as the observed line shapes for both absorption and dispersion.

where the effect of the RF field is accounted for by the *polarization* of the nuclear spins. We know that when an electric (or magnetizing) field E (or H) is applied in a region containing matter, the material becomes polarized (or magnetized). We write

$$\mathbf{P} = \chi_e \mathbf{E} \qquad \mathbf{M} = \chi_\mu \mathbf{H}, \tag{7.34}$$

where χ_e and χ_μ are the electric and magnetic susceptibilities. The polarization is due primarily to the alignment of the permanent electric (magnetic) dipole moments of the atoms or molecules in the direction of the applied field. Materials that have such dipole moments and exhibit large polarization should be called *paraelectric* (or for large magnetization, they are indeed called *paramagnetic*).

The refractive index of light is related to the electric and magnetic susceptibilities, since

$$\epsilon = (1 + \chi_e)\epsilon_0 \qquad \mu = (1 + \chi_\mu)\mu_0$$

and

$$n = \frac{c}{c'} = \frac{1/(\sqrt{\epsilon_0 \mu_0})}{1/(\sqrt{\epsilon\mu})} = \sqrt{(1 + \chi_e)(1 + \chi_\mu)}.$$

The refractive index and therefore also the susceptibilities are a function of the frequency, as is evident from the familiar phenomenon of the dispersion of light. Thus the susceptibility at optical frequencies differs from the static one and is a function of the frequency.[12] Frequently the transmission of light through matter is accompanied by absorption that may be strongest at a particular resonant frequency. We may account for the absorption by attributing an imaginary part to the susceptibility.

The same formalism can be used as well for the description of nuclear magnetic resonance phenomena. The static susceptibility arising from the *nuclear* moments in an otherwise diamagnetic material differs from zero, but is very small and difficult to measure. For the radiofrequency susceptibility, we write

$$\chi(\omega) = \chi'(\omega) - i\chi''(\omega),$$

[12] For optical frequencies and for almost all materials, χ_μ is 0 and the variation in n arises entirely from χ_e.

where both $\chi'(\omega)$ and $\chi''(\omega)$ exhibit a resonant behavior when ω reaches $\omega_0 = \gamma B_0$. The real part $\chi'(\omega)$ is given by

$$\chi'(\omega) = \frac{1}{2}\,\chi_0\omega_0 T_2 \left[\frac{(\omega_0-\omega)T_2}{1+(\omega_0-\omega)^2T_2^2+\gamma^2 H_1^2 T_1 T_2}\right], \tag{7.35}$$

while the imaginary part $\chi''(\omega)$ is given by

$$\chi''(\omega) = \frac{1}{2}\,\chi_0\omega_0 T_2 \left[\frac{1}{1+(\omega_0-\omega)^2T_2^2+\gamma^2 H_1^2 T_1 T_2}\right]. \tag{7.36}$$

Here χ_0 is the static magnetic susceptibility defined as in Eq. (7.34)

$$M_0 = \chi_0 H_0,$$

and T_1 and T_2 are the familiar relaxation times introduced before; the term $\gamma^2 H_1^2 T_1 T_2$ appearing in the denominator is a measure of the saturation as defined in Eq. (7.31).

Equations (7.35) and (7.36) are shown in Fig. 7.5 under the assumption that $\gamma^2 H_1^2 T_1 T_2 \ll 1$; they have the typical behavior of a dispersion and a power resonance curve. We also note that Eq. (7.35) is proportional to the derivative, with respect to ω, of Eq. (7.36). By adjusting the detection equipment, we may observe experimentally either of those curves, or a combination of both, as a function of $\omega_0 - \omega$. Experimentally we can vary

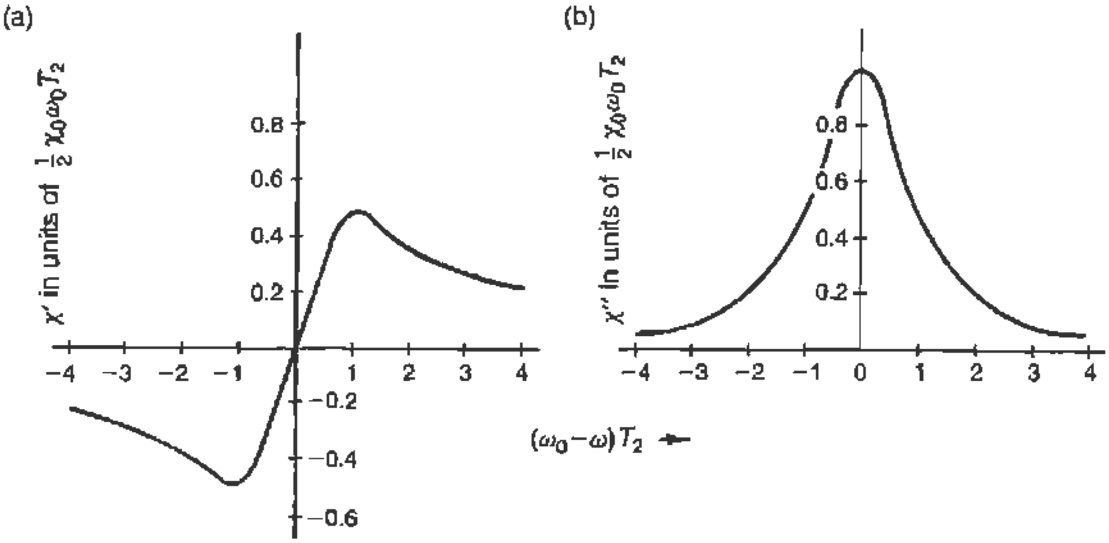


FIGURE 7.5 The radiofrequency magnetic susceptibilities near resonance. (a) The real part of the susceptibility exhibits a typical dispersion shape (Eq. (7.35)). (b) The imaginary part of the susceptibility exhibits a typical absorption shape (Eq. (7.36)).

$\omega_0 - \omega$ either by sweeping the magnetic field (changes $\omega_0 = \gamma B_0$) at fixed RF frequency, or by sweeping the RF frequency ω, while keeping the field B_0 fixed.

7.3.4. Free Induction Decay and Pulsed NMR[13]

It is convenient to consider again the classical interpretation of NMR discussed in Section 7.2.2. Refer to Fig. 7.4b and assume that the RF field is applied along the x' axis in the rotating frame, for a short time t, such that $\omega_1 t = \gamma H_1 t = \pi/2$. Then the net magnetization vector **M** will be rotated into the x'–y' plane; in fact it will be along the y' axis. In the laboratory frame this situation corresponds to a magnetization vector rotating in the x–y plane with angular frequency $\omega_0 = \gamma B_0$ around the z axis. A coil is fixed in the laboratory frame with its axis in the x–y plane. Then the rotating magnetization will *induce* an RF signal in the coil at frequency ω_0. Recall that now $\mathrm{M}(t) = \mathrm{M}_x \cos \omega t + M_y \sin \omega t$. This sequence is shown in Figs. 7.6a and 7.6b.

How long will the signal persist after time t? First of all because the spins are in contact with the lattice there will be a tendency for **M** to return into alignment with the z axis (recall that there is no RF field after time t). This relaxation process is characterized by the time T_1, the spin–lattice relaxation time introduced in Eq. (7.29). Usually, however, T_1 is fairly long and the individual spins that contribute to **M** become *dephased* either because of field inhomogeneities or because of the spin–spin interaction. When the spins are completely dephased (i.e., when they are pointing uniformly in all directions in the x–y plane) $d\mathbf{M}/dt$ through the coil vanishes and so does the induced signal. This effect occurs on a time scale T_2, which is usually shorter than T_1. Thus we observe a decaying exponential as shown in Fig. 7.6c. In general the decay constant is designated by T_2^* and contains the effects of the spin–spin interaction, magnetic field inhomogeneity, and spin–lattice relaxation

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_1} + \gamma \Delta B_0. \tag{7.37}$$

[13] This section, too, can be omitted on a first reading without loss of continuity. However, it provides insight on the interpretation of transient effects and of the modern NMR techniques that are based on pulsed excitation rather than continuous wave (CW) measurements.

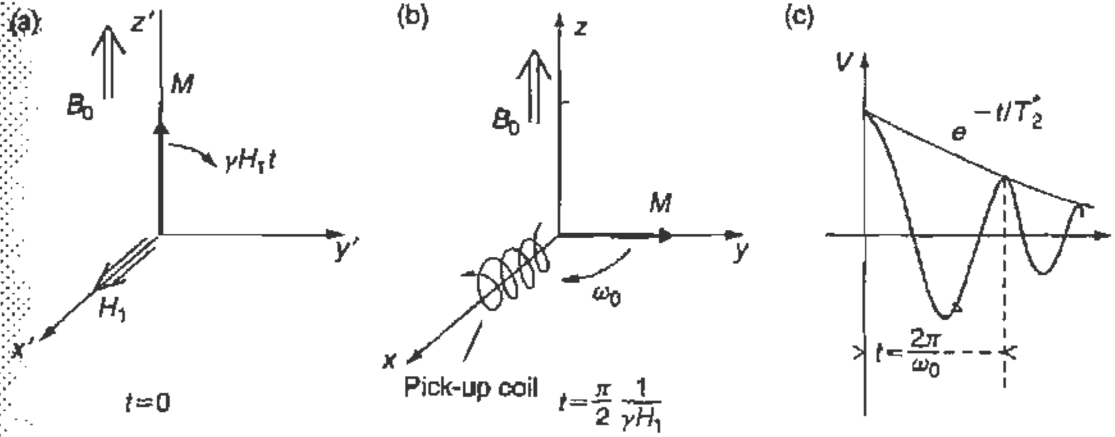


FIGURE 7.6 Free induction decay following a $\pi/2$ RF pulse. (a) The magnetization vector M in the rotating frame of reference before the application of the RF ($t = 0$). (b) After the $\pi/2$ pulse, the M vector will precess in the stationary frame with angular velocity ω_0. (c) The induced signal in a stationary coil in the x–y plane will have period $T = 2\pi/\omega_0$ and will decay exponentially with time constant T_2^*.

Therefore the free induction decay (FID) signal contains information on both the resonant frequency ω_0 (namely on γ) and on T_2 for the sample being investigated.

Note that if one performs a Fourier transform on the FID signal, which is acquired in the time domain, we obtain the spectrum of *all* the resonant frequencies of the sample. This is much more convenient and efficient than searching for each resonant line separately.

We now briefly return to the quantum-mechanical description of these phenomena. It was mentioned in Section 7.2.1 that the response of a two-level system to a resonant perturbation can be solved *exactly* in quantum mechanics.[14] If at $t = 0$ the spin is in state $m = -\frac{1}{2}$ the probability of finding it at time t in state $m = +\frac{1}{2}$ is

$$P_{+1/2} = \sin^2(\omega_1 t/2), \tag{7.38a}$$

with $\omega_1 = \gamma H_1$. The probability for the spin to remain in state $m = -\frac{1}{2}$ is

$$P_{-1/2} = \cos^2(\omega_1 t/2), \tag{7.38b}$$

as it must be since for a two-level system it must hold that $P_{-1/2} + P_{+1/2} = 1$.

[14] See footnote 7 of this chapter.

First we reconcile the result of Eq. (7.38a) with our perturbative calculation for the transition rate obtained in Eq. (7.16). The transition rate is, of course, the time derivative of the probability and we have

$$\frac{dP_{-1/2\to+1/2}}{dt} = \frac{\omega_1}{2}\sin\omega_1 t = \frac{\gamma H_1}{2}\sin\omega_1 t.$$

The perturbative calculation is valid when $\omega_1 t \ll 1$, and therefore we can expand $\sin\omega_1 t$ to first order to find that

$$\frac{dP_{-1/2\to+1/2}}{dt} = \frac{\gamma^2 H_1^2}{2}\, t. \tag{7.39}$$

This result seems different from Eq. (7.16) but we realize that Eqs. (7.38) are valid as long as the initial and final states are not otherwise disturbed over the time interval t. The maximum such time is given by $T_2 = \frac{1}{2}\, g(\nu_0)$ (see Eq. (7.32)). Thus

$$\frac{dP_{-1/2\to+1/2}}{dt} = \frac{\gamma^2 H_1^2}{4}\, g(\nu_0)$$

as expected from Eq. (7.16); we recall that in deriving Eqs. (7.38) it was assumed that $\omega = \omega_0$. Note also that

$$\left|\frac{dP_{+1/2\to-1/2}}{dt}\right| = \left|\frac{dP_{-1/2\to+1/2}}{dt}\right|$$

as repeatedly emphasized for the transition rate.

It is clear that applying a π-pulse ($\omega_1 t = \pi$) to a spin $\frac{1}{2}$ in the state $m = -\frac{1}{2}$ will make $P_{+1/2} = 1$ and $P_{-1/2} = 0$; namely the spin will flip states, as we also concluded from the classical analogy in Section 7.2.2. However, what is the result of a $\pi/2$-pulse ($\omega_1 t = \pi/2$)? Then we find that

$$P_{+1/2} = \frac{1}{2} = P_{-1/2}.$$

Namely the spin is in a coherent superposition of the $m = +\frac{1}{2}$ and $m = -\frac{1}{2}$ states. It is described by a wave function

$$\psi = \frac{1}{\sqrt{2}}\left[\left|m_z = +\frac{1}{2}\right\rangle + \left|m_z = -\frac{1}{2}\right\rangle\right]. \tag{7.40}$$

This represents a spin oriented in the x–y plane, and in the presence of a magnetic field B_0 along the z axis it will precess[15] in this plane with angular frequency $\omega_0 = \gamma B_0$. The quantum and classical descriptions lead to precisely the same conclusions.

We conclude with the remark that the same formalism is used for atoms when two states of energy E_1 and E_2 are connected by an electric-dipole moment d. If such an atom is subject to an oscillatory electric field $\mathcal{E}_1$ at the resonant frequency between the states, $\omega_0 = (E_f - E_i)/\hbar$, transitions will occur. Of course ω_0 is now an optical frequency rather than RF frequency. If the atom is initially in the state $|i\rangle$ and the optical field is switched on at $t = 0$, the probability[16] for finding the atom in the state $|f\rangle$ at time t is

$$P_f(t) = \sin^2\left(\frac{\mathcal{E}_1 d}{2\hbar}\, t\right), \tag{7.41a}$$

and for finding it in the state $|i\rangle$

$$P_i(t) = \cos^2\left(\frac{\mathcal{E}_1 d}{2\hbar}\, t\right). \tag{7.41b}$$

These are of course the exact analogues to Eq. (7.38).

The precession frequency for the atomic case $\Omega_1 = \mathcal{E}_1 d/2\hbar$ is called the Rabi frequency. With the availability of lasers one can achieve strong enough electric fields to generate $\pi/2$, π, etc., optical pulses. In this way atoms can be placed in specific quantum states. Such manipulation of single atoms has recently found applications in quantum cryptography, and it could eventually lead to quantum computing.

7.4. EXPERIMENTAL OBSERVATION OF THE NUCLEAR MAGNETIC RESONANCE OF PROTONS

7.4.1. General Considerations

To observe nuclear magnetic resonance we need a sample, a magnet, a source of electromagnetic radiation of the appropriate frequency, and a detection system.

[15] See Das and Melissinos (1986) cited in Footnote 7 of this chapter.

[16] Here we gloss over the fact that d is really the matrix element of the electric-dipole operator between the initial and final states.

The magnetic field should be fairly homogeneous, and therefore it is advisable to choose a good magnet with polefaces at least 4 to 6 in. in diameter. As discussed in Section 7.3.2 inhomogeneities in the magnetic field broaden the line and reduce the peak amplitude; to obtain reasonable results, the inhomogeneities over the volume of the sample should be less than 1/1000. The choice of the field strength is arbitrary, provided the resulting frequency lies in a convenient radiofrequency band. However, since the signal-to-noise ratio increases (improves) as $\nu_0^{3/2}$, high fields are preferable; commonly, magnetic fields of 0.5 to 1 T are used; and for protons this corresponds to frequencies of 20 to 40 MHz.

The sample can be any material containing an ample supply of protons; paraffin, water, mineral oil, or any organic substance containing hydrogens will, in general, give a proton nuclear magnetic resonance signal. Some care must be exercised to avoid materials with long spin–lattice relaxation times T_1, since they will saturate at very low levels of radiofrequency power and therefore give weak signals (see Eq. (7.33)). Similarly it is profitable to have a narrow line; hence materials with long spin–spin relaxation time T_2 are chosen. Liquids will meet this condition, and in most instances the width of the line will be determined by the magnet inhomogeneity ($T_2 = 3 \times 10^{-4}$ s will give for protons a line width of 10^{-5} T). Plain tap water makes a good sample, or tap water doped with 1 wt% manganese nitrate $Mn(NO_3)_2$ or copper sulfate.

The size of the sample is limited by the area over which the magnet is homogeneous, but also by practical considerations of the coil used to couple the radiofrequency to the sample. In usual practice a 1-cm^3 sample is adequate; it is contained in a small tubular glass container, around which is wrapped a radiofrequency coil as shown in Fig. 7.7a. The whole assembly is then inserted into the magnet gap and should be secured firmly, since vibration is picked up by the coil and appears as noise in the detector.

In deriving the probability for a transition between the m sublevels, and in all our previous discussion, we have assumed the existence of a *rotating* field at the angular frequency ω close to ω_0. In practice, a magnetic field oscillating linearly as $A \sin \omega t$ is established in the interior of the radiofrequency coil (Fig. 7.7a). Linear harmonic motion, however, is equivalent to two rotations in opposite direction of amplitude $A/2$ as shown in Fig. 7.7b since

$$A \cos \omega t \mathbf{n}_x = \frac{A}{2}(\cos \omega t \mathbf{n}_x + \sin \omega t \mathbf{n}_y) + \frac{A}{2}(\cos(-\omega t)\mathbf{n}_x + \sin(-\omega t)\mathbf{n}_y), \tag{7.42}$$

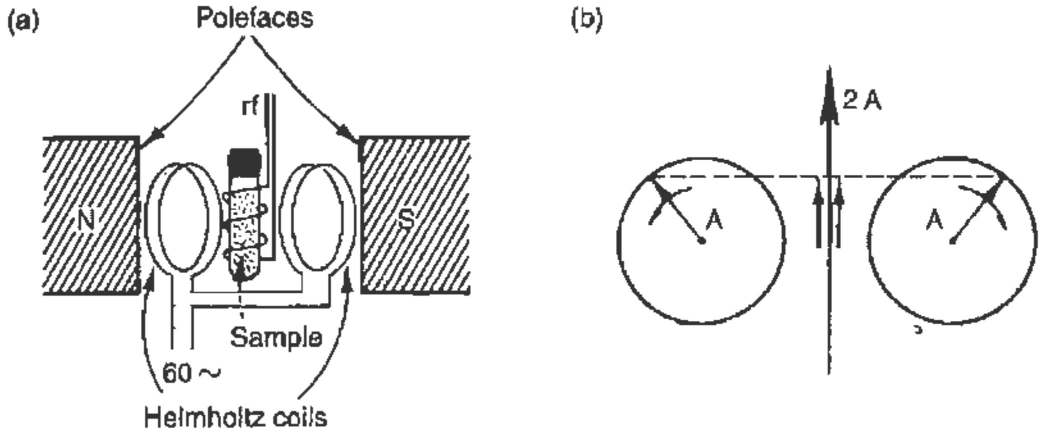


FIGURE 7.7 (a) Schematic arrangement of a nuclear magnetic resonance apparatus. The sample is placed in a homogeneous magnetic field and radiofrequency is coupled to it by means of the coil. The Helmholtz coils are used to modulate the constant magnetic field. (b) A linearly oscillating field of frequency ω is equivalent to two fields rotating in opposite directions with the same frequency ω.

where $\mathbf{n}_x$ and $\mathbf{n}_y$ are unit vectors in the x and y directions. The component rotating in the same direction as the precessing spins will be in resonance and may cause transitions; the other component is completely out of phase and has no effect on the sample.

When the radiofrequency reaches the resonance value ω_0, energy is absorbed from the field in the coil and this fact is sensed by the detector. Because of the low signal levels involved and the difficulty of maintaining a very stable level of radiofrequency power it is advantageous to *traverse the whole resonance curve* in a relatively short time. This can be achieved either by "sweeping" the frequency of the radiofrequency oscillator while maintaining the magnetic field constant, or by "sweeping" the magnetic field while the frequency remains fixed. In early NMR experiments as well as in this laboratory the choice is to sweep the field with a pair of Helmholtz coils,[17] as indicated in Fig. 7.7a, because it is easy and does not require fancy frequency generators. The sweep coils are fed with a slowly varying current,[18] which results in a modulation of the magnetic field B. If this sweep covers the value of B_0, which is in resonance with the fixed frequency of the oscillator, a resonance signal modulated at the frequency of

[17] A pair of coils of diameter d, spaced a distance $d/2$ apart and traversed by current in the same direction, produce a very homogeneous field at the geometrical center of the configuration.

[18] In the absence of a sweep generator and audio amplifier the 60-Hz line voltage can be used through a variac and an isolation transformer.

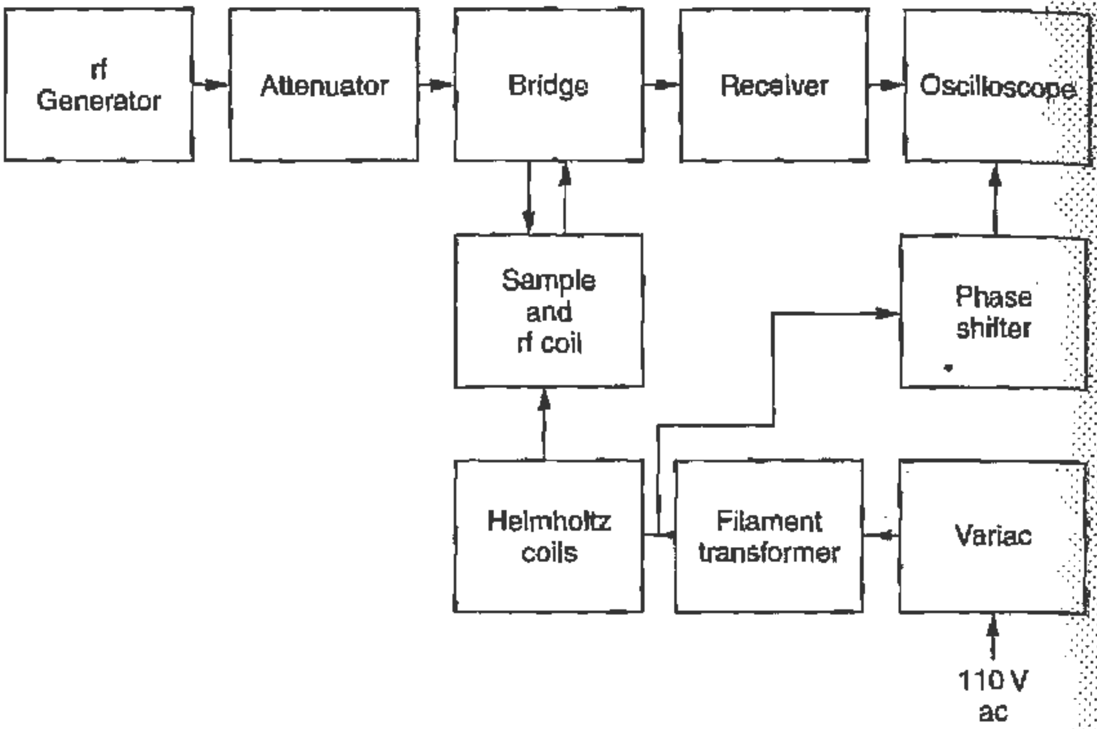


FIGURE 7.8 Block diagram of the nuclear resonance measuring apparatus.

the sweep will appear at the detector. A modulated signal has the advantage of easier amplification and improvement in the signal-to-noise ratio by using a narrow bandwidth detector.

The radiofrequency oscillator and detection circuit can be of several designs. Today, commercial frequency generators are used to provide the RF drive and low-noise amplifiers for the detector. A *single* coil is used as both a transmitter and receiver. A block diagram of a CW NMR apparatus as used in this laboratory is shown in Fig. 7.8. The signal was detected by a bridge circuit; this arrangement has great sensitivity but can be used without retuning only over a fairly narrow frequency range.

Commercial magnetometers often use a "marginal oscillator" circuit where the oscillator and detector are combined in one unit. In this design the RF power is kept low so as to allow the direct observation of the absorption, as well as to avoid saturation of the sample. To cover a wide frequency range the coil containing the sample is changed since it is part of the resonant circuit that sets the oscillator frequency. A unit suitable for laboratory demonstrations is available from Klinger Educational Products, as well as from other sources.

7.4.2. Detection of Nuclear Magnetic Resonance with a Bridge Circuit

The coil in which the sample is located is part of a resonant circuit with high Q. The Q value, or quality factor, of a device is defined as 2π times the ratio of the time-averaged energy stored to energy dissipated, in one cycle. For a coil of inductance L and resistance R,

$$Q = \frac{2\pi\omega L}{R}. \tag{7.43}$$

When resonance is reached, the real part of the magnetic susceptibility (Eq. (7.35)) changes, and thus the inductance of the coil also changes. Alternatively, an increase in the imaginary part of the susceptibility (Eq. (7.36)) corresponds to the absorption of power from the field and thus to increased dissipation and therefore increased resistivity of the coil. This small change in the Q value can be detected with a bridge circuit, as shown in Fig. 7.9.

The radiofrequency voltage is applied between points a and g (see Fig. 7.9a), and therefore radiofrequency current flows through the load L and the dummy branch D; if the bridge is balanced, no voltage should appear at the point d (since b and c were in phase and of the same amplitude, and the signal from c and d is shifted by $\lambda/2$). Any slight unbalance of the bridge produces a small voltage at d. The actual bridge circuit is shown in (b) of the figure. The $R'C'$ elements are effectively generating

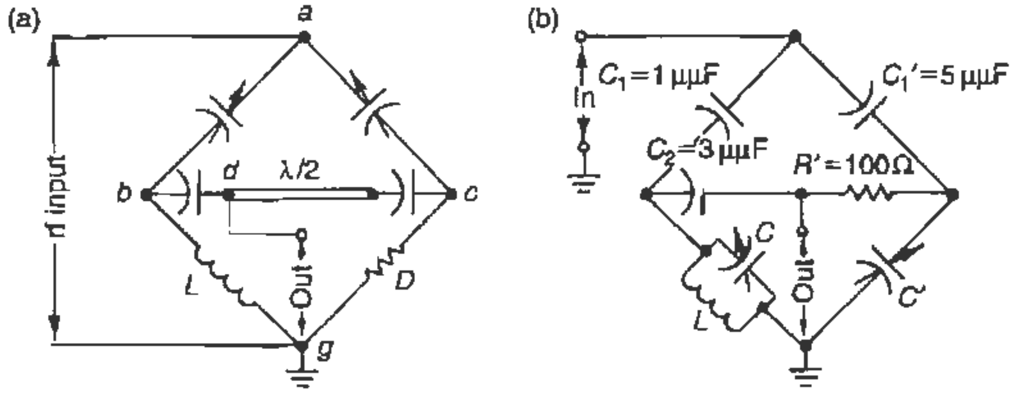


FIGURE 7.9 A radiofrequency bridge circuit that can be used for the detection of nuclear magnetic resonance. (a) Schematic arrangement; note that L is the radiofrequency coil. The $\lambda/2$ line ascertains cancellation at the output of the signals from b and c. (b) A practical radiofrequency bridge circuit. For resonance conditions see Eqs. (7.44) of the text.

the $\lambda/2$ phase shift and L is the sample coil. The conditions for balance are

$$\begin{aligned} &\text{Resistive balance:} \quad && \omega^2 C_1 C_2 \left(1 + C'/C_1{}'\right) R' R_\mathrm{p} = 1 \\ &\text{Reactive balance:} \quad && C + C_1 + C_2 \left(1 + C_1/C_1{}'\right) = 1/L\omega^2, \qquad (7.44) \end{aligned}$$

where R_p is the parallel resistance of the coil. The bridge is balanced either in the resistive mode, when the change in the Q of the coil will appear as an absorption curve as in Fig. 7.5b, or the bridge may be balanced in the reactive mode, when the signal appears as a dispersion curve as in Fig. 7.5a.

The experimental results obtained with this arrangement by a student are shown in Fig. 7.10. The sample was 1 cm^3 of water doped with manganese nitrate [$\mathrm{Mn(NO_3)_2}$]. In Fig. 7.10a the bridge was balanced in the reactive mode, whereas in Fig. 7.10b it was balanced resistively. The sweep, derived from the 60-Hz line voltage, corresponds to approximately 10^{-4} T/division at the center of the oscilloscope trace.

The exact frequency at resonance can be measured quite precisely with a (crystal-controlled) "wave meter" to better than 1 part in 10^6. The magnetic field is measured either with a Hall probe magnetometer or with a rotating coil flux-meter.

From the experimental curves of Fig. 7.10 it is found that the frequency at resonance is

$$\nu_0 = 28{,}141.48 \pm 0.63 \text{ kHz}.$$

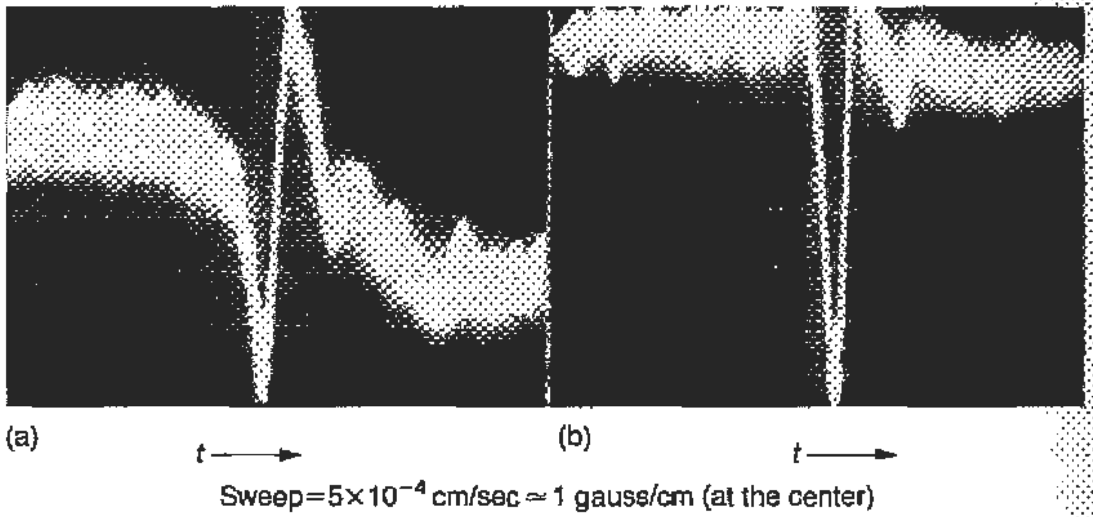


FIGURE 7.10 Results obtained from the nuclear magnetic resonance of protons using a bridge circuit: (a) Dispersion curve and (b) absorption curve. The oscilloscope sweep was linear at 0.5 ms/cm, which corresponds to approximately 10^{-4} T/cm at the center of the sweep.

Using a rotating coil flux-meter at the field position previously occupied by the sample, the magnetic field at resonance is found to be

$$B_0 = 0.6642 \pm 0.0020\ \text{T},$$

and hence

$$\gamma = \frac{2\pi\nu_0}{B_0} = (26.618 \pm 0.08) \times 10^7\ \text{rad/s-T} \quad , \qquad (7.45)$$

in good agreement with the accepted value

$$\gamma = 26.73 \times 10^7\ \text{rad/s-T}.$$

Clearly, it is much easier to measure ratios of nuclear moments to high accuracy than to establish their absolute value to the same accuracy.

To obtain the g factor of the proton—that is, the connection between magnetic moment and the nuclear magneton—we recall that

$$\mu = gI\mu_N.$$

Thus from Eq. (7.4)

$$g = \frac{\gamma\hbar}{\mu_N} = \frac{\gamma}{2\pi}\frac{1}{\mu_N/h} = 5.56 \pm 0.02,$$

where we used the derived value of γ (Eq. (7.45)) and μ_N/h from Eq. (7.2). We have measured the proton magnetic moment of the proton to an accuracy of 0.4%.

7.4.3. Measurement of T_2^*

In this laboratory no pulsed NMR experiments were carried out. However, under certain conditions one can observe the free induction and its decay with a CW apparatus. This happens if the field is swept rapidly enough through the resonance, in which case wiggles such as those shown in Fig. 7.11 appear.

The interpretation follows the discussion of Section 7.3.4. Far from resonance the field seen in the rotating frame is B_0, i.e., along the z axis. As resonance is approached the B_0 field is canceled in the rotating frame and only H_1 is present. This results in rotating the **M** vector into the x'–y' plane. After the resonance is traversed the effect of H_1 is again minimal, but the magnetization remains in the x'–y' plane, and it induces a signal at

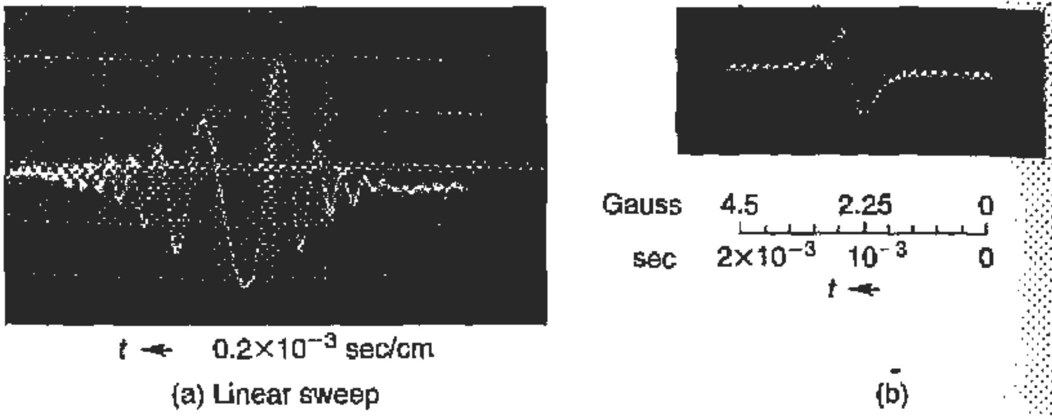


FIGURE 7.11 Nuclear magnetic resonance signals of protons obtained with a marginal oscillator circuit. (a) The sample is water-saturated with LiF. (b) The sample is water-doped with manganese nitrate. A linear sweep of the same speed is used in both cases.

a frequency $\omega(t) = \gamma B_0(t)$, which differs from ω_0. The two frequencies $\omega(t)$ and ω_0 beat against each other, and this gives rise to the wiggles. One can clearly see that the frequency difference increases (the beating period shortens) as the field is further away from resonance. The effect that is relevant for our measurement is the exponential decay of the envelope of the beat oscillations.

We still must explain the wiggles that appear in Fig. 7.11a before the resonance is crossed. These are present because the spins have not dephased by the time the sweep is restarted and continue to rotate in the x–y plane. Indeed they are absent from the trace of Fig. 7.11b where the water sample was doped with manganese nitrate as compared to water-doped with LiF in the sample used for Fig. 7.11a. The shorter T_2 in part (b) of the figure leads to more rapid dephasing.

If a linear sweep is assumed, the beat signal has the form

$$e^{-t/T_2^*} \cos\left[\frac{1}{2}\,\gamma\,\frac{dH}{dt}\,t^2\right], \tag{7.46}$$

where $t = 0$ when the resonance is traversed. Note also that the beat frequency increases with time since

$$\omega_b = \frac{1}{2}\,\gamma\,\frac{dH}{dt}\,t.$$

From a measurement of the wiggle envelope, information about T_2^* can be obtained. This is shown in Fig. 7.12 where the data are well fitted by an

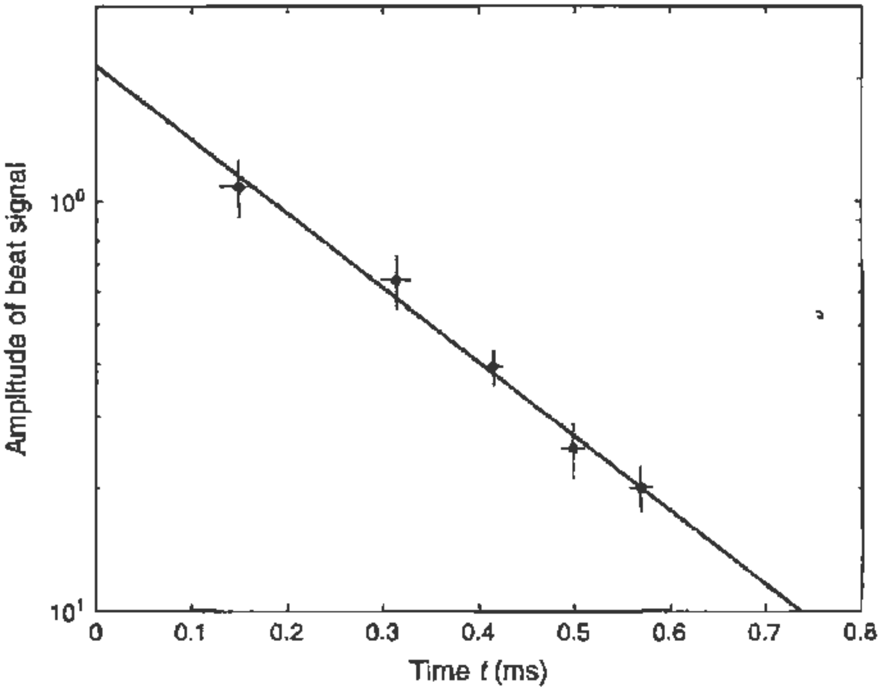


FIGURE 7.12 Semilog plot of the amplitude of the "wiggles" of the resonance signal shown in Fig. 7.11a plotted against time. It yields an exponential decay of the amplitude with a time constant $T_2^* = 2.4 \times 10^{-4}$ s.

exponential yielding

$$T_2^* = 2.4 \times 10^{-4}\ \text{s}.$$

When we convert the measured value of T_2^* into a magnetic field (see Eq. (7.37)), we find that

$$2\Delta B_0 = \frac{2}{T_2^* \gamma} = 3.2 \times 10^{-5}\ \text{T},$$

namely, that an inhomogeneity of the magnetic field, over the size of the sample, of 0.32 G is sufficient to cause the wiggles observed in Fig. 7.11a. We also conclude that T_2 for this sample is longer than 2.4×10^{-4}s.

7.4.4. The Effect of T_1

In Fig. 7.13 we show a very simple marginal oscillator circuit[19] that is adequate for demonstrating NMR signals. The first transistor supplies constant

[19] J. R. Singer and S. D. Johnson, *Rev. Sci. Instrum.*, **30**, 92 (1959).

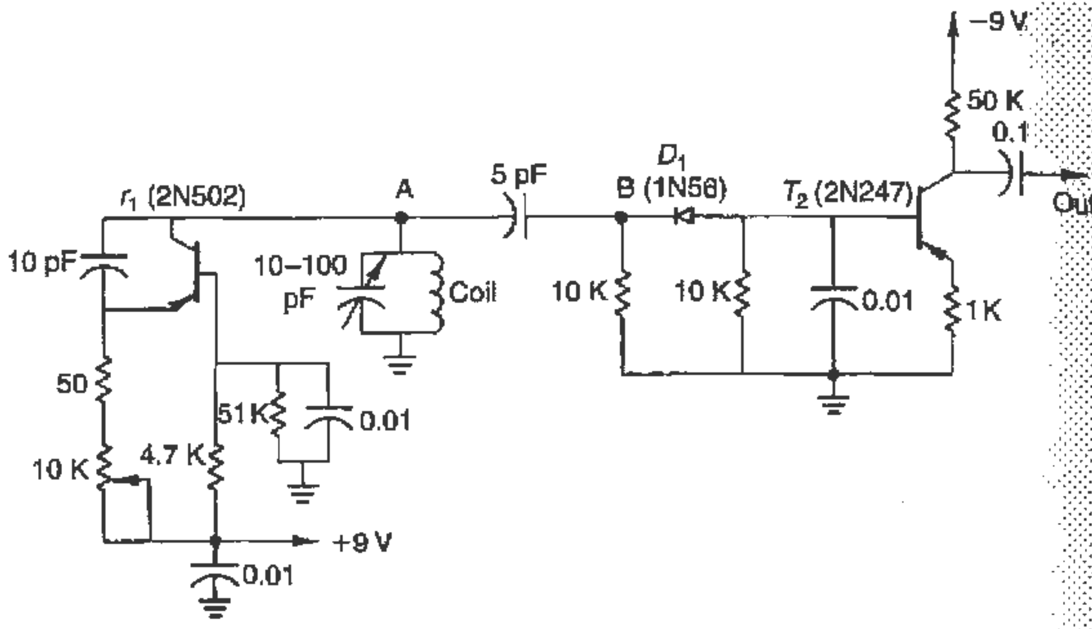


FIGURE 7.13 A simple transistorized nuclear magnetic resonance circuit.

current to the coil attached to point A. Any change in the Q of the coil appears as a change in voltage at that point, which is then amplified by the output transistor. The circuit will oscillate in the range of 2–80 MHz, depending on the resonance set by the coil LC circuit.

Data obtained by a student using a 1-cm^3 sample of water doped with manganese nitrate are shown as a function of RF amplitude in Fig. 7.14. The amplitude is controlled by the 10-kΩ potentiometer in the oscillator loop of the circuit of Fig. 7.13. The data were obtained in a field $B_0 = 0.8$ T ($\nu_0 = 33.83$ MHz). The RF level as measured across the coil is indicated for each of the traces shown in the figure. Note that the NMR signal increases with increasing RF power until the RF amplitude reaches approximately 0.5 V. Beyond this point the signal decreases because the sample is saturated. From a knowledge of the Q of the coil one can convert the RF amplitude to the corresponding value of the rotating field H_1 and thus use the data to find the spin–lattice relaxation time T_1.

Note also that once the sample is saturated there is sufficient magnetization left in the x–y plane to begin showing a beat signal (wiggles) after passage through resonance (see Fig. 7.11b). For convenience the time scale on the oscilloscope trace in Fig. 7.14 was set to cover a full cycle of the 60-Hz sinusoidal sweep.

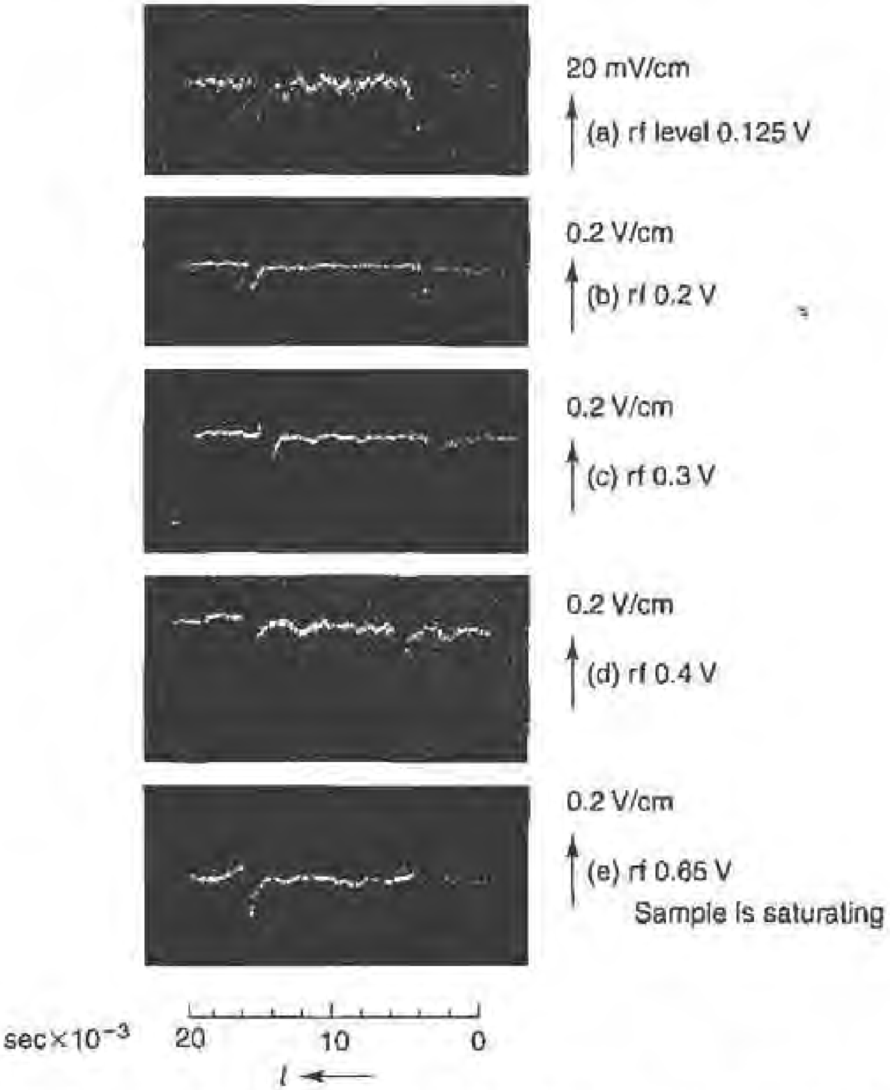


FIGURE 7.14 Nuclear magnetic resonance signals from protons obtained with the circuit shown in Fig. 7.13 as a function of the amplitude of the radiofrequency. Note that initially the output signal increases with increasing radiofrequency amplitude but at a level of approximately 0.5 V the sample is saturated and the signal begins to decrease. The signal of 0.5 V is shown in Fig. 7.11b.

7.5. ELECTRON SPIN RESONANCE

7.5.1. General Considerations

So far we have discussed transitions between the energy levels of a proton or a nucleus in the presence of an external magnetic field. Transitions between the energy levels of a quasi-free electron in an external magnetic field can also be observed. We refer to this case as *electron spin resonance* (ESR) as already mentioned in Section 7.1. We expect the transition frequency for ESR to be approximately $\mu_B/\mu_N \sim 2000$ times higher than that for NMR. Namely for similar laboratory fields the resonance frequency is in the microwave region.

Atoms or molecules that have $J \neq 0$ in the ground state will exhibit ESR effects. Such atoms or molecules are paramagnetic (the atomic spins can be oriented by an external magnetic field), hence the use of the term electron paramagnetic resonance. In solids, however, it is much more difficult to find electronic states with $J \neq 0$: this is due to the fact that in the chemical binding of atoms into molecules, the valence electrons get paired off, so that each atom appears to have a completely closed shell. For example, in NaCl, the sodium has a $^2S_{1/2}$ electron ($n = 3,\ l = 0$) outside closed shells, and the chlorine has a $^2P_{3/2}$ electron hole ($n = 3,\ l = 1$) inside closed shells. However, in the NaCl molecule, the sodium appears as a Na^+ ion, and hence presents a closed shell configuration, whereas the chlorine appears as a Cl^- ion again with completely closed shells. Consequently, the NaCl molecule is completely diamagnetic.

Nevertheless, it is known from the work on static magnetic susceptibilities especially at low temperatures that certain salts show strong paramegnetism. Namely they contain ions with permanent magnetic moments on the order of μ_B. In particular, compounds containing ions of the "transition elements" of the periodic table are frequently found to be paramagnetic. An example is the compound copper sulfate ($\mathrm{Cu(SO)_4}$), in which the double valence results in a Cu^{2+} ion. For copper the $n = 1, 2$, and 3 shells are completely filled and one electron is in the 4s state, so that Cu^{2+} has a hole in the 3d shell; thus the ground state of the Cu^{2+} ion has $l = 2$, $s = \frac{1}{2}$, and, consequently, $J \neq 0$, so that it *does* possess a magnetic-dipole moment. In an external magnetic field, the ground state will be split into sublevels and resonance between them can be established and is indeed observed. The actual situation, however, is more complicated due to the influence of the electric field of the crystalline lattice.

Electronic magnetic moments can also be found in solids when the chemical bond is broken, as in organic free radicals. Especially, the organic salt DPPH, diphenyl-picryl-hydrazil ($\mathrm{(C_6H_5)_2N{-}NC_6H_2(NO_2)_3}$) shows a very strong and narrow resonance line, with a g factor very close to 2.00 (the free electron value) and it is therefore frequently used as a standard. The structure of the molecule is shown in Fig. 7.15, and the "free-electron" behavior comes from the single electron bond in one of the nitrogens. Electron spin resonance is also observed in other materials where unpaired electrons may

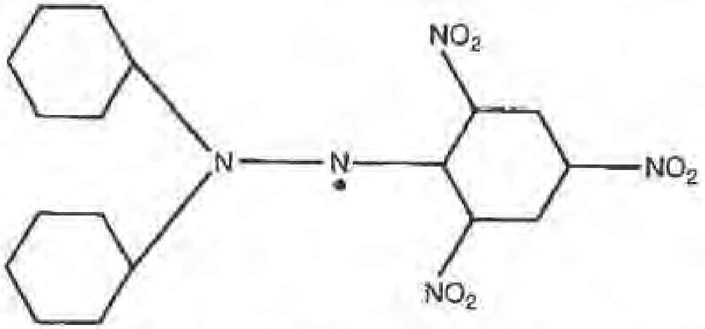


FIGURE 7.15 Chemical structure of DPPH (diphenyl-picryl-hydrazil), $(C_6H_5)_2N{-}NC_6H_2{-}(NO_2)_3$.

exist, such as crystals with lattice defects, in ferromagnetic materials, and in metals and semiconductors.

The much higher frequency of the ESR transitions is advantageous because the energy absorbed from the microwave field for every transition is much higher than that in the NMR case, thus leading to a much improved signal-to-noise ratio. Furthermore the separation between the energy levels is much larger, so that they remain resolved despite their large intrinsic width.

The resonance condition is detected, as in the case of nuclear magnetic resonance, by the absorption of energy, and for this reason solids and liquids are much easier to study than gases with their very low densities. Much of our previous discussion on transition probabilities and relaxation mechanisms is equally applicable to electron paramagnetic resonance. However, the population difference between the energy levels (see Eq. (7.26)) is much larger because of their greater energy spacing. A difficulty with ESR is that the width of the resonance line may be prohibitively large, since both the spin–lattice and spin–spin interactions are stronger than in the nuclear magnetic resonance case. In order to reduce the line width, the sample may be cooled to low temperatures (lengthens the spin–lattice relaxation time) and/or the paramagnetic ions are diluted in a diamagnetic salt (lengthens the spin–spin interaction time by effectively increasing the distance between the spins).

When measuring electron paramagnetic resonance lines in solids, a great variety of g factors are obtained. This is due to the differences in the coupling of the unpaired electron's spin with the orbital angular momentum; the strength of this coupling depends very much on the position (in energy) of the adjacent levels of the ion as they are modified by the crystalline field. Further, the electron paramagnetic resonance lines show hyperfine structure characteristic of the interaction of the nucleus with the ionic energy levels; this structure in turn can be used to positively identify small traces of an element contained in some unknown sample. Similarly, the organic free radicals show characteristic lines (g factors) that can be used to identify them and show hyperfine structure as well. In fact, a radical that has no structure (like DPPH) may exhibit such effects when the sample is prepared in a liquid solution.

7.5.2. The Electron Spin Resonance Spectrometer

In this laboratory ESR is observed using an X-band spectrometer. The term *X-band* refers to the frequency of the microwave source, which is in the 10-GHz region ($\lambda_f \simeq 3$ cm). Microwave components and "plumbing" are readily available. A schematic of the spectrometer is shown in Fig. 7.16 and at first appears quite elaborate.[20] However, the basic components shown in blocks separated by the dotted lines can be easily understood. The connections between components are made with an X-band waveguide, which is rectangular copper tubing with inner dimensions $a = 0.900$ in., $b = 0.400$ in.

The microwave source (block A) consists of a Varian X-13 klystron powered by a Hewlett-Packard 716B power supply. The klystron frequency can be controlled by the KLSP modulator, and this feature is used to "lock" the klystron frequency onto that of the external reference cavity shown in block B. Instructions for tuning the klystron to an appropriate mode and locking the frequency are provided with the instruments, and after a while one becomes familiar with the procedure. The sample cavity is shown in block C together with a phase shifter and tuner. There are also provisions for measuring the wavelength. Detection is accomplished in block D, by the equivalent of a microwave bridge, which uses a "magic tee" to compare the sample signal with the reference frequency. Block E indicates the magnet power supply and a set of Helmholtz sweep coils, which are driven by an audio amplifier at a ramp generated by a function generator. Finally in block F is shown the audio part of the detector where a lock-in detector can be used when the field is modulated. Otherwise the main field can be ramped under computer control, which also records the signal vs field.

[20] A very simple ESR demonstration apparatus operating in the RF range, and thus at very weak field, is available from Klinger Educational Products.

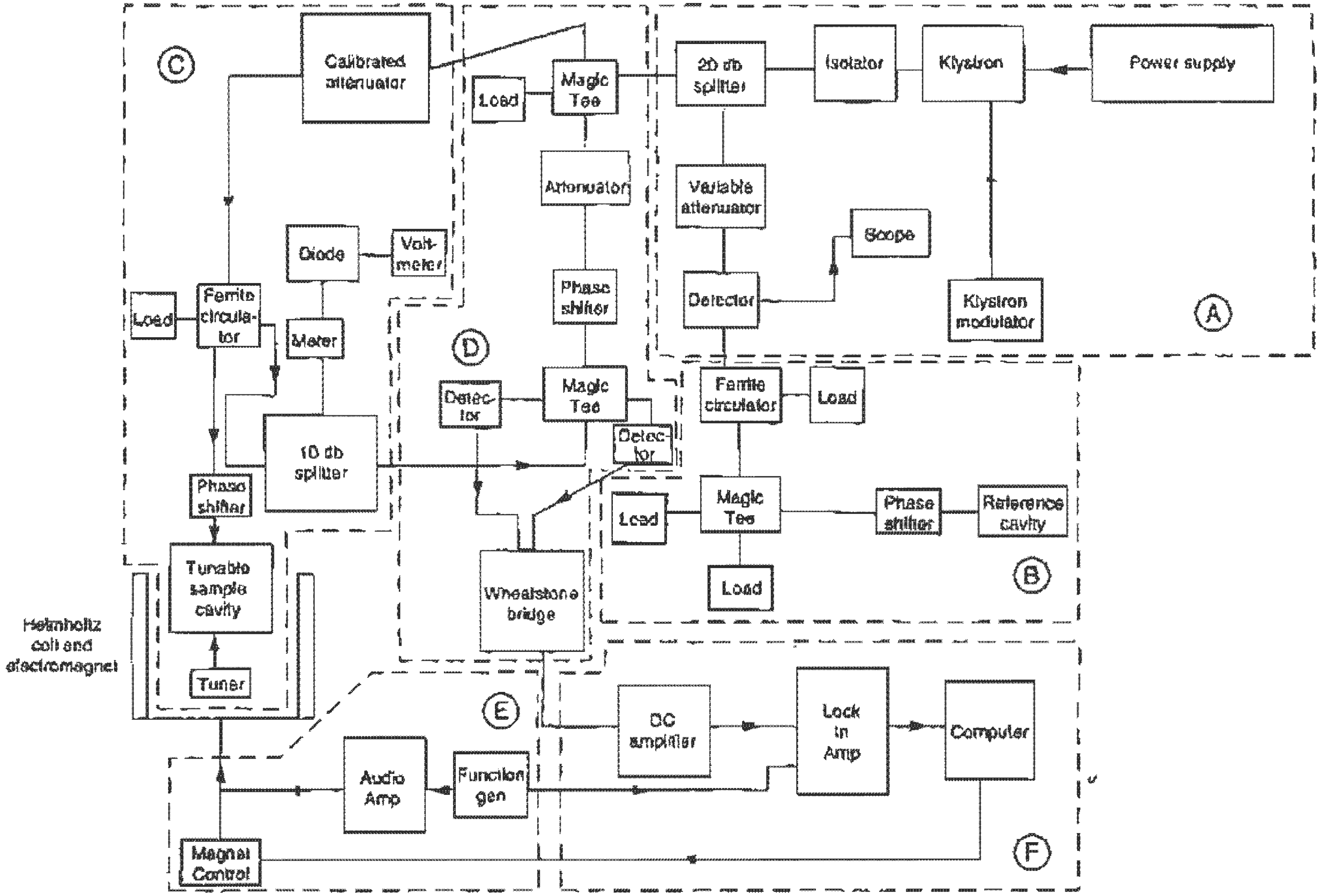


FIGURE 7.16 Schematic of the X-band ESR spectrometer.

We now elaborate on some of these components:

(a) *Propagation in the Waveguide.* Only certain modes will propagate without attenuation and the wavelength λ_g in the guide is given by

$$\frac{1}{\lambda_g^2} = \frac{1}{\lambda_f^2} - \frac{1}{\lambda_c^2} = \frac{1}{\lambda^2} - \frac{(m/a)^2 + (n/b)^2}{4}, \tag{7.47}$$

where λ_f is the free space wavelength and a and b are the inner dimensions of the guide; m and n are integers. Since

$$a = 2.29 \text{ cm}, \qquad b = 1.02 \text{ cm}, \qquad \text{and} \qquad \lambda_f \approx 3.2 \text{ cm}$$

we find that only the $m = 1$, $n = 0$ mode can propagate, and

$$\lambda_g = 4.5 \text{ cm}.$$

In this mode, the electric field is completely transverse to the axis of the guide; this is called the TE_{10} mode. The field lines for the traveling TE_{10} wave are shown in Fig. 7.17 where the density of field lines is proportional to the field strength.

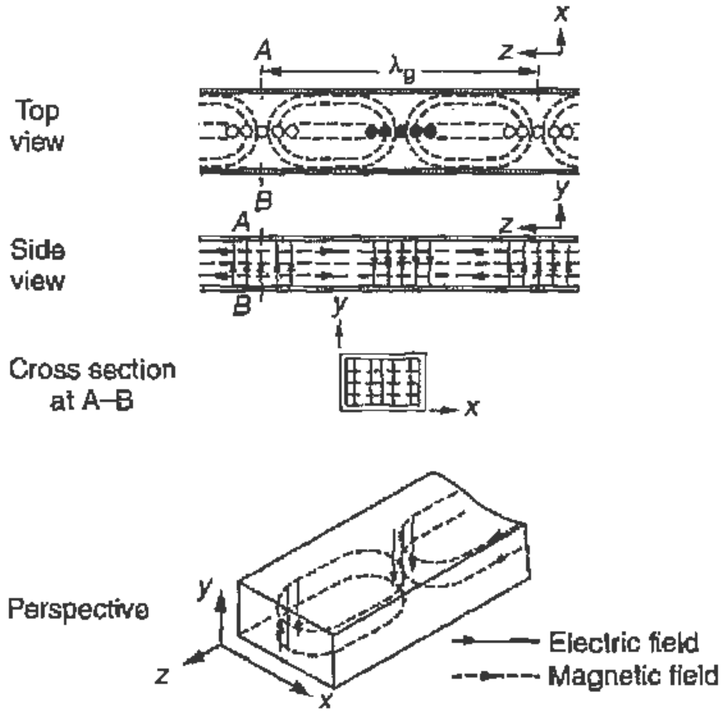


FIGURE 7.17 Configuration of electric and magnetic field lines for a traveling wave in a rectangular waveguide. λ_g is the wavelength in the guide.

(b) *The Microwave Cavity and Sample.* The cavity can be made from a part of the waveguide ending with a shorting stub, to set up a standing wave. The sample is placed so as to be located in the middle of the magnet polefaces, and then the (shorting) sliding stub is adjusted so that maximum B field exists at the sample. From the configuration of the *standing wave* pattern, maximum B field occurs at a distance x from the short, where

$$x = \left(\frac{1}{2} + \frac{p}{2}\right)\lambda_g$$

with p an integer. Since the microwave field must be normal to B_0 it is preferable to place the guide in the magnet with its wide side parallel to the polefaces.

(c) *The Magic Tee.* This is the heart of the bridge circuit, and is used to compare (interfere) microwave signals. It can be used in different configurations but in the spectrometer used here it is set up as shown in Fig. 7.18.

Let E_1 be the reference field and E_2 the signal field. The power at D_R and D_L are

$$P_R = |E_1 + E_2|^2 = |E_1|^2 + |E_2|^2 + 2\mathrm{Re}\left(E_1 E_2^*\right)$$
$$P_L = |E_1 - E_2|^2 = |E_1|^2 + |E_2|^2 - 2\mathrm{Re}\left(E_1 E_2^*\right).$$

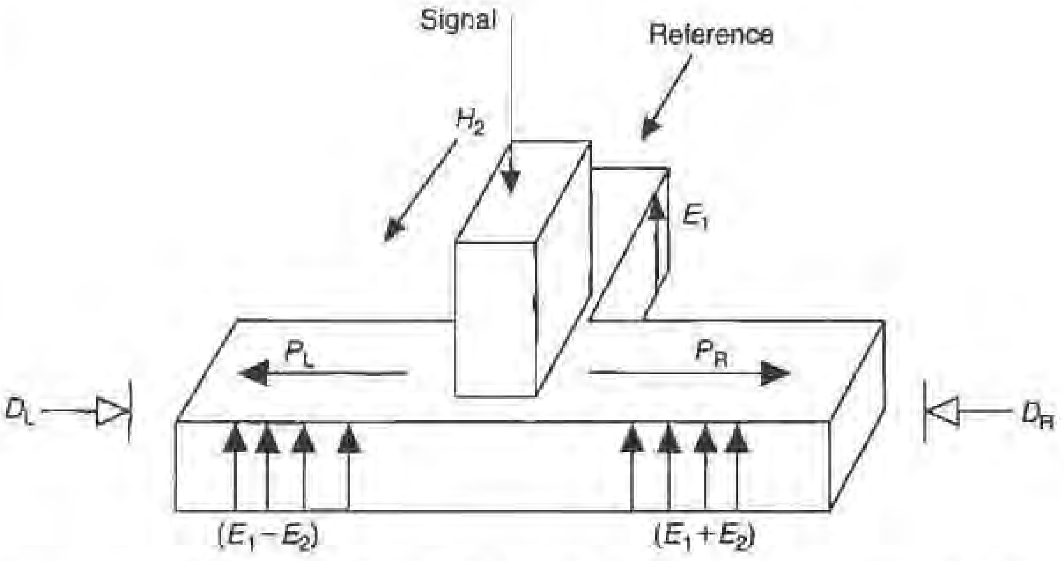


FIGURE 7.18 The magic tee used in the ESR spectrometer. A reference field and signal field are mixed within the tee to provide a sum or difference in the output arms of the tee.

If these two power levels can be subtracted, we have a signal, S, equal to

$$S = 4\mathrm{Re}\left(E_1 E_2^*\right).$$

Let

$$E_1 = E_\mathrm{R} e^{i\theta}, \qquad E_\mathrm{R} \text{ is real.}$$
$$E_2 = E_0(1+\chi), \qquad E_0 \text{ is real.}$$

Then recalling from Section 7.3.3 that $\chi(\omega) = \chi'(\omega) - i\chi''(\omega)$ we find that

$$S = (4E_\mathrm{R}E_0)[\cos\theta + \chi'\cos\theta + \chi''\sin\theta]. \tag{7.48}$$

By selecting the phase of the reference signal, we can select the desired curve. For $\theta = 0$ we obtain

$$S = (4E_\mathrm{R}E_0)[1 + \chi'(\omega)].$$

Since only $\chi'(\omega)$ is modulation dependent the signal follows the dispersive curve (Eq. (7.35) or Fig. 7.5a). For $\theta = \pi/2$ we obtain

$$S = 4(E_\mathrm{R}E_0)\chi''(\omega),$$

namely the absorptive part (Eq. (7.36) or Fig. 7.5b). If the reference phase is not set properly the signal is a mixture of the two curves as given by Eq. (7.48).

(d) *Detection.* One can use bolometers in the two arms of the magic tee. These are devices where the resistance changes as a function of incident power and are quite sensitive. It is, however, simpler to use in the magic tee microwave diodes similar to those used elsewhere in the spectrometer for diagnostic purposes.

(e) *Lock-In Detection.* If more sensitivity is required one modulates the B_0 field and sends the difference of the signals from the two arms of the magic tee to the lock-in detector. When the modulation width is much less than the line width the detected signal represents the *derivative* of the absorption (or dispersion) curve. This can be seen from the sketch of Fig. 7.19.

7.5.3. Experimental Results

Results obtained by students are shown below. The magnetic field was modulated at 1 kHz and the lock-in detector was used. The modulation

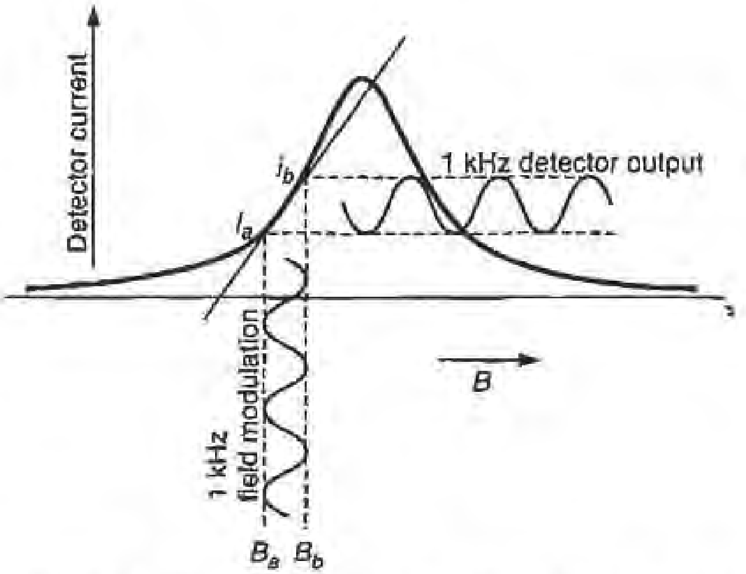


FIGURE 7.19 Effect of small-amplitude field modulation. The output is proportional to the derivative of the absorption curve and is maximum at the points of inflection.

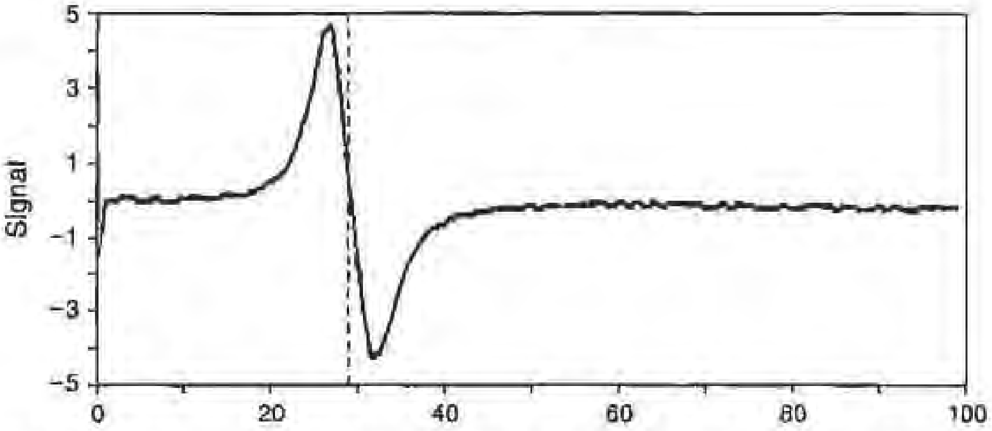


FIGURE 7.20 Resonance signal for DPPH as a function of the magnetic field. A small modulation was applied to allow lock-in detection, and therefore the signal gives the derivative of the absorption curve.

amplitude was kept low so that the derivative of the absorbtion line was observed. The field was swept through the resonance by slowly ramping the magnet current. The frequency was measured by using the wavemeter, and the magnetic field by using a Hall probe.

Figure 7.20 shows the results for DPPH. The field measured at the two ends of the sweep[21] was $B(0) = 0.3370$ T and $B(100) = 0.3480$ T.

[21] The number in parentheses refers to the markings on the x axis of the computer plot.

The field on resonance is

$$B_0 = 0.3402 \pm 0.005 \text{ T},$$

where the error arises from the error in the Hall probe calibration. The frequency was found to be

$$\nu_0 = 9.578 \pm 0.010 \text{ GHz},$$

the error reflecting an estimate of the accuracy of the wavemeter calibration. Thus

$$g_{\text{DPPH}} = \frac{h\nu_0}{\mu_{\text{B}} B_0} = \frac{1}{14.01 \text{ GHz/T}} \frac{9.578 \text{ GHz}}{0.3402 \text{ T}} = 2.01 \pm 0.03 \qquad (7.49)$$

in good agreement with the accepted value

$$g_{\text{DPPH}} = 2.0036.$$

The width of the line is fairly narrow, of order $\Delta B = 8 \times 10^{-4}$ T at full width.

Figure 7.21 shows data for a $CuSO_4$ sample under the same conditions. The frequency is the same as before but the sweep of the field is much wider. It extends from $B(0) = 0.2690$ T to $B(100) = 0.3750$ T. The central field is found to be

$$B_0 = 0.3146 \pm 0.005 \text{ T},$$

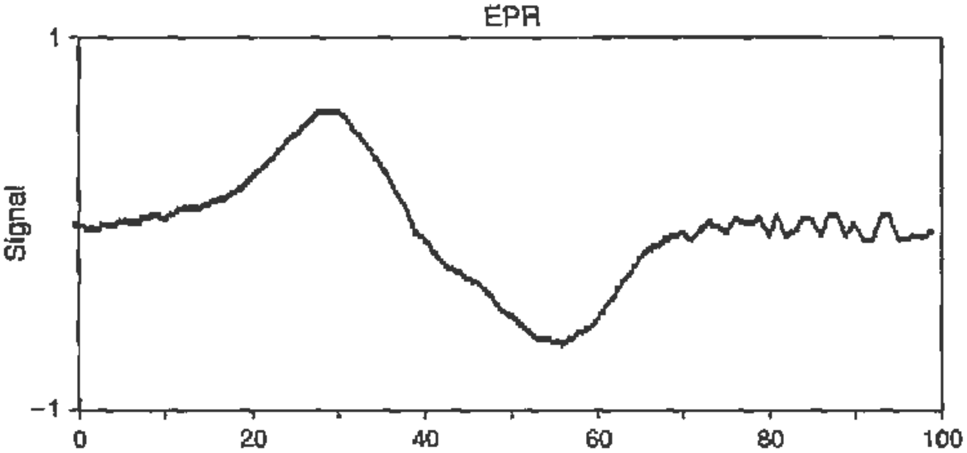


FIGURE 7.21 As described in legend to Fig. 7.20 but for a $Cu(SO_4){\cdot}7H_2O$ sample. Note the large width of the line.

so that

$$g_{\mathrm{CuSO_4}} = \frac{h\nu_0}{\mu_{\mathrm{B}} B_0} = \frac{1}{14.01\ \mathrm{GHz/T}}\,\frac{9.578\ \mathrm{GHz}}{0.3146\ \mathrm{T}} = 2.17 \pm 0.05,$$

where the increased error is from locating the center of the line. This result lies between the known values of the two g factors of the Cu^{2+} ion.[22] What is strikingly different from the DPPH sample is the width of the line, which is $\Delta B_0 = 290 \times 10^{-4}$ T. This is a clear indication of the effects of the crystalline fields in broadening the energy levels of the Cu^{2+} ion.

7.6. REFERENCES

A. Abragam, *The Principles of Nuclear Magnetism*, Oxford Univ. Press, Oxford, 1961. An outstanding work on nuclear magnetic resonance, where the treatment is theoretical and advanced, but very complete and clear.

E. R. Andrew, *Nuclear Magnetic Resonance*, Cambridge Univ. Press, Cambridge, UK, 1956. A shorter text containing experimental details as well; it is very useful to students in this course.

C. H. Townes and A. L. Shawlow, *Microwave Spectroscopy*, McGraw-Hill, New York, 1955. An extensive and comprehensive work on the subject, mainly treating the molecular spectra obtained in gases.

G. E. Pake, *Paramagnetic Resonance*, Benjamin, Elmsford, NY, 1962.

D. J. E. Ingram, *Spectroscopy at Radio and Microwave Frequency*, Butterworth, Stoneham, MA, 1955. Very helpful for the study of paramagnetic resonance in solids and crystalline materials.

E. Fukushima and S. B. W. Roëder, *Experimental Pulse NMR*, Addison-Wesley, Reading, MA, 1981.

[22] For a crystal the g factor depends on the orientation of the crystal axis with respect to the magnetic field. The sample used here was crystalline (powder), and therefore one cannot observe the two g factors, $g_{\parallel}$ and $g_{\perp}$.

Particle Detectors and Radioactive Decay

8.1. GENERAL CONSIDERATIONS

The terms *radiation* and *particle* used in this chapter require clarification. The term *radiation* here designates electromagnetic energy propagating in space (crossing a given area in unit time), but specifically of a frequency higher than that of the visual spectrum, namely, X-rays and gamma rays. Visible, infrared, microwaves, and radiofrequency waves are not included. Because of the quantum-mechanical aspects of the electromagnetic field, such radiation can be described by a flux of (neutral) quanta, the photons, with an energy $E = h\nu$ and a momentum $p = h\nu/c$, where ν is the frequency of the radiation. These quanta interact with electric charges, and the probability for such interactions is of the same order as that for the interaction of two charges.

The term *particle* here encompasses all entities of matter (energy) to which can be assigned discrete classical and quantum-mechanical properties, such as rest mass, spin, charge, lifetime, and so on. The use of the term "particle" is not always clear: for example, we speak of a hydrogen molecule, whereas we refer to the nucleus of the hydrogen atom, the proton, as a particle. Similarly, the electron, the neutron, the (almost) massless neutrino, the π meson, etc., are referred to as particles; the same term is frequently used for a fission fragment, a helium nucleus, or a heavy ion. The visualization of a particle is that of a massive point describing a certain trajectory under the influence of external forces and initial conditions; this provides a useful model for many calculations.

Since particles have dimensions on the order of fermis (10^{-13} cm), they cannot be "seen" even by electron microscopes,[1] but their impact on certain materials, or passage through them, can be noticed readily. Even more remarkably, in certain substances and under specific conditions the whole trajectory of a charged particle can become visible and be permanently recorded. Thus, a particle detector, or radiation detector, is a device that produces a signal (intelligible to the experimenter) when a particle or photon arrives; if the device reveals to the experimenter the whole trajectory of the particle, it is called an image-forming detector.

All detectors are based on the electromagnetic interaction of the charge of the incoming particle with the atoms or molecules of the detector. The different types of interaction (ionization is the most common) and the different principles of amplification of this interaction distinguish the different types of detector. Neutrons, however, are detected through the interaction of the charged particles of the detector to which they transfer energy. This occurs either through elastic collisions of the neutrons with protons (hydrogenous materials), or through neutron capture in certain nuclei, or through the production of fission by the neutron: for example, $n+{}^{10}\mathrm{B}\rightarrow{}^{7}\mathrm{Li}+\alpha$.

In the following discussion we will be concerned with signal-producing devices, which we classify as follows:

(a) Gaseous ionization instruments, encompassing the ionization chamber, the proportional counter, and the Geiger counter,
(b) Scintillation counters,
(c) Solid-state detectors, and
(d) Other detectors.

Such detectors can be designed so as to respond to the passage or arrival of a *single* particle or quantum. They can also be used as integrating devices

[1] High-energy electron-scattering experiments (which serve as a sort of microscope) have, however, revealed much about the electromagnetic structure of the proton and neutron.

(as is frequently done with ionization chambers), giving a signal proportional to $N\overline{E}$, where N is the total number of particles crossing the instrument per unit time and $\overline{E}$ the average energy deposited by each particle.

In evaluating a detector, the following properties are taken into consideration:

(a) Sensitivity, which defines the minimum energy that must be deposited in the detector so as to produce a signal; related to it is the signal-to-noise ratio at the system's output.

(b) Energy resolution, in certain detectors, which are large enough to stop the particle; the signal may be proportional to the initial energy of the particle. In other cases the velocity of the traversing particle can be measured, as in Cherenkov counters, or in dE/dx (ionization per unit length) detectors.

(c) Time resolution, which characterizes the time lag and time jitter from the arrival of the particle until the appearance of the signal, and the distribution in time (duration) of the output pulse; related to it is the dead time of the device, that is, the period during which no (correct) signal will be generated for the arrival of a second particle.

(d) Efficiency, which specifies the fraction of the flux incident on the counter that is detected. It usually is fairly high for charged particles, but can be as low as a few percent for neutral particles and for photons.

Particle detectors play a most important role in nuclear physics, and in many of the experiments described in this text some type of particle detector is used. Just as the spectrograph was the paramount instrument of atomic physics, so the Geiger counter and, later, the NaI scintillation counter have been the paramount instruments of nuclear physics.[2]

In the following sections, we first present a brief discussion of the interaction of charged particles and of photons with matter. Then gaseous ionization instruments are described with specific emphasis on the Geiger counter. This is followed by a description of the scintillation counter and the measurement of nuclear gamma-ray spectra. The following section deals with solid-state detectors and the measurement of the specific ionization of polonium alpha rays in air. Other detectors are mentioned, and some specific experiments using these detectors are described.

[2] It is interesting that the first particle detector ever to be used (by Rutherford in his alpha-particle scattering experiments in 1910) was a scintillating screen, a technique that came again into prominence after 40 years.

Finally, note that precautions should be taken when handling radioactive sources. We recommend that the reader review the material on radiation safety in Appendix D before undertaking the measurements described in this chapter.

8.2. INTERACTIONS OF CHARGED PARTICLES AND PHOTONS WITH MATTER

8.2.1. General Remarks

As already mentioned the interaction of charged particles and photons with matter is electromagnetic and results either in a gradual reduction of energy of the incoming particle (with a change of its direction) or in the absorption of the photon. Particles such as nuclei, protons, neutrons, and π-mesons, are subject to a nuclear interaction as well, which is, however, of much shorter range than the electromagnetic one. The nuclear interaction may become predominant only when the particles have enough energy to overcome Coulomb-barrier effects. A nuclear mean free path, which is approximately 60 g/cm^2, is the distance over which the probability for a nuclear interaction is of order unity.

Heavy charged particles lose energy through collisions with the atomic electrons of the material, while electrons lose energy both through collisions with atomic electrons and through radiation when their trajectory is altered by the field of a nucleus (*bremsstrahlung*—see Section 8.2.6). Photons lose energy through collisions with the atomic electrons of the material, either through the photoelectric or the Compton effect; at higher energies photons interact by creating electron–positron pairs in the field of a nucleus.

A brief review of definitions will be helpful.

(a) *Cross Section.* We define the cross section, σ, for scattering from a *single* target particle as

$$\sigma = \frac{\text{scattered flux}}{\text{incident flux per unit area}}. \tag{8.1}$$

Thus σ has dimensions of area (usually cm^2) and can be thought of as the area of the scattering center projected on the plane normal to the incoming beam. If the density of scatterers is n (particles/cm^3), there will be $n\,dx$

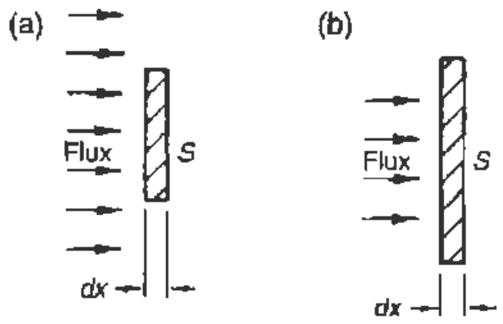


FIGURE 8.1 Scattering of an incoming flux of particles by a target: (a) Area covered by flux is larger than the target area and (b) area covered by flux is smaller than the target area.

scatterers per unit area in a thickness dx of material, and the probability $dP = I_s/I_0$ of an interaction in the thickness dx is

$$dP = \frac{\sigma(I_0/S)}{I_0}(Sn\,dx) = \sigma n\,dx, \tag{8.2}$$

where S is the area covered by the scattering material and I_0 is the total flux incident on the target; thus I_0/S is the flux per unit area as shown[3] in Fig. 8.1a. The result of Eq. (8.2) is not surprising since dP *must* be proportional to n and dx:

$$dP \propto n\,dx,$$

σ is then the factor that transforms this proportionality into an equality. Nuclear cross sections are on the order of 10^{-24} cm^2 (one barn), as expected given the geometrical size (cross section) of the nucleus

$$\sigma_{\text{geom}} = \pi R^2 = 3.14 \times 10^{-26} A^{2/3}\text{cm}^2.$$

(b) *Differential Cross Section.* For a single scatterer we define[4]

$$\frac{d\sigma(\theta,\phi)}{d\Omega} = \frac{\text{flux scattered into element } d\Omega \text{ at angles } \theta, \phi}{\text{incident flux per unit area}}.$$

It follows that

$$\int_0^{2\pi} d\phi \int_0^{\pi} \frac{d\sigma}{d\Omega} \sin\theta d\theta = \sigma,$$

[3] Occasionally confusion arises because the area of the incoming beam may be smaller than the area presented by the target as shown in Fig. 8.1b. The definition of Eq. (8.1) is valid in either case and always leads back to Eq. (8.2).

[4] See the discussion on "solid angle" in Section 9.1.

where the integration is over all angles. If after the scattering process the particle emerges with variable energy, then

$$\frac{d\sigma(\theta,\phi,E)}{d\Omega\, dE} = \frac{\text{flux with energy } E, \text{ within } dE, \text{ scattered into } d\Omega \text{ at angles } \theta,\phi}{\text{incident flux per unit area}}.$$

It follows that

$$\int_0^\infty \frac{d^2\sigma(\theta,\phi,E)}{d\Omega\, dE}dE = \frac{d\sigma(\theta,\phi)}{d\Omega},$$

where the integration is over all possible energies of the scattered flux.

(c) *Absorption Coefficient.* To obtain the probability for scattering in a length x of some material, we consider an incident flux per unit area I_0; $I(x)$ represents the flux at a distance x into the material. According to Eq. (8.2)

$$-dI(x) = I(x)dP = I(x)\sigma n\, dx; \tag{8.3}$$

thus

$$\frac{dI}{I} = -\sigma n\, dx, \qquad I(x) = I_0 e^{-\sigma n x}.$$

If we designate by $P(x)$ the probability for scattering in a length x, we have

$$\begin{aligned} P(x) &= 1 - (\text{probability for survival in a length } x) \\ &= 1 - e^{-\sigma n x} = 1 - e^{-\kappa x}, \end{aligned}$$

where $\kappa = \sigma n$ is the absorption coefficient. Similarly $\lambda = 1/\sigma n$, which has dimensions of length, is called the absorption length, or mean free path.

The density of scattering centers n is given by

$$\begin{aligned} n &= \rho N_0/A && \text{if we consider scattering by nuclei} \\ n_e &= \rho N_0 Z/A && \text{if we consider scattering by electrons} \\ n_N &= \rho N_0 && \text{if we consider scattering by nucleons,} \end{aligned} \tag{8.4}$$

where N_0 is Avogadro's number 6.023×10^{23} and ρ is the density of the material in grams per cubic centimeter; Z and A are the atomic and mass number, respectively.

Often we wish to express the absorption in terms of the equivalent matter traversed, namely, $\xi = \text{g/cm}^2$. Then the thickness of the material can be expressed by $d\xi$, where

$$d\xi = \rho dx.$$

The *mass* absorption coefficient is defined by

$$\mu = \frac{\kappa}{\rho}, \tag{8.5}$$

so that the fraction of a beam *not* absorbed is

$$\frac{I}{I_0} = e^{-\mu\xi}. \tag{8.6}$$

Similarly, if the region of interaction is very thin, the scattered flux is given directly by

$$I_s = I_0 \sigma n \, dx, \qquad \text{for example, for nuclei,} \qquad I_s = I_0 \frac{N_0}{A} \sigma \, d\xi.$$

8.2.2. Energy Loss of a Charged Particle

When a charged particle collides with atomic electrons, as we have already seen in the Frank–Hertz experiment (Section 1.3), it can transfer energy to them only in discrete amounts. It can either excite an electron to a higher atomic quantum state or impart to the electron enough energy so that it will leave the atom; the latter process is the ionization of the atom. Since in our present considerations the incoming particles have considerable energy, the process of ionization is by far the prevailing one, and we will use this term in the discussion.

Let us consider then an atomic electron at a distance b from the path of a heavy charged particle, of charge ze, mass M, and velocity v, as shown in Fig. 8.2a. If we assume that the electron does not move appreciably during the passage of the heavy particle, we can easily obtain the impulse transferred to it due to the electric field, $\mathbf{E}$, of the passing heavy particle:

$$\begin{aligned} I_\perp = \int_{-\infty}^{+\infty} F_\perp(t)\, dt &= e \int_{-\infty}^{+\infty} E_\perp(t)\, dt \\ &= e \int_{-\infty}^{+\infty} E_\perp(t) \frac{dt}{dx}\, dx = \frac{e}{v} \int_{-\infty}^{+\infty} E_\perp(x)\, dx. \end{aligned}$$

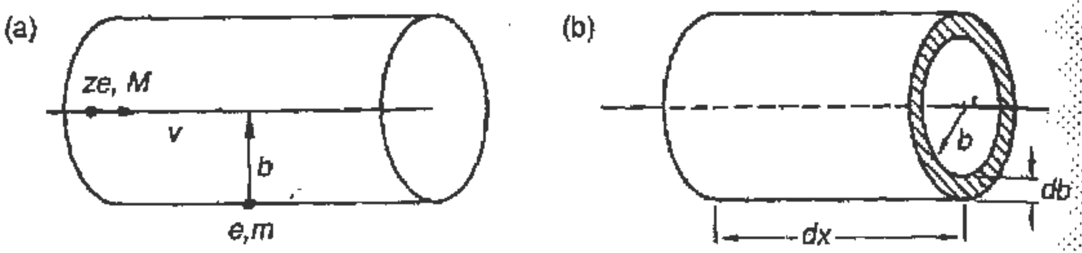


FIGURE 8.2 (a) A particle of charge ze, mass M, and velocity v passes by an electron with an impact parameter b. (b) The differential number of electrons with an impact parameter b in the interval db is given by the volume of the cylindrical shell $2\pi b db\, dx$.

We use only the component of the electric field normal to the particle's trajectory since the longitudinal component averages to 0 when integrated from $-\infty$ to $+\infty$. However, from Gauss's law, integrating over a cylinder of radius b, coaxial with the trajectory (see Fig. 8.2a) we have

$$4\pi ze = \oint \mathbf{E}\cdot\mathbf{dS} = \int_{-\infty}^{+\infty} E_\perp 2\pi b\, dx \qquad \text{and} \qquad \int_{-\infty}^{+\infty} E_\perp dx = \frac{2ze}{b},$$

hence

$$I_\perp = \frac{2ze^2}{vb}.$$

Since the electron was originally at rest, its momentum after the collision, $p = I_\perp$, and the energy transferred is

$$E(b) = \frac{p^2}{2m} = \frac{2z^2e^4}{mv^2b^2}. \tag{8.7}$$

Thus E is a function of the impact parameter b. To obtain the total energy lost by the heavy particle per unit path length, we must count how many electrons it encounters and average over the impact parameters.

From Fig. 8.2b we see that in a cylindrical ring of radius b, width db, and unit height dx, there are contained[5] $n_e 2\pi b db\, dx$ electrons; hence

$$dE(b) = \frac{4\pi n_e dx z^2 e^4}{mv^2}\frac{db}{b}$$

and

$$-\frac{dE}{dx} = \frac{4\pi z^2 e^4}{mv^2} n_e \ln\left[\frac{b_{\max}}{b_{\min}}\right], \tag{8.8}$$

[5] n_e is the electron density as also given by Eq. (8.4).

where because of the logarithm we had to use finite limits on b rather than 0 and ∞. The finite limits are imposed by physical considerations: for $b_{\max}$ we consider the distance where the time of passage of the heavy particle's field becomes of the same order as the period of rotation of the atomic electron in its orbit. Thus

$$\tau = \frac{b}{v} = \frac{1}{\nu} \qquad \text{or} \qquad b_{\max} = \frac{v}{\nu}. \tag{8.9}$$

For the minimum value[6] we equate b to the DeBroglie wavelength of the electron

$$b_{\min} = \frac{\hbar}{p} = \frac{\hbar}{mv}. \tag{8.10}$$

We then obtain

$$-\frac{dE}{dx} = \frac{4\pi z^2 e^4}{mv^2} n_e \ln\left[\frac{mv^2}{\hbar\nu}\right]. \tag{8.11}$$

The frequencies of the atomic electrons ν are, however, different for each orbit, so that a suitable average must be taken; we thus replace $\langle\hbar\nu\rangle$ with an average ionization potential $\bar{I}$. Finally, inclusion of relativistic effects and a precise calculation give

$$-\frac{dE}{dx} = \frac{4\pi z^2 e^4}{mv^2} n_e \left[\ln\frac{2mv^2}{\bar{I}(1-\beta^2)} - \beta^2\right] \tag{8.12}$$

for the energy loss of heavy particles due to ionization.

In Eq. (8.12), $\beta = v/c$, and we see that the energy loss is only a function of the velocity, v, of the charge ze of the incoming particle, and of the electron density, n_e, of the scattering material. Note that in Eq. (8.12), m is the mass of the *electron* while the mass of the incoming particle does not appear at all.

Before further investigating Eq. (8.12), we should note the following effects:

(a) Equation (8.12) was derived on the assumption that the incoming particle is not deflected, and thus it is valid only for *heavy particles*; for electrons the term in the parentheses must be slightly modified.

[6] An alternate approach is to set $b_{\min}$ such that maximum energy is transferred to the electron. Because of momentum conservation we have $p_{\max} = 2mv$ leading to $b_{\min} = ze^2/mv^2$.

(b) Electrons also lose energy through their interaction with the nucleus, and this is the prevailing mechanism at high energies. That is, the electron's trajectory is bent by the field of the nucleus, which implies an acceleration (since the velocity vector changes), and from electrodynamics we know that accelerated charges radiate. This radiation, called "bremsstrahlung," is discussed in Section 8.2.6.

(c) For extremely relativistic particles, $v \approx c, \beta \approx 1$, Eq. (8.12) predicts a continuous rise in dE/dx proportional to $\ln \gamma^2$ where $\gamma = E/mc^2 = 1/(1-\beta^2)^{1/2}$. Such a fast rise, however, is not observed experimentally. This is due to polarization of the medium: the electrons that are being set into motion by the field of the incoming particle move so as to reduce the effect of the external field. Consequently a much slower rise with energy results; the correct expression is[7]

$$-\frac{dE}{dx}\bigg]_{\text{ion}} = \frac{4\pi z^2 e^4}{mc^2} n_e \left[\ln \frac{2\gamma\, mc^2}{I'} + \frac{1}{2}\right], \tag{8.13}$$

where

$$I' = \hbar\omega_p = \sqrt{\frac{4\pi n_e z e^2}{m}}.$$

For silver bromide $I' \approx 48$ eV.

(d) For low-energy particles we obtain from Eq. (8.12)

$$-\frac{dE}{dx} \propto \frac{z^2}{v^2} = \frac{z^2 M}{2E},$$

where M is the mass of the incoming particle and E its kinetic energy. The above expression (when applicable) is useful since a measurement of dE/dx and of E identifies the incoming particle

$$E\left(\frac{dE}{dx}\right) \propto z^2 M.$$

(e) In image-forming devices and particularly in nuclear emulsions, the density of developed silver bromide grains can be used as a measure of the particle's velocity because of the dependence of Eq. (8.12) and Eq. (8.13) on β. However, the density of the track depends only on energy

[7] See J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Section 13.1, Wiley, New York, 1999.

transfers ≤ 5 keV, since when an atomic electron acquires more energy, its own track becomes visible and separated from the primary particle's track; such electrons are called knock-ons or *delta rays*. The energy-loss expression for energy transfers ≤ 5 keV does not exhibit at all the relativistic rise of Eq. (8.13), but for high values of γ, stabilizes at a plateau 1.2 times the minimum value.

The energy loss of a heavy particle in a typical absorber, such as nuclear emulsion, as a function of the logarithm of its kinetic energy (in units of rest energy) is given in Fig. 8.3. Strictly speaking, this curve holds only for a given absorber and all singly charged particles, since we know from Eqs. (8.12) and (8.13) that dE/dx is a function only of the *velocity* of the incoming particle and its charge. (Note that K.E.$/mc^2 = \gamma - 1$, which has a one-to-one correspondence to β.) However, the general behavior of this curve holds for all absorbers.

We do recognize four regions of interest: (a) near the stopping point where a Bragg curve is applicable (see Fig. 8.32); (b) the low-energy region where the $1/v^2$ dependence of Eq. (8.12) dominates, and tends asymptotically toward the value $1/c^2$; (c) the relativistic region, where because of the rise of the logarithmic term, a minimum appears approximately at $\gamma = 1$; and (d) tbe screened region in which Eq. (8.12) becomes applicable. Had polarization effects not been included, the rise of the dE/dx curve in this last region would be steeper than indicated in Fig. 8.3. The lower curve in Fig. 8.3 (energy transfers ≤ 5 keV) is applicable to the grain density in nuclear emulsions.

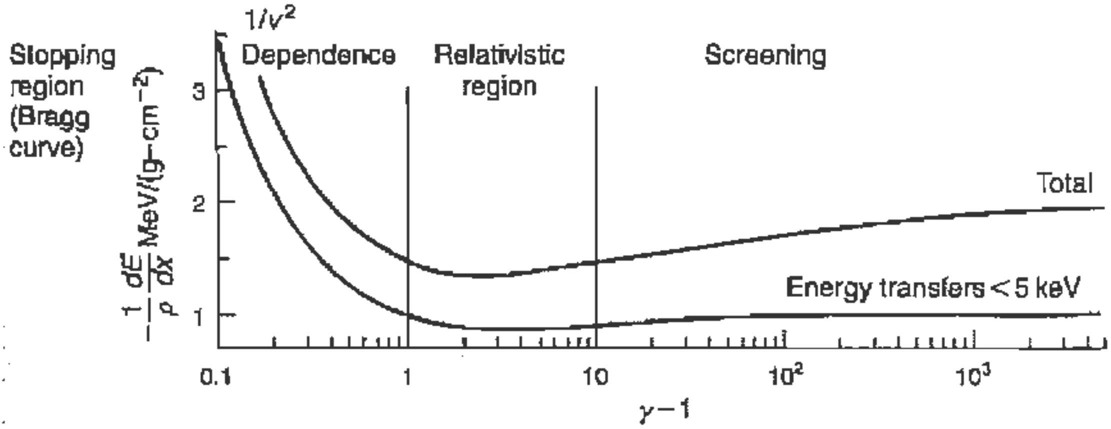


FIGURE 8.3 The universal energy-loss curve for a singly charged particle plotted in MeV/(g–cm^{-2}) against $\gamma - 1$. Note the upper curve for the total energy loss and the lower curve for energy loss involving only energy transfers smaller than 5 keV.

If we choose to calibrate the abscissa of Fig. 8.3 in units of energy (MeV) of the particle rather than by $\gamma - 1$, we will not have a universal curve any more, but for each particle the energy-loss curve will be shifted horizontally by m_1/m_2, in such fashion that $dE/dx]_{m_1} = dE/dx]_{m_2}$ when the corresponding kinetic energies T_1 and T_2 result in the same value of $\gamma - 1$:

$$\frac{T_1}{m_1} = \frac{T_2}{m_2} = c^2(\gamma - 1).$$

This is shown in Fig. 8.4, which gives the absolute value of energy loss $-dE/d\xi$ (in MeV/(g–cm^{-2})) in air for protons (curve 1) and π-mesons ($m_\pi = 140$ MeV; curve 2), where the latter is shifted to the left by a factor $m_\pi/m_p = 0.150$.

Further, if we consider particles of different z, the energy loss will differ by the ratio $(z_1/z_2)^2$. In this fashion we obtaiu curve 3 in Fig. 8.4, the energy loss of alpha particles in air, which is shifted (with respect to curve 1) to the right by a factor of $m_\alpha/m_p = 4$ and upward by a factor of $(z_\alpha/z_p)^2 = 4$.

If we now turn our attention to the dependence of dE/dx on the absorber material, it is clear that it will vary rapidly, due to its dependence on n_e. If instead we use $-dE/d\xi$ (the energy loss per g/cm^2 of material) the variation is much slower, since

$$n_e = \rho N_0 \frac{Z}{A}$$

and

$$d\xi = \rho\, dx.$$

Thus

$$-\frac{dE}{d\xi} = N_0 \frac{Z}{A} z^2 f(\beta, \bar{I}),$$

so that the energy loss per g/cm^2 is larger for low Z materials, neglecting the small dependence on $\bar{I}$, the average ionization potential. Curve 4 of Fig. 8.4 gives $-dE/d\xi$ for protons in lead, which is indeed lower than that in air, but not by a large amount.

An approximate universal figure for the energy loss of a relativistic singly charged particle in any materials is 2 MeV/(g-cm^{-2}).

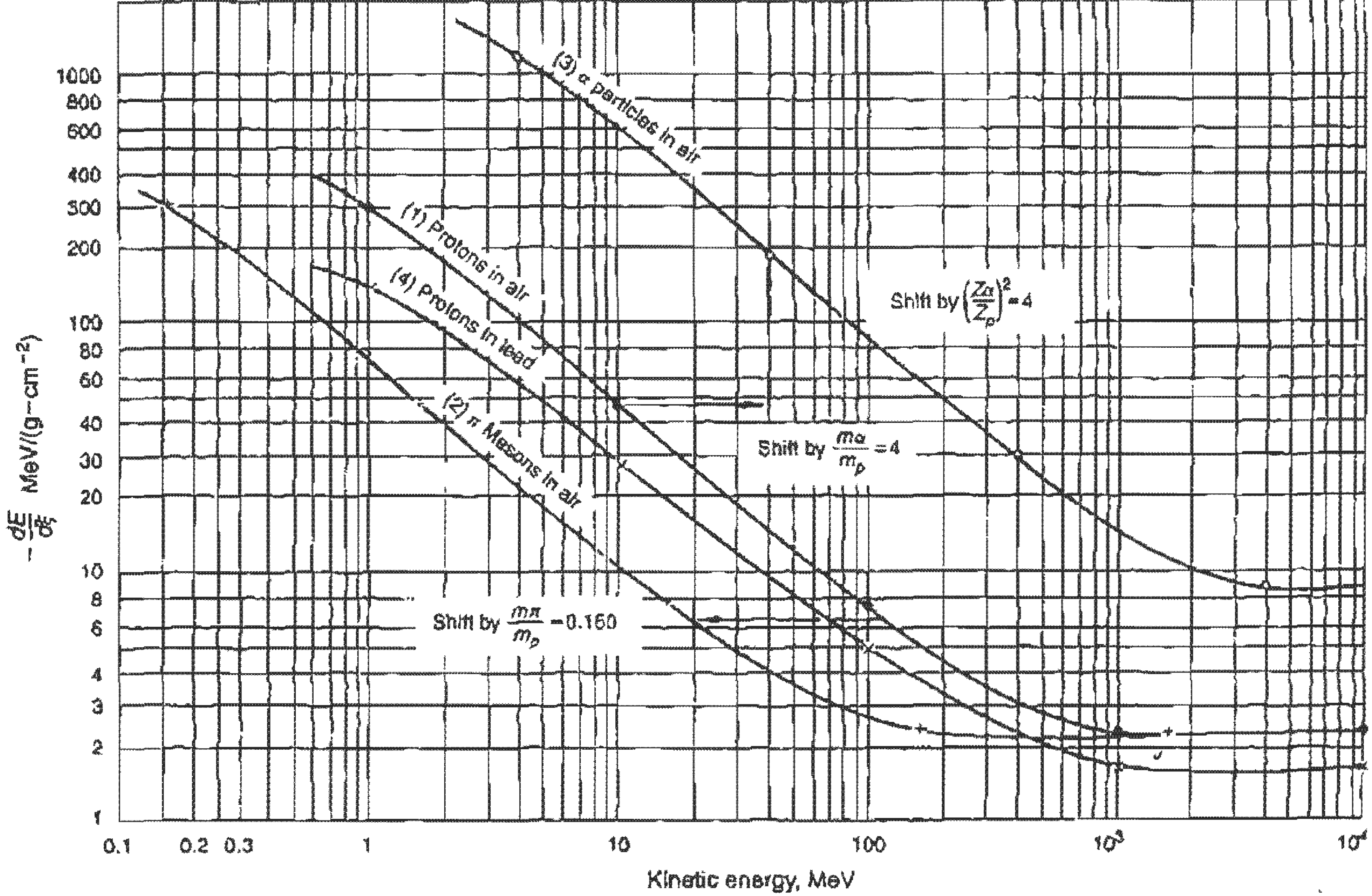


FIGURE 8.4 Energy-loss curves for different charged particles in air and in lead. Note how all the curves are related to each other.

8.2.3. Range of a Charged Particle

Since the exact expression for the energy loss of a charged particle is known, it is possible by integration to find what total length of material an incoming particle of given energy will traverse before coming to rest; this is called its range R, and we can set

$$E_0 = \int_0^R \frac{dE}{dx} dx$$

or conversely, since $dE/dx = z^2 n_e f_1(\beta)$ and[8] $dE = M f_2(\beta)\, d\beta$ (M is the mass of the incoming particle),

$$R = \int_0^R dx = -\frac{1}{z^2 n_e} \int_{E_0}^0 \frac{dE}{f_1(\beta)} = \frac{M}{z^2 n_e} \int_0^{\beta_0} \frac{f_2(\beta)}{f_1(\beta)} d\beta = \frac{M}{z^2 n_e} F(\beta_0). \tag{8.14}$$

That is, for the same velocity the range is proportional to the mass of the incoming particle, inversely proportional to the square of its charge, and inversely proportional to the electron density of the stopping material. Extensive tabulations of range curves for different particles and different absorbers are available.[9] Also various empirical formulas have been devised; for example, for electrons, Feather's expression gives for the range of electrons of aluminum (in g/cm^2)

$$R = 0.543E - 0.160, \qquad E > 0.8 \text{ MeV}, \tag{8.15}$$

where E is the initial kinetic energy of the electron (in MeV).

As suggested above, it is highly preferable to express the range in grams per squared centimeter, because then the dependence on the absorber material is slow (since $n_e/\rho = N_0 Z/A$), resulting in a larger range (in g/cm^2) in heavy elements.

Figure 8.5 gives the range (in g/cm^2) of protons, π-mesons, and alpha particles as a function of their kinetic energy for air. As explained for Fig. 8.4, the π-meson curve is obtained from the proton curve by shifting to the left by the factor $m_\pi/m_p = 0.15$ to reach the same β_0, but also by multiplying the ordinate values by $m_\pi/m_p = 0.15$; for the alpha particles the curve is obtained by shifting to the right by the factor $m_\alpha/m_p = 4$ and

[8] Nonrelativistically we have the simple relation $dE = Mc^2 \beta\, d\beta$.

[9] See, for example, the compilation by the Particle Data Group (2000; see Section 8.7).

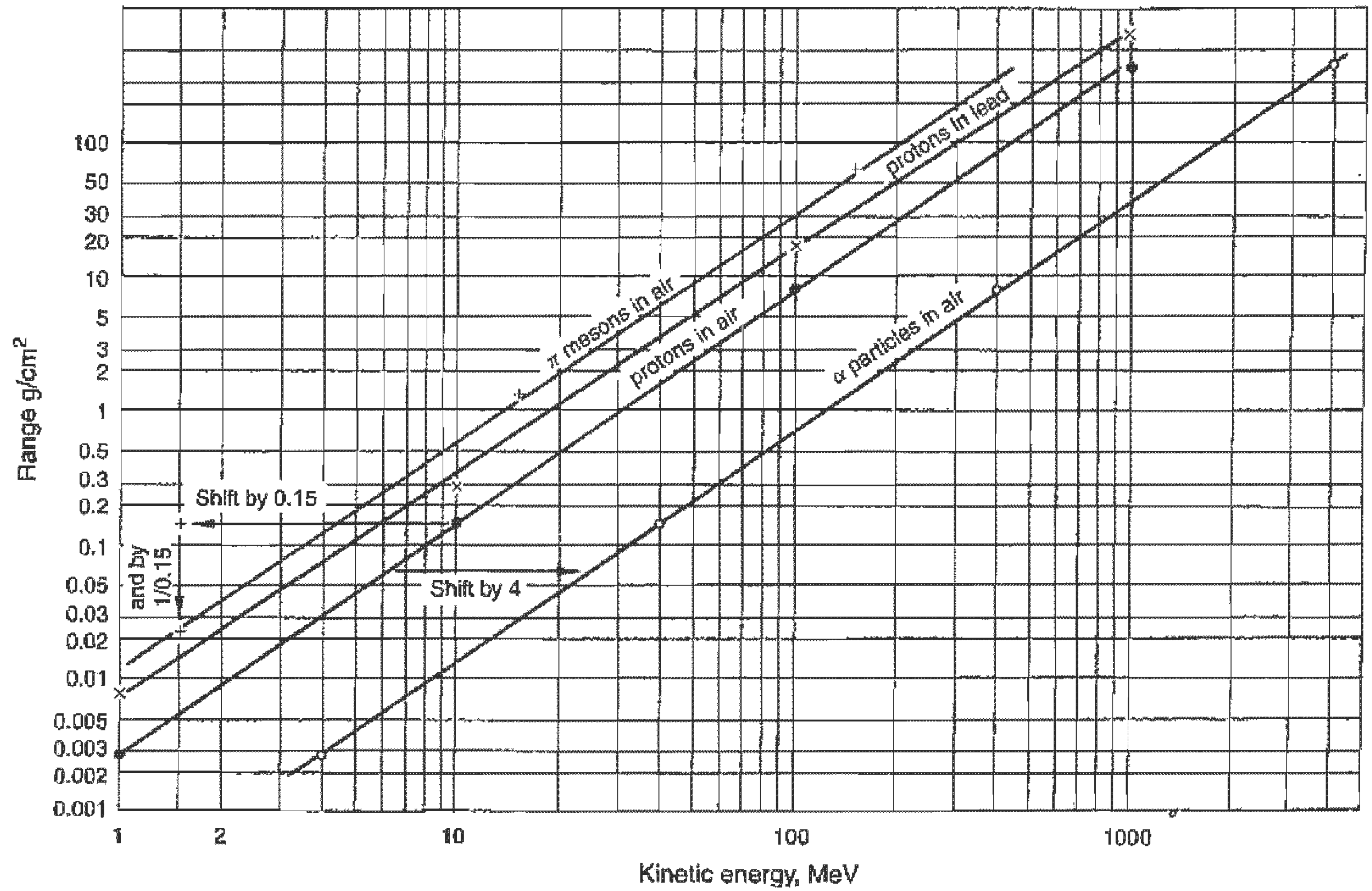


FIGURE 8.5 Range curves for different particles in air and in lead. Note how the different curves are related to each other.

multiplying the ordinate (first) by $m_\alpha/m_p = 4$ (due to the different mass) and then by $(z_p/z_\alpha)^2 = 1/4$, hence leaving it unshifted.

Finally, the range of protons in lead is also given. The concept of range loses its meaning, however, when the amount of material that the particle must traverse before coming to rest is on the order of a nuclear mean free path as explained in the introduction to this section.[10]

8.2.4. Multiple Scattering

In discussing the passage of a charged particle through matter, we have neglected up to now its interaction with the electric field of the nucleus, because indeed the energy transfer to the nucleus is minimal. However, when a particle of charge ze, mass m, and velocity v passes by the vicinity of a nucleus of charge Ze, it will be scattered (Fig. 8.6) with the Rutherford cross section

$$\frac{d\sigma}{d\Omega} = \frac{1}{4}\left(\frac{e^2 Zz}{mv^2}\right)^2 \frac{1}{\sin^4 \theta/2}, \tag{8.16}$$

showing that the probability for small-angle scattering is predominant. For such small angles we approximate the angle of deflection by

$$\theta = \frac{\Delta p}{p} = \frac{2Zze^2}{pvb}, \tag{8.17}$$

where p is the momentum of the particle and b is the impact parameter.

During its traversal of the material, the incoming particle suffers many small-angle scatterings. It can be shown that the resultant scattering angle θ, after traversal of a finite thickness of material D, has a Gaussian[11] distribution about the mean $\theta = 0$; the probability for a scattering through an angle Θ within the interval $d\Theta$ is

$$P(\Theta)\, d\Theta = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\Theta}{\sigma}\right)^2\right].$$

The standard deviation is $\sigma = \sqrt{\overline{\theta^2}}$ (the root mean square scattering angle).

[10] For heavy ions, energy loss due to collisions with the nuclei must also be considered.

[11] See Chapter 10.

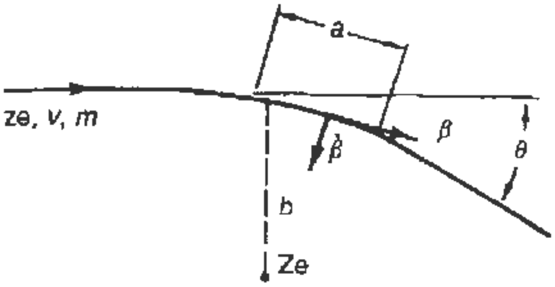


FIGURE 8.6 Deflection of a charged particle when passing in the vicinity of a nucleus. Note the scattering angle θ.

For the mean square scattering angle we have

$$\overline{\theta^2} = \frac{8\pi z^2 Z^2 e^4}{v^2 p^2} n D \ln\left(\frac{a_0 v p}{2Z^{4/3} z e^2}\right). \tag{8.18}$$

where a_0 is the Bohr radius. We further simplify Eq. (8.18) in order to exhibit the dependence of $\overline{\theta^2}$ on the incoming particle's charge z, velocity β, and momentum p, as well as on the material's thickness D, the density of nuclei $n = \rho N_0/A$, and the atomic number Z: we obtain

$$\sqrt{\overline{\theta^2}} = \frac{z}{p\beta}\sqrt{DZ^2 n}F,$$

where F is a slowly varying function of the parameters of the incoming particle and the scattering material (it contains the logarithmic term and constants). Furthermore $1/(Z^2 n)$ is proportional to the "radiation length" L_{rad} of the material (defined in Section 8.2.6 below), which is frequently tabulated, so that we finally write

$$\overline{\theta^2} = |\theta|_{\text{rms}} = z\frac{21.2(\text{MeV}/c)}{p\beta}\sqrt{\frac{D}{L_{\text{rad}}}}(1+\epsilon), \tag{8.19}$$

where $|\theta|_{\text{rms}}$ is in radians, and p must be expressed in MeV/c; ϵ is a small correction[12] depending both on the scattering material and on β/z of the incoming particle. When we are interested in the rms *projected* angle, the numerical factor in Eq. (8.19) must be replaced by 15 (MeV/c).

[12] Calculated from Moliére theory; see U.C.R.L. Report 8030 by W. Barkas and A. H. Rosenfeld for tables of ϵ.

8.2.5. Passage of Electromagnetic Radiation (Photons) through Matter

As mentioned in the introduction to this section photons lose energy or are absorbed in matter by one of the following three mechanisms:

(a) Photoelectric effect, which predominates at low energies,

(b) Compton effect, which predominates at medium energies (below a few MeV), and

(c) Pair production of electrons and positrons, which is dominant in the high-energy region.

The relative importance of these processes and the energies at which they set in are best seen in Fig. 8.7, which gives the cross section for the interaction of a photon as a function of its energy (in units of the electron's rest mass). We will now briefly consider each process separately.

(a) *Photoelectric Effect.* We speak of the photoelectric effect when the photon is completely absorbed and all its energy is transferred to an atomic electron. Consequently the photon must have enough energy to excite the bound electron from its quantum state to a higher state or into the continuum; the latter process (ionization of the atom) is much more probable. Since the binding energy of the inner electrons in atoms is on the order of kiloelectronvolts, as the frequency of the photon is increased and it reaches the value of the binding energy of a particular shell,[13] a new "channel" opens, and we expect a sudden rise in the absorption cross section. Apart from the onset of new channels, the overall variation of the photoelectric effect is a rapid decrease as the third power of the photon frequency (as $\nu^{-7/2}$), thus resulting in the curve shown on the left in Fig. 8.7. The cross section for the photoelectric effect is derived in Heitler (1954),[14] from which we give the nonrelativistic value for the ejection of *one* electron from the K shell, when the photon energy is not too close to

[13] Note that $n = 1$ electrons are said to be in the K shell, $n = 2$ in the L shell, $n = 3$ in the M shell, etc.

[14] W. Heitler, *The Quantum Theory of Radiation*, 3rd ed., pp. 207 and 208, Oxford Univ. Press, Oxford, 1984.

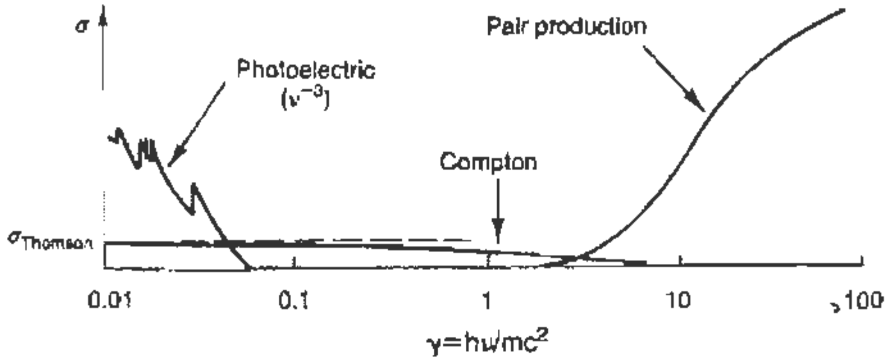


FIGURE 8.7 The cross section for the interaction of photons with matter as a function of their energy (expressed in units of the electron's rest mass).

the absorption edge,

$$\sigma_{\text{P}} = \sigma_{\text{T}} \frac{Z^5}{(137)^4} 2\sqrt{2} \left[\frac{h\nu}{mc^2} \right]^{-7/2} \ (\text{cm}^2). \tag{8.20}$$

Note the dependence on the Z of the nucleus, indicating that L shell and higher-shell ejection is less probable because of the screening of the nuclear charge. Here σ_{T} is the classical Thomson cross section, which is derived from the simplified assumption of a plane polarized elecromagnetic wave scattering from a free electron (it is assumed that the displacement of the electron is much smaller than the wavelength); we obtain

$$\sigma_{\text{T}} = \frac{8\pi}{3} \left[\frac{e^2}{mc^2} \right]^2 = \frac{8\pi}{3} r_0^2, \tag{8.21}$$

where $r_0 \equiv e^2/mc^2$ is the classical radius of the electron $= 2.8 \times 10^{-13}$ cm. Note that the Thomson cross section is independent of the frequency of the incoming photon.

(b) *Compton Effect.* In the Compton effect, the photon scatters off an atomic electron and loses only part of its energy. This phenomenon, which is one of the most striking quantum effects, is described in detail in Section 9.2; the cross section for Compton scattering is given by the Klein–Nishina (K–N) formula, shown in an expanded scale in Fig. 8.8. The energy of the photon is given on the abscissa in units of the electron rest mass[15] $\gamma = h\nu/mc^2$, and the ordinate gives the ratio of the Compton cross section σ_{C} to the classical Thomson cross section σ_{T}.

[15] Not to be confused with the usual definition of γ for a charged particle $\gamma = E/mc^2$, introduced in Eq. (8.13).

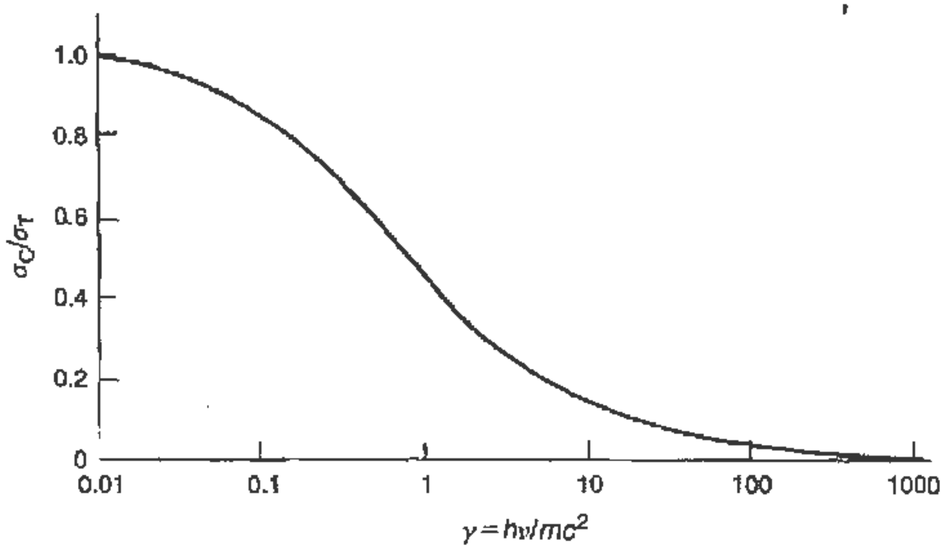


FIGURE 8.8 The ratio of the Compton scattering cross section, σ_C, to the constant Thomson cross section, σ_T, as a function of photon energy expressed in units of the electron's rest mass.

We give below the asymptotic approximations to the (K–N) Compton scattering cross section:

For low energies:

$$\sigma_C = \sigma_T \left(1 - 2\gamma + \frac{26}{5}\gamma^2 + \cdots\right) \qquad \gamma = h\nu/mc^2 \ll 1$$

For high energies:

$$\sigma_C = \frac{3}{8}\sigma_T \frac{1}{\gamma}\left(\ln 2\gamma + \frac{1}{2}\right) \qquad \gamma = h\nu/mc^2 \gg 1. \tag{8.22}$$

(c) *Pair Production.* In pair production a photon of sufficiently high energy is converted into an electron–positron pair. For a free photon conservation of energy and momentum would not be possible in this process, so pair production must take place in the field of a nucleus (or of another electron), which will take up the balance of momentum. Clearly the threshold for this process is $2mc^2$ (where m is the mass of the electron), hence 1022 keV. The cross section for pair production rises rapidly beyond the threshold, and reaches a limiting value for $h\nu/mc^2 \approx 1000$ given by[16]

$$\sigma_{\text{pair}} = \frac{Z^2}{137} r_0^2 \left[\frac{28}{9} \ln \frac{183}{Z^{1/3}} - \frac{2}{27}\right] \quad (\text{cm}^2). \tag{8.23}$$

[16] See Heitler (1984), p. 260.

Since both the photoelectric and Compton effect cross sections decrease as the photon energy rises, pair production is the predominant interaction mechanism for very high-energy photons.

It is advantageous to introduce the mean free path (L_{pair}) for pair production; when a photon traverses a material with density of nuclei n,

$$L_{\text{pair}} = \frac{1}{n\sigma_{\text{pair}}} = \frac{1}{(28/9)(Z^2 n/137)r_0^2 \ln(183/Z^{1/3})}, \tag{8.24}$$

where we have dropped the small term 2/27. Thus, the attenuation of a beam of I_0 photons will proceed as

$$I(x) = I_0 e^{-x/L_{\text{pair}}}. \tag{8.25}$$

In conclusion, Fig. 8.9 gives the total absorption coefficient for a photon traversing lead as a function of its energy (in units of the electron rest mass). Note that

$\kappa_{\text{P}} = \sigma_{\text{P}} 2n$ because there are 2 K-shell electrons per nucleus
$\kappa_{\text{C}} = \sigma_{\text{C}} n_e$ electron density
$\kappa_{\text{pair}} = \sigma_{\text{pair}} n$ density of nuclei.

The dashed curves in Fig. 8.9 indicate the relative contributions of each of the three interaction mechanisms.

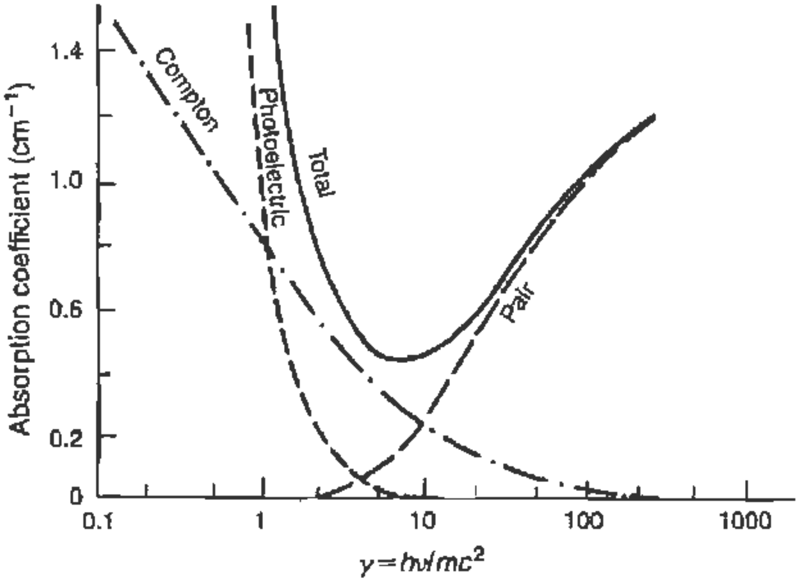


FIGURE 8.9 The relative contribution of the three effects responsible for the interaction of photons with matter. The absorption coefficient in lead is plotted against the logarithm of photon energy (in units of the electron's rest mass).

8.2.6. Interaction of Electrons with Matter (Bremsstrahlung)

Since electrons carry charge, their interaction with matter must follow along the lines given in Section 8.2. Because of their small mass, however, their interaction with the nucleus results in significant energy loss by radiation; this process, called "bremsstrahlung," becomes the dominant mode of energy loss for high-energy electrons.

We can obtain an estimate of the cross section for "bremsstrahlung" from a classical nonrelativistic model. Consider an electron (charge e, mass m, and velocity v) passing by the vicinity of a nucleus of charge Ze, and let us assume that in the collision process the nucleus does not move (Fig. 8.6). The scattering angle of the electron is given by Eq. (8.17), and the change in the velocity vector of the electron is

$$\Delta v = \frac{2Ze^2}{mvb}. \tag{8.26}$$

The radiation formula for an accelerated charge[17] is

$$P(t) = \frac{dE}{dt} = \frac{2}{3}\frac{e^2}{c}\left[(\dot{\boldsymbol{\beta}})^2 - (\boldsymbol{\beta} \times \dot{\boldsymbol{\beta}})^2\right]. \tag{8.27}$$

So for our case, since $\dot{\boldsymbol{\beta}}$ is normal to $\boldsymbol{\beta}$,

$$dE(t) = \frac{2}{3}\frac{e^2}{c}\,|\dot{\boldsymbol{\beta}}|^2\,dt. \tag{8.28}$$

By a general theorem of Fourier analysis, if

$$E = \frac{2}{3}\frac{e^2}{c}\int_{-\infty}^{+\infty} |A(t)|^2\,dt,$$

then also

$$E = \frac{2}{3}\frac{e^2}{c}\int_{-\infty}^{+\infty} |A(\omega)|^2\,d\omega,$$

[17] See Jackson (1999), p. 666. In Eq. (8.27) γ was set equal to 1; similarly Eq. (8.28) should include a term $(1-\beta^2) = 1/\gamma^2$, which was also set equal to 1.

where

$$A(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} A(t) e^{i\omega t}\, dt \tag{8.29}$$

is the Fourier transform amplitude of $A(t)$.

Using then Eq. (8.29), we obtain in analogy with Eq. (8.28) the frequency spectrum of the radiation[18]

$$dE(\omega) = \frac{2}{3}\frac{e^2}{c}\left[|A(\omega)|^2 + |A(-\omega)|^2\right] d\omega = \frac{4}{3}\frac{e^2}{c}\, |A(\omega)|^2\, d\omega. \tag{8.30}$$

To evaluate $dE(\omega)$ we must perform the integral indicated in Eq. (8.29) with $A(t) = |\dot{\boldsymbol{\beta}}|$. We assume that the acceleration $\Delta\beta$ occurs in a very brief interval of time, on the order of $\tau = a/v$, where a is the characteristic distance over which the force is appreciable[19]; then

$$A(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} |\dot{\boldsymbol{\beta}}|\, e^{i\omega t}\, dt = \begin{cases} \frac{1}{\sqrt{2\pi}}\Delta\beta & \omega\tau < 1 \\ 0 & \omega\tau > 1. \end{cases} \tag{8.31}$$

If $\omega\tau > 1$, there will be several oscillations of the exponential term over the region where $|\dot{\boldsymbol{\beta}}|$ is different from zero, and the integral will average to zero.

The integral results in a rectangular spectrum for the emitted radiation, as shown in Fig. 8.10, with

$$\frac{dE}{d\omega} = \begin{cases} = \frac{2e^2}{3\pi c}\frac{4Z^2e^4}{c^2m^2v^2b^2} & \omega\tau < 1 \\ = 0 & \omega\tau > 1. \end{cases} \tag{8.32}$$

Next we integrate over all impact parameter b to obtain the total radiated energy at frequency ω when the electron passes by a nucleus

$$\chi(\omega) = \int_{b_{\text{min}}}^{b_{\text{max}}} \frac{dE(\omega)}{d\omega} 2\pi b\, db,$$

where we can set $b_{\text{max}} = a = \tau v$ and in view of $\omega\tau \sim 1$ we also let $b_{\text{max}} \sim v/\omega$; from classical considerations (see footnote 6)

$$b_{\text{min}} = \frac{Ze^2}{mv^2}.$$

[18] Because $A(t)$ is real, $A(\omega) = A^*(-\omega)$.

[19] See, for example, W. K. H. Panofsky and M. Phillips, *Classical Electricity and Magnetism*, p. 304, Addison-Wesley, Reading, MA, 1955.

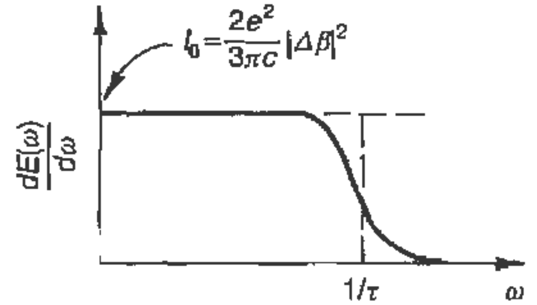


FIGURE 8.10 Idealized bremsstrahlung spectrum resulting from the sudden acceleration $|\Delta\beta|$ of a charged particle.

The cross section σ_{brems}, giving the probability of emission of a photon of energy $\hbar\omega$ in the interval $d(\hbar\omega)$, is related to $\chi(\omega)$ through

$$(\hbar\omega)\sigma_{\text{brems}}(\omega)d(\hbar\omega) = \chi(\omega)\,d\omega,$$

resulting in the classical nonrelativistic bremsstrahlung cross section

$$\sigma_{\text{brems}} = \frac{16}{3}\frac{Z^2e^2}{\hbar c}\left(\frac{e^2}{mc^2}\right)^2\left(\frac{c}{v}\right)^2\frac{1}{\hbar\omega}\ln\left(\frac{mv^3}{Ze^2\omega}\right). \tag{8.33}$$

The average energy loss per path length, $-dE/dx$, is obtained by integrating over all photon energies (the square pulse) and multiplying by the density of nuclei:

$$-\frac{dE}{dx} = \int n(\hbar\omega)\sigma_{\text{brems}}d(h\omega) = n(\hbar\omega)(\hbar\omega_{\text{max}})\sigma_{\text{brems}}.$$

Substituting $1/137 = e^2/\hbar c$, $r_0 = e^2/mc^2$, and $(\hbar\omega_{\text{max}}) = E_0$, the energy of the electron, we obtain

$$-\left(\frac{dE}{dx}\right)_{\text{av}} = \frac{16}{3}\frac{Z^2n}{137}E_0r_0^2\left(\frac{c}{v}\right)^2\ln\left(\frac{mv^3}{Ze^2\langle\omega\rangle_{\text{av}}}\right). \tag{8.34}$$

Equation (8.33) is a fair approximation; the correct quantum-mechanical result, including the screening of the nucleus by the atomic electrons, is given by[20]

$$-\left(\frac{dE}{dx}\right)_{\text{av}} = \frac{Z^2n}{137}E_0r_0^2\left(4\ln\frac{183}{Z^{1/3}} + \frac{2}{9}\right). \tag{8.35}$$

[20] See Heitler (1984), p. 253.

The mean free path for bremsstrahlung by an *electron*, called the "radiation length," is defined as

$$L_{\text{rad}} = \frac{1}{n\sigma_{\text{brems}}} = \frac{1}{4(Z^2n/137)r_0^2 \ln(183/Z^{1/3})}, \qquad (8.36)$$

which is obtained from Eq. (8.35) by setting $L_{\text{rad}} = dx$, when $-dE/E_0 = 1$; the term $\frac{2}{9}$ (small as compared to the ln) was dropped.

To show at what electron energies bremsstrahlung becomes important, we note that

$$\frac{(dE/dx)_{\text{rad}}}{(dE/dx)_{\text{ioniz}}} = \frac{ZE(\text{MeV})}{800}.$$

This is shown in Table 8.1, where we give for some common absorbers, L_{rad}, as well as the electron energy at which bremsstrahlung loss becomes equal to ionization loss.

Equation (8.36) is amazingly similar to Eq. (8.24), by which we defined L_{pair}. We have

$$L_{\text{rad}} = \frac{7}{9} L_{\text{pair}},$$

indicating that in matter the mean free path of a high-energy electron is of the same order as the mean free path of a high-energy gamma ray; this is the reason for the phenomenon of the electromagnetic cascade, first observed in cosmic rays.

If a very high-energy electron is incident in the atmosphere, it will soon (after approximately 330 m) emit one or more high-energy gamma rays. These gamma rays will soon again (after approximately 330 m) produce electron–positron pairs. Each of the secondary electrons and positrons will again radiate, and so on, until most of the energy of the primary electron

TABLE 8.1 Radiation Length of Electrons in Different Materials

Material	L_{rad}	Electron energy for $(dE/dx)_{\text{rad}} = (dE/dx)_{\text{ioniz}}$
Air	330 m	120 MeV
Aluminum	9.7 cm	52 MeV
Lead	0.52 cm	7 MeV

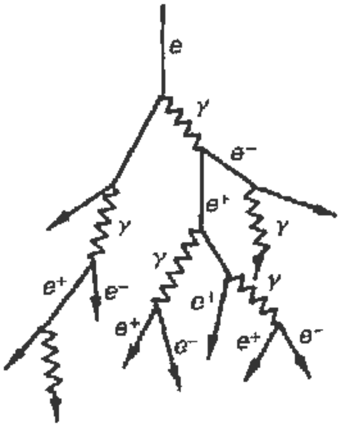


FIGURE 8.11 Formation of an electromagnetic cascade. Note that high-energy electrons (positrons) radiate gamma rays and the gamma rays later convert into electron–positron pairs and so forth.

(or gamma ray) has been transferred to many less energetic electrons (Fig. 8.11).

In another connection we have already used L_{rad} in Eq. (8.19) for multiple scattering; from Table 8.1 we see that in heavy materials scattering will be much more pronounced. Note that multiple scattering is the same for particles of the same momentum. Thus, at low energies a light particle will scatter much more than a heavier particle of the same kinetic energy ($p = \sqrt{2Tm}$). This is clearly seen when observing the tracks of low-energy protons and electrons in an image-forming device; the former ones are, in general, straight, whereas the latter ones suffer multiple scattering through large angles.

8.3. GASEOUS IONIZATION DETECTORS; THE GEIGER COUNTER

8.3.1. General

As mentioned earlier, most particle detectors are based in one form or another on the energy lost by the charged particle due to ionization of the medium it traverses. In a large class of instruments the detecting material is a gas; the ionization potentials are on the order of 10 eV, but on the average, for example in air, the charged particle loses 30 to 35 eV for each electron–ion pair formed.[21] By collecting the free charges that were thus

[21] This is due to additional interactions such as excitation and elastic scattering.

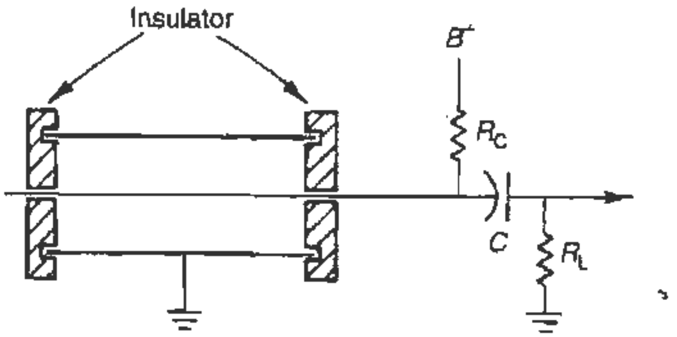


FIGURE 8.12 Diagrammatic arrangement of a cylindrical Geiger counter; the central wire is charged to B^+ through R_C while the cylindrical envelope is held at ground. The output signal appears across R_L.

created, it is possible to obtain an electrical pulse, signaling the passage of the charged particle.

The simplest type of gaseous detector consists of a cylindrical chamber with a wire stretched along its center, as shown in Fig. 8.12. The chamber walls act as the negative electrode, and positive voltage is applied to the central electrode. Under the influence of the electric field, the electrons are collected at the center while the positive ions move toward the walls. It is desirable to collect the free charges before they recombine in the gas; this is mainly a function of the pressure of the gas and of the applied voltage.[22]

If, however, the voltage is sufficiently raised, the electrons gain enough energy to ionize through collision further atoms of the gas, so that there is a significant multiplication of the free charges originally created by the passage of the particle. In Fig. 8.13, Curve 1 gives the number of electron–ion pairs collected as a function of applied voltage when an electron (minimum ionizing) traverses the counter; Curve 2 gives the same data, but for a much more heavily ionizing particle. Thus the ordinate is proportional to the pulse height of the signal that will appear after the coupling capacitor C (in Fig. 8.12).

Referring to Fig. 8.13, we see the following regions of operation of a gaseous counter: in region II the voltage is large enough to collect all the electron–ion pairs, yet not so large as to produce any multiplication. A detector operated in this region is called an *ionization chamber*. As the voltage is further raised, region III is reached, where multiplication of the original free charges takes place through the interaction of the electrons as they move through the gas toward the collecting electrode. However, over

[22] It is also, of course, a function of the specific gas or mixture of gases used.

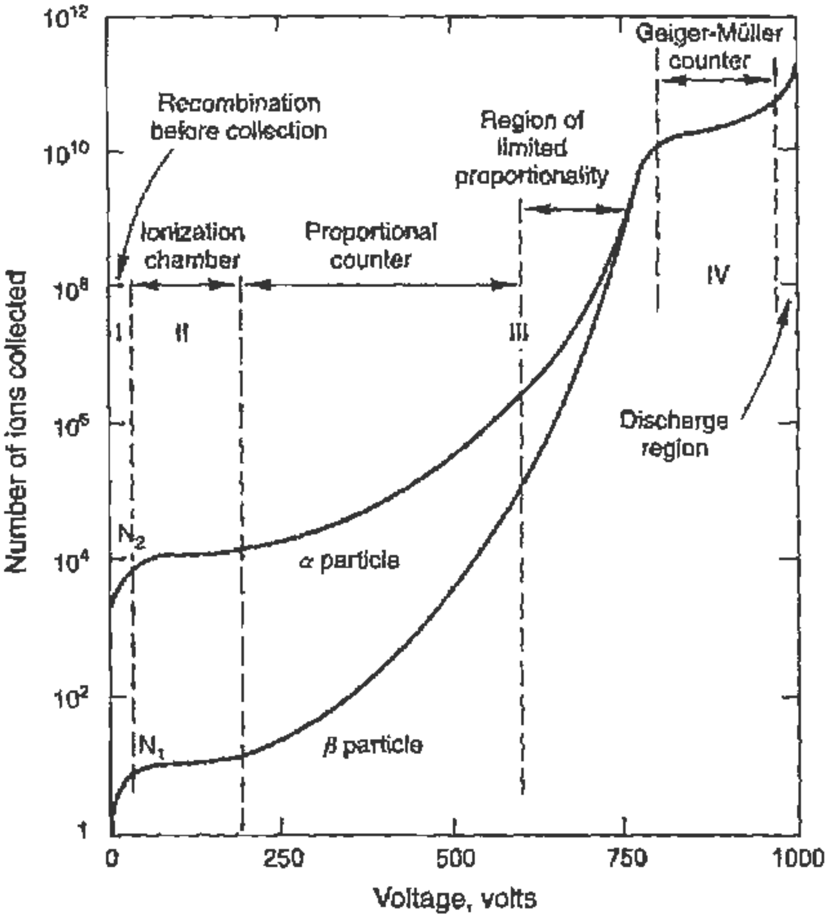


FIGURE 8.13 The number of electron–ion pairs collected when a charged particle traverses a gaseous counter of average size plotted against the voltage applied between the electrodes. Curve 1 is for a minimum ionizing particle, whereas curve 2 refers to a heavily ionizing particle. Note the three possible regions of operation as (a) an ionization counter, (b) a proportional counter, and (c) a Geiger counter.

a considerable range of voltage, the total number of collected electron–ion pairs is fairly *proportional* to the original ionization caused by the traversal of the charged particle.[23] A detector operated in this region is called a *proportional counter*; it has an advantage over the ionization counter in that t signals are much stronger, achievable gains being on the order of 10^2 to 1(Finally, further increase of the high voltage leads to region IV, where ve large multiplications are observed, and where the number of collec electron–ion pairs is independent of the original ionization. This is region of the *Geiger–Müller counter*, which has the great advantage c very large output pulse, so that its operation is simple and reliable. Inde

[23] The proportionality does not have to be a linear function of the applied voltage.

at such high voltages, once a few electron–ion pairs are formed the electrons produce more ionization at such a rapid rate that regenerative action sets in, the whole gas becomes ionized, and a discharge takes place. At that point, the resistance between the central electrode and the chamber wall becomes negligible, and the counter acts as a switch that has been closed between the high-voltage source and ground; this discharges capacitor C through resistor R_L (Fig. 8.12). Since C was charged at B^+ (on the order of 1000 V), very large output signals may be obtained. For example, if the number of electron–ion pairs collected is 10^{10} (as given by Fig. 8.13) and $C = 0.001\ \mu\text{F}$, we obtain

$$V = \frac{Q}{C} = \frac{1.6 \times 10^{-19} \times 10^{10}}{10^{-9}} = 1.6\ \text{V}. \tag{8.37}$$

By scaling this result according to the graphs in the figure, it is easy to appreciate the difficulties involved in the amplification of proportional-counter and ionization-counter signals.

The disadvantages of the Geiger counter are the loss of all information on the ionizing power of the charged particle that traversed the counter, and the long time necessary for restoring the gas to its neutral state after a discharge has taken place. However, the simplicity and good efficiency of the device for single-particle detection have made it a very common nuclear radiation detector.

8.3.2. The Ionization Chamber

The main difficulty with ionization counters is their very low signal output. If they are used, however, in an intense flux of radiation as an integrating device, high signal levels can be reached; in that case the output signal corresponds to the total number of electron–ion pairs formed (per unit time) by the radiation. In this fashion ionization chambers are frequently used for monitoring X-ray radiation or high levels of radioactivity; in such applications they are far superior to Geiger counters, since the rates are so high that a Geiger would be completely jammed.

When an absolute measurement of the created free charges is made, as with an electrometer, ionization chambers may also serve as standards of ionizing radiation. Most commercial instruments, however, amplify the output pulse and are directly calibrated in roentgens (or fractions of roentgens) per hour. For use in the laboratory an ionization counter

Model 2526 ("cutie-pie") manufactured by the Nuclear-Chicago Company is suggested for radioactivity surveys and as an X-ray monitor in the range of 0–2500 mR/h.

Below we describe a very rudimentary "student-type" ionization chamber that was used in this laboratory for measuring the range of alpha particles emitted by ^{210}Po. Figure 8.14 is a sketch of the apparatus; it consists of a flask with a 5-in. outer diameter, its inside wall having been coated with a conducting material (such as aqua-dag or silver). A rubber stopper inserted at the mouth of the flask acts as a support, electrical insulator, and vacuum lock. Through the stopper is fastened a brass rod at the tip of which has been attached a 20-μCi ^{210}Po source,[24] which is thus located at the center of the flask. A 180-V battery is connected between the flask walls and the rod supporting the source, and the ionization current is measured with a Keithley electrometer.

The energy of the ^{210}Po alpha rays is 5.25 MeV, and their range in air at stp is 3.93 cm; hence the alphas stop before reaching the walls of the flask and deposit all their energy in the gas. By using the number of approximately 30 eV per electron–ion pair, mentioned at the beginning of Section 8.3.1, we would expect per alpha particle a total of

$$5.25 \times 10^6/30 = 170{,}000 \text{ electron–ion pairs}$$

(the true number in this case being closer to 110,000).

Since

$$20\ \mu\text{Ci} = 20 \times 10^{-6} \times 3.7 \times 10^{10} = 7.4 \times 10^5 \text{ alpha particles/s}$$

if all the electrons were collected, the ionization current should be

$$I = 1.6 \times 10^{-19} \times 7.4 \times 10^5 \times 1.1 \times 10^5 = 1.3 \times 10^{-8}\ \text{A}, \qquad (8.38)$$

which is readily measurable.

If now the flask is slowly evacuated, the alpha particles will traverse a longer path before stopping; however, as long as the alphas stop in the gas, the same number of electron–ion pairs is formed and the ionization current should remain flat and independent of pressure. When the density of the air in the flask becomes so low that the alphas reach the wall *before* losing *all* their energy in the gas, fewer electron–ion pairs are formed and the ionization current will drop monotonically with decreasing pressure.

[24] See Appendix D.

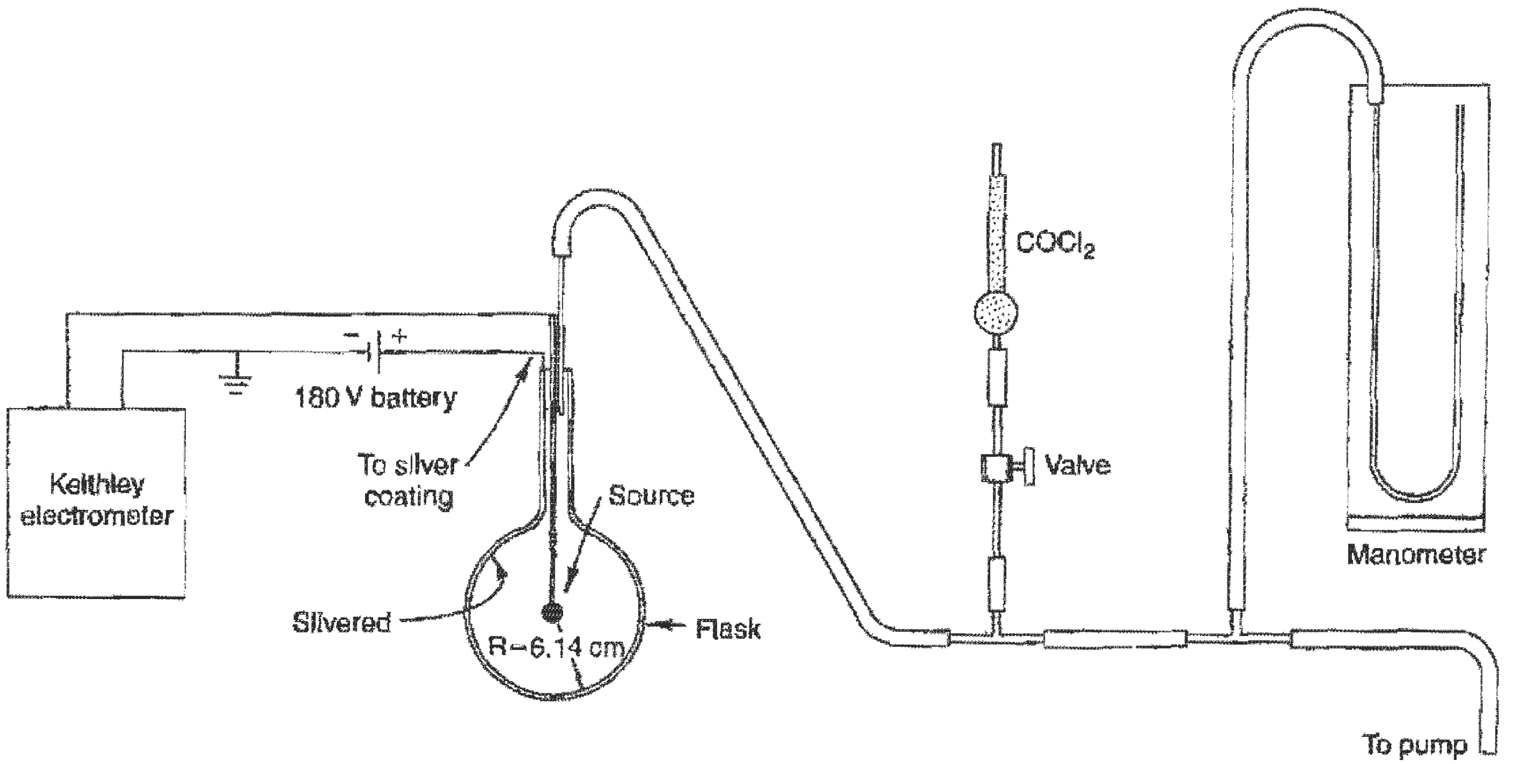


FIGURE 8.14 A simple arrangement for the determination of the range of alpha particles in air by measuring the ionization current as a function of chamber pressure.

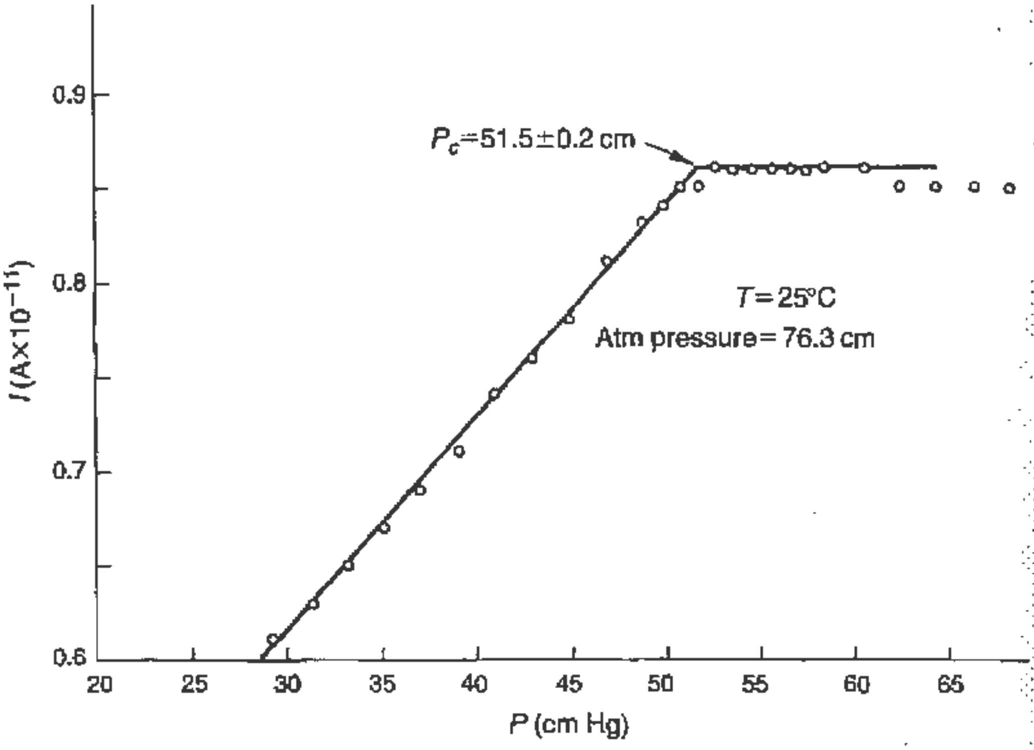


FIGURE 8.15 The results of the measurement referred to in Fig. 8.14. The ionization current is plotted against residual air pressure and a decrease in current begins at $P =$ 51.5 cm Hg. This corresponds to a range of 4.02 cm in air at stp.

Data obtained in this fashion by a student are shown in Fig. 8.15. Indeed, the expected qualitative behavior of the ionization current is observed; from the breaking point we conclude that at a pressure of 51.5 ± 1 cm Hg, the range of ^{210}Po alpha rays in air is $R = 6.14$ cm. Hence at stp (760 mm Hg, 15°C)

$$R_{\text{stp}} = R \times \frac{P}{P_{\text{stp}}} \times \frac{T_{\text{stp}}}{T} = 6.14 \times \frac{51.5}{76.0} \times \frac{288}{298} = 4.02 \pm 0.1 \text{ cm}$$

in good agreement with the accepted value of $R_{\text{stp}} = 3.93$ cm.

From the ordinate of Fig. 8.15 we note, however, that the ionization current is three orders of magnitude lower than the estimate given by Eq. (8.38); this is due to the recombination of the electron–ion pairs, which proceeds at a fast rate because of the long path in the air, the high pressure, and the low value of the electric field accelerating the electrons.[25] This is an example of a low-efficiency integrating ionization chamber.

[25] Some loss is also due to self-absorption in the source, and the geometrical solid angle is only 2π.

8.3.3. The Proportional Counter

We will not describe in detail the proportional counter[26] but only give the results obtained by a student using such a detector in connection with the experiment on the Mössbauer effect (see Chapter 9). The advantage of proportional counters lies in the detection of very low-energy X-rays or gamma rays, which can hardly penetrate a scintillation crystal, and when in addition good energy resolution is required. This is the case in the Mössbauer experiment from ^{57}Fe, where it is necessary to identify a 14.4-keV gamma ray in a strong background of 123-keV gamma rays and 5-keV X-rays.

The proportional counter used was[27] Amperex type 300-PC. It was filled with a xenon methane mixture at a pressure of 38 cm Hg. The equipment used for amplification and pulse-height measurement[28] is shown in Fig. 8.16, and the counter was operated at 2100 V. Figure 8.17 gives the results obtained, where the number of pulses is plotted against the discriminator channel. The large peak at Channel 12 is the 5-keV X-ray; the small peak at Channel 26 represents the sought-after 14.4-keV gamma ray.

As we know from Section 8.2.5 (Fig. 8.7) the predominant interaction of low-energy gamma rays in the gas is the photoelectric effect. The cross section for 5-keV quanta is on the order of 6×10^{-24} cm^2, so that if the counter represents approximately 50 mg/cm^2 of material, the efficiency for gamma-ray detection might be as high as

$$6 \times 10^{23} \times 50 \times 10^{-3} \times \frac{Z}{A} \times 6 \times 10^{-24} \approx 10\%.$$

Using the data from the 5-keV peak, we obtain for the resolution of this proportional counter,

$$\Delta E/E = 1.7/12 = 14\%,$$

where for ΔE we chose the half-width of the peak at half-maximum (after background subtraction).

[26] For an extensive discussion of proportional and ionization counters, see the *Encyclopedia of Physics*, Vol. 45, *Nuclear Instrumentation II*, Springer-Verlag, Berlin, 1958, articles by H. W. Fulbright, pp. 1–50, and by S. C. Curran, pp. 174–221.

[27] Manufactured by the Amperex Corporation and obtainable from Scientific Sales, Inc., Long Island, N.Y.

[28] For a more detailed discussion of pulse-height spectra see Section 8.4.

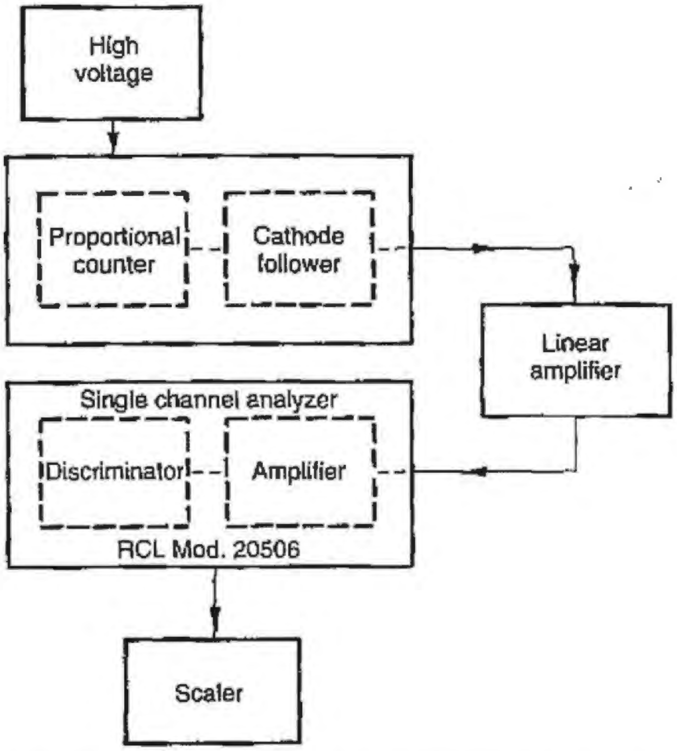


FIGURE 8.16 Block diagram for pulse-height measurements using a proportional counter.

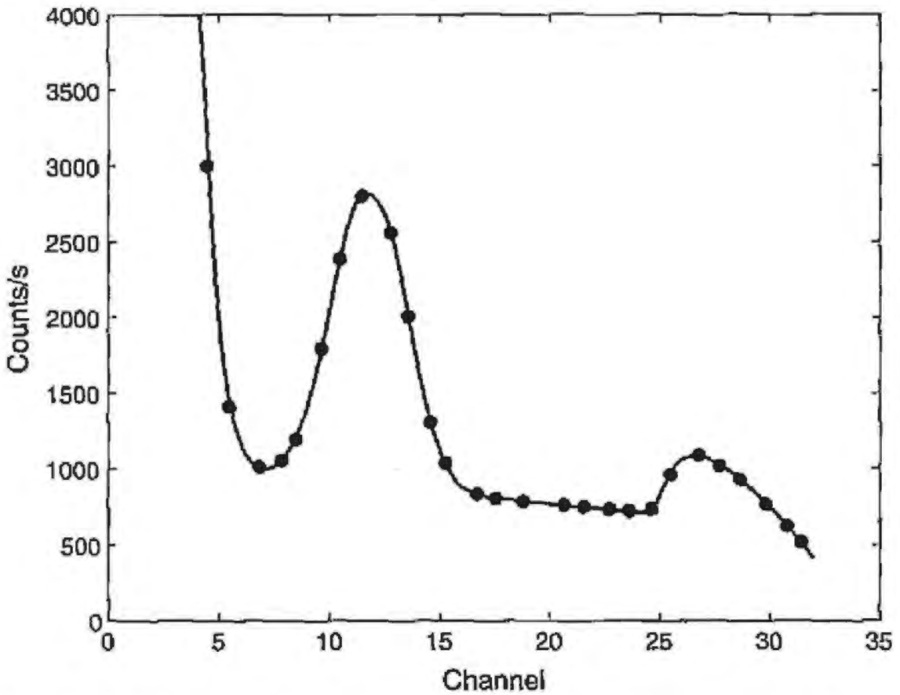


FIGURE 8.17 Pulse-height spectrum of the low-energy gamma radiation from ^{57}Fe as obtained with a commercial proportional counter. The pronounced peak at channel 12 is the 5-keV X-ray while the smaller peak at channel 26 is the 14.4-keV gamma-ray line used in the Mössbauer effect.

8.3.4. The Geiger Counter; Plateau and Dead Time

It has been pointed out in Section 8.3.1 that a gaseous counter operates in the Geiger region when the voltage between electrodes is sufficiently large; that is, the traversal of a charged particle initiates a discharge in the gas, and as a result a pulse appears at the output that is independent of the original ionization. If the voltage is further increased, spontaneous discharges occur, making the device useless as a particle detector.

Because the principle of operation is simple, Geiger counters are simply constructed, the geometry of Fig. 8.12 being typical. For certain applications, the thickness of the walls is an important consideration, and Geiger counters may be built with special thin windows (usually mica of few mg/cm^2). Glass envelopes for Geiger counters are fairly common, and various pressures as well as mixtures of gases are used.

Another important consideration for Geiger counters is the "quenching" of the discharge initiated by the traversal of a charged particle. Until the gas is returned to its neutral state, the passage of a charged particle will not produce an output pulse; this is the period of time during which the counter is "dead." The quenching of the discharge can be achieved through the external circuit (for example, in Fig. 8.12 the charging resistor R_C will introduce such a voltage drop that the discharge will extinguish itself), through the addition of special impurities (such as alcohol) to the gas of the counter, or by both methods used together. The circuitry necessary for the operation of a Geiger counter is also extremely simple. A single stage of amplification and pulse shaping is usually sufficient to drive any scaler.

In order to operate a Geiger counter properly, the high-voltage source must be set in the "plateau" region (Fig. 8.13, region IV), where a similar output is consistently obtained for all charged particles traversing the counter. We may then define the *efficiency* of the detector as the ratio of the number of output pulses over the total flux traversing the counter; since the pulse heights are all equal in the plateau region, we do expect the efficiency to remain constant in that same region. Clearly any particle detector should be operated in a region where the efficiency is "flat" with respect to variation of operating parameters. The efficiency of Geiger counters is 90% or higher for charged particles, but for photons it is much lower, being only on the order of 1–2%.

It is difficult to make absolute efficiency measurements for Geiger counters. A "standard" calibrated source of radioactive material may be used, and the output count compared with the expected flux from a knowledge of the solid angle subtended by the Geiger counter. If the counter is placed at several distances from the source, the consistency of the measurements may also be checked through the $1/r^2$ dependence. However, a *relative* measurement of the efficiency as a function of the high voltage is easy to make; if it yields a *flat plateau*, this is an indication that the detector operates at high efficiency (close to 100%) for the particular type of radiation that is incident. Geiger-counter plateaus are usually a few hundred volts wide and have a small slope, on the order of 1–2% per 100 V.

To determine the plateau, either a radioactive source or the cosmic-ray flux may be used; since this flux is on the order of 10^{-2} particles/cm^2-s, it takes several minutes to accumulate 1000 counts for a counter of average size. As explained in Chapter 10, the emission of radiation is a random process, so that the standard deviation[29] of any measurement is given by the square root of the number of counts, and thus the measurement should be interpreted as

$$1000 \pm 31 = 1000 \times (1 \pm 0.03) \text{ counts}$$

or in common parlance, 1000 counts give 3% statistics. The high voltage should be well stabilized, usually to a few parts in one thousand.

Figure 8.18 gives the plateau found by a student for the RCL[30] type 10104 Geiger counter. A 10-μCi ^{60}Co source was used for the measurements, and the standard deviation at each point is shown by the size of the dot. The plateau begins at 1100 V and is approximately 250 V wide; the discharge region begins at 1400 V.

The slopc of the plateau, from Fig. 8.18, is

$$150/3200 = 5\% \text{ per } 100 \text{ V}.$$

Next we turn our attention to the dead time of the Geiger counter already mentioned. Indeed, once a discharge has been initiated, the counter will not register another pulse unless the discharge bas extinguished itself, and until, in addition, thc counter has "recovercd"—that is, returned to a neutral state. During the rccovery period, the counter will generate an output pulse, but of a smaller-than-normal amplitude depending on the stage of recovery.

[29] If this measurement is repeated many times, in 68% of the cases we will obtain $\overline{N} - \sigma > N > \overline{N} + \sigma$, where $\overline{N}$ is the average of all measurements. See Chapter 10 for the definition of σ.

[30] Radiation Counter Laboratories, Inc., 512 West Grove Street, Skokie, Ill.

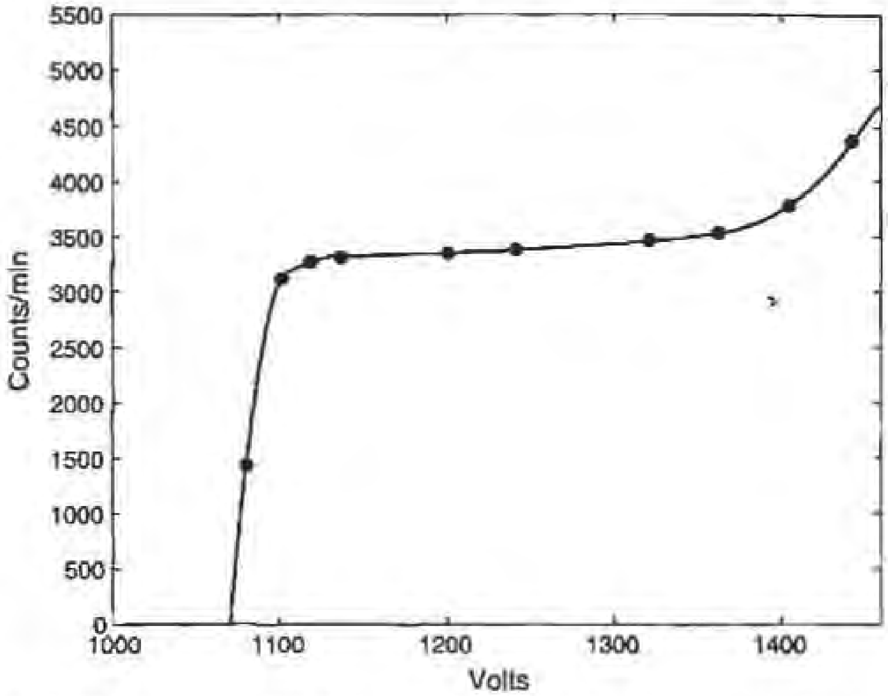


FIGURE 8.18 Plateau curve of a Geiger counter. Note that the plateau region extends for 250 V and has a slope of the order of 5% per 100 V.

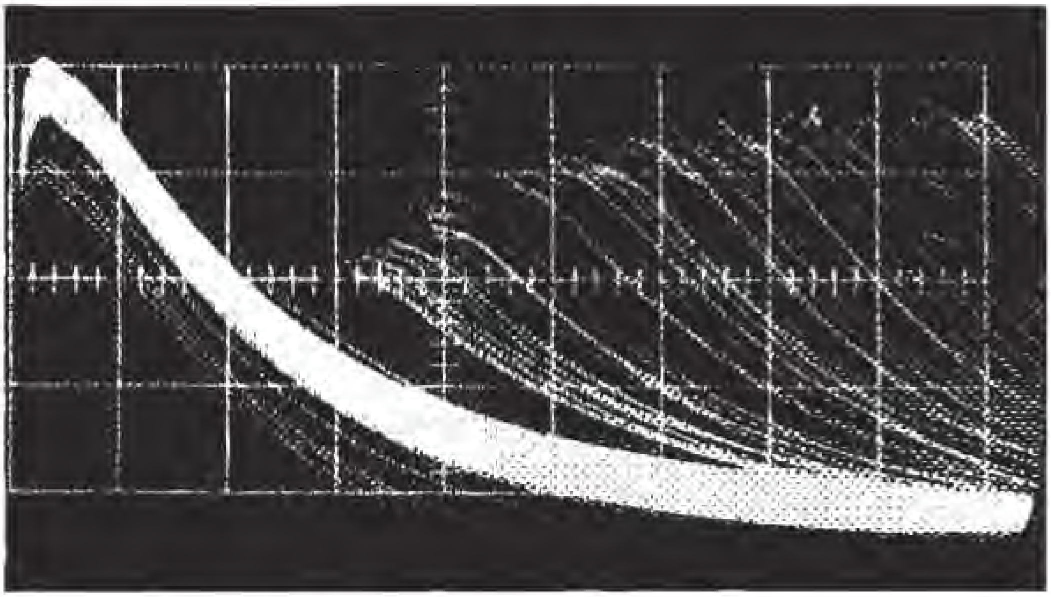


FIGURE 8.19 Multiple-exposure photograph of oscilloscope traces obtained from a Geiger counter exposed to a high flux of radiation. Note the effect of the "dead time" of the counter and the gradual buildup (recovery) of the output pulses.

This phenomenon of recovery can be clearly seen in Fig. 8.19, obtained by a student. The Geiger counter was exposed to a high flux of radiation; the trace of an oscilloscope is triggered when the output pulse appears. The horizontal scale is 100 μs/cm so that the shape of the output pulse and its exponentially decaying tail can be seen in detail. If now a second particle arrives within 1 ms of the previous one, it will appear on the same oscilloscope trace since the scope will not trigger again until the sweep is completed (the screen is 10 cm wide). The picture shown in Fig. 8.19 was obtained by making a multiple exposure of such traces. The correlation of pulse height against delay in arrival time and the exponential dependence of the recovery are clearly noticeable. If we consider that the counter is inoperative until the output is restored to 63% of its original value $(1-1/e)$, the data of Fig. 8.19 give a value for the dead time τ on the order of

$$\tau = 400\ \mu\text{s}. \tag{8.39}$$

Pulses, however, seem to appear after an interval

$$\tau \approx 300\ \mu\text{s}. \tag{8.40}$$

The dead time of a counter may also be obtained by an "operational" technique, such as by measuring the counting loss when the detector is subjected to high flux. If the dead time is τ (s), and the counting rate R (counts/s), the detector is inoperative for a fraction $R\tau$ of a second; the true counting efficiency is then $1 - R\tau$.

Consider two sources S_1 and S_2, which when placed at distances from the counter D_1 and D_2 give a true rate (counts/s) R_1, R_2. The counter, however, registers rates $R_1' < R_1$, $R_2' < R_2$ due to dead-time losses, and when both sources are simultaneously present, it registers $R_{12}' < R_1' + R_2'$ due to the additional loss accompanying the higher flux. Now,

$$\begin{aligned} R_1' &= R_1(1 - R_1'\tau) \\ R_2' &= R_2(1 - R_2'\tau) \\ R_{12}' &= (R_1 + R_2)(1 - R_{12}'\tau). \end{aligned}$$

We solve by writing

$$\frac{R_{12}'}{1 - R_{12}'\tau} = \frac{R_1'}{1 - R_1'\tau} + \frac{R_2'}{1 - R_2'\tau},$$

which reduces to a quadratic equation in τ with the solution

$$\tau = \frac{1 \pm \sqrt{1 - R'_{12}(R'_1 + R'_2 - R'_{12})/R'_1 R'_2}}{R'_{12}}.$$

This can be expanded in the small quantity $(R'_1 + R'_2 - R'_{12})$ to give the approximate expression

$$\tau \approx \frac{(R'_1 + R'_2 - R'_{12})}{2R'_1 R'_2}. \tag{8.41}$$

We now apply Eq. (8.41) to data obtained by students with the same counter used for Fig. 8.19. In practice, source S_1 is first brought to the vicinity of the counter and R'_1 is obtained, next S_2 is also brought in the area and R'_{12} is obtained, and finally S_1 is removed and R'_2 is measured: thus no uncertainties due to source position can arise. They obtain

$$R'_1 = 395 \pm 3 \text{ counts/s}$$
$$R'_{12} = 655 \pm 3 \text{ counts/s}$$
$$R'_2 = 334 \pm 3 \text{ counts/s},$$

yielding $\tau = 282 \pm 20$ μs, in better agreement with Eq. (8.40) than with Eq. (8.39).

The rather long dead time of the Geiger counter is a serious limitation restricting its use when high counting rates are involved; the ionization counter and proportional counter have dead times several orders of magnitude shorter.

8.4. THE SCINTILLATION COUNTER

8.4.1. General

As we saw, in gaseous-ionization instruments, the electron–ion pairs were directly collected; in the scintillation counter the ionization produced by the passage of a charged particle is detected by the emission of weak scintillations as the excited molecules of the detector return to the ground state. The fact that certain materials emit scintillations when traversed or struck by charged particles has been known for a long time, Rutherford being the first to use a ZnS screen in his alpha particle scattering experiments.

The scintillation counters used currently were developed in the 1950s and consist of an organic or inorganic crystal coupled to a sensitive photomultiplier that responds to the light pulses. Anthracene or stilbene crystals make excellent scintillators, but organic compounds embedded in transparent plastic, such as polystyrene, are now widely used because of ease in handling and machining and availability in large sizes. Such materials are commercially available[31] under the general description of "plastic scintillators." The active materials are compounds, such as "PPO," 2-5-diphenyloxazole, or diphenylstilbene, or others, and are also available in liquid form.

Organic scintillators have an extremely fast response, on the order of 10^{-9} s, which can be matched by good photomultipliers. On the other hand, because of the low density and low Z, their efficiency for gamma-ray conversion is not high. To detect gamma rays, inorganic crystals, such as NaI or CsI, are used instead, activated with some impurity, for instance Tl (1 part in 10^3). Inorganic crystals have an excellent efficiency for gamma-ray conversion, due to their high Z; from Eq. (8.20) we recall that the photoelectric effect is proportional to Z^5 and from Eq. (8.23) pair production is proportional to Z^2. However, the light output from inorganic crystals is spread over a much longer time interval, on the order of 10^{-6} s. Such inorganic crystals are also available commercially,[32] appropriately encased since they are damaged by humidity; they come in sizes up to several cubic inches.

The light output of scintillators is proportional (as a matter of fact, linear) to the energy lost by the particle that traverses the detector; thus, by pulse-height analyzing the electrical output of the photomultiplier, the scintillation counter may be used as a spectrometer. This procedure is discussed in detail in the following section, where it is seen that energy resolution on the order of 10% or better is achievable.

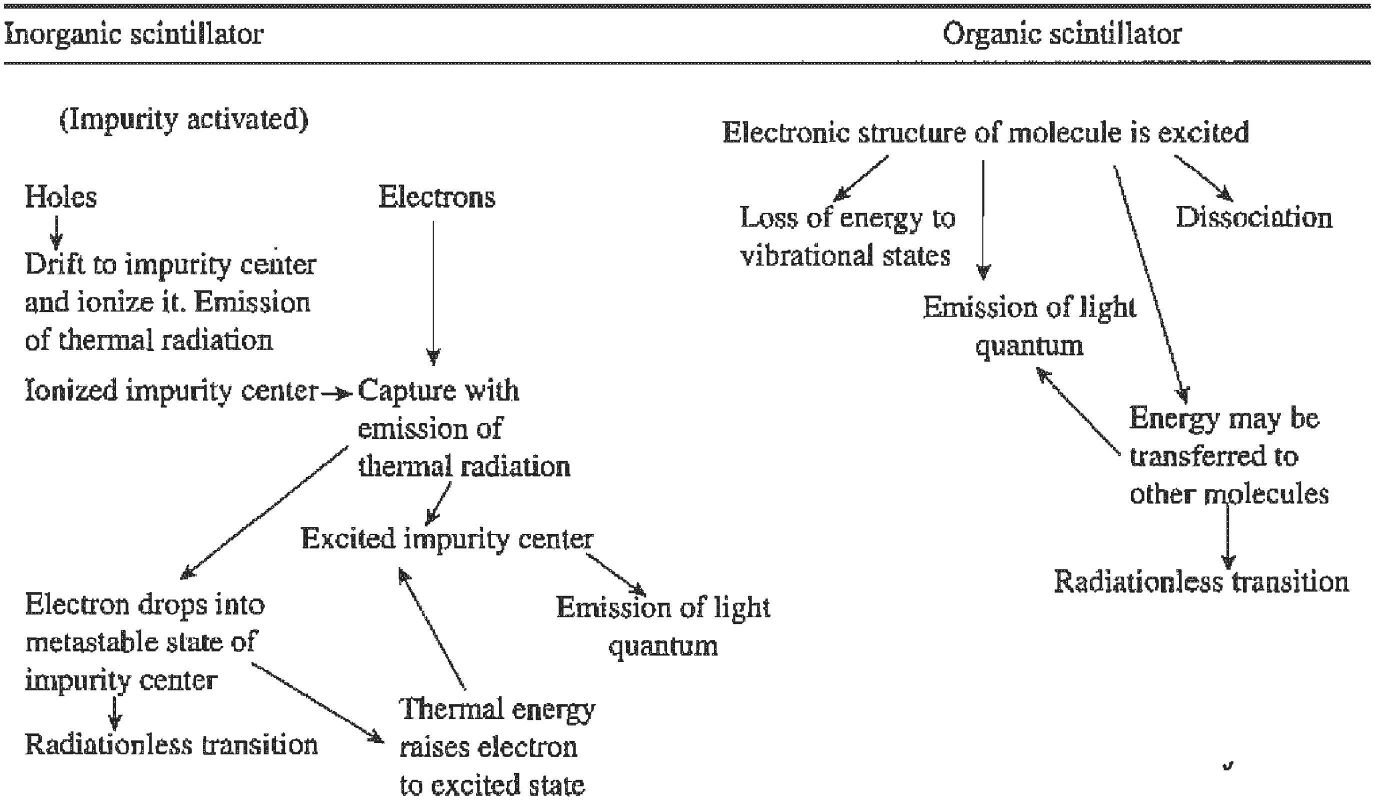
The mechanism of emission of the photons in the scintillator material is rather involved. Table 8.2 gives a chart of the processes involved in the emission of light in organic and inorganic crystals. In inorganic materials it is the migration of the electrons through the lattice (until they excite an impurity center) that is responsible for the long duration of the light pulse.

Even though the efficiency for transferring the energy lost by ionization to the photons in the visible region is on the average low, $\epsilon \approx 1.5\%$, a scintillator still provides ample light output. Consider the case of a plastic scintillator 1-cm thick, traversed by a minimum-ionizing particle: $dE/dx = 2 \times 10^6$ eV per g/cm^2; if we take the average photon energy

[31] For example, from Pilot Chemicals Inc., 36 Pleasant St., Watertown, MA.

[32] For example, from Harshaw Chemical Corp., Cleveland, OH.

TABLE 8.2 The Series of Processes Leading to the Emission of Light When a Charged Particle Traverses a Scintillator Material[a]


Inorganic scintillator
Organic scintillator
(Impurity activated)
Holes
Electrons
Drift to impurity center and ionize it. Emission of thermal radiation
Ionized impurity center
Capture with emission of thermal radiation
Excited impurity center
Emission of light quantum
Electron drops into metastable state of impurity center
Radiationless transition
Thermal energy raises electron to excited state
Electronic structure of molecule is excited
Loss of energy to vibrational states
Emission of light quantum
Dissociation
Energy may be transferred to other molecules
Radiationless transition

[a] After J. Sharpe, *Nuclear Radiation Detectors*, Methuen, London, 1965 (Courtesy of the Publishers).

as 3 eV, we obtain 10^4 photons. The efficiency of a photomultiplier cathode for converting photons into electrons is on the order of 0.1, and the geometric efficiency for collecting the photons onto the photocathode is usually high, so that on the order of 1000 electrons are released. With modern techniques, however, it is possible to detect the release of a few photoelectrons, or even of a single one.

Clearly the scintillator material must be transparent to the visible radiation and optical coupling to the photomultiplier must be provided. This is achieved either directly or through a "lightpipe," which is an appropriately shaped piece of lucite or other medium of high refractive index that traps and guides the light due to total internal reflection at its surfaces. At the surfaces where the lightpipe is joined to the scintillator or to the photomultiplier, optical contact is achieved by the use of either viscous fluids or special glues.[33] Obviously the whole assembly must be light tight; this is frequently achieved by wrapping black electrical tape around the scintillator, lightpipe, and phototube.

Because of its great stability and ease of operation, as well as because of its time and energy resolution, the scintillation counter has become the most frequently used detector in nuclear physics, especially for high-energy particles.

8.4.2. Experiment on the Determination of the Energy of Gamma Rays with a Scintillation Counter

If atoms are quantum-mechanical systems and a typical manifestation of this fact is the emission of spectral lines of light, it should be expected that nuclei, when excited, would emit similar line spectra.

Since the nuclear radius is three to five orders of magnitude smaller than that of atoms, the forces that bind the nucleus (against the repulsion of the positive charges confined in its volume) must be correspondingly stronger than the forces that bind the atomic electrons to the nucleus. As a consequence, the energy levels and the quanta of energy emitted in a nuclear transition are also orders of magnitudes larger than those of atomic transitions. Indeed, the quanta of electromagnetic radiation emitted in a nuclear transition fall in the gamma-ray region, and new techniques are needed for their detection and for the measurement of their (wavelength) energy.

[33] In the first category, Corning 200,000 centipoise fluid or clear vacuum grease; in the latter, R 363, PS 28 acrylic glue, etc.

Further, because of the larger spacing between energy levels, it is not easy to excite a nucleus from its ground state by the simple means of electric discharges or arc sources such as are used for atoms; instead, beams of neutrons or high-energy gamma rays, or high-energy charged particles, are required. However, in distinction to atomic transitions where the de-excitation probability is on the order of 10^8/s, some nuclear transitions have a very small "decay" probability, as small as 10^{-7}/s, corresponding to a lifetime of 100 days. Thus, it is possible to excite a sample of nuclei inside a nuclear reactor, or by subjecting them to cyclotron bombardment, or by other means, and subsequently bring them to the laboratory for measuring their spectrum or for other uses. Indeed, some of the nuclei that have very long lifetimes can be found in nature in their excited state; these are the naturally radioactive elements.

We now know that the appropriate detector for measurements of the energy of gamma rays is an inorganic crystal. When a gamma ray of energy <1 MeV enters the detector, it will interact either by the photoelectric effect or the Compton effect. In the former case it is fair to assume that the ejected photoelectron will deposit all its energy in the scintillator; in the Compton effect, however, the scattered photon may or may not convert in the scintillator (depending on the size and geometry of the detector).

The pulse-height spectrum for gamma rays of a given energy will consist of a peak at an energy corresponding to that of the gamma ray and a continuum below the peak, corresponding to Compton-scattered gamma rays that escaped from the crystal before totally converting. This can be seen in Fig. 8.20 and those that follow. Clearly the larger the size of the crystal, the larger the percentage of the output counts that will lie in the photopeak; thus, the gamma-ray line will become more pronounced.

Most of the data reported here were obtained with a NaI–Tl activated crystal,[34] 2 in. in diameter and 2 in. wide, coupled directly to a photomultiplier tube.[35] (Photomultiplier tubes and high-voltage bias schemes are discussed in Appendix E.2.) The output pulse is fed to an Ortec[36] Model 570 amplifier, and its output is fed to a Canberra multiport multichannel analyzer (MCA). The MCA is controlled and read out through a

[34] Bicron Corporation, http://www.bicron.com/.

[35] The crystal and photomultiplier tube assembly is a commercial package from Canberra Industries, http://www.canberra.com/, Model 802-3. The photomultiplier tube "base" was constructed from a commercial socket and simple components.

[36] http://www.ortec-online.com/.

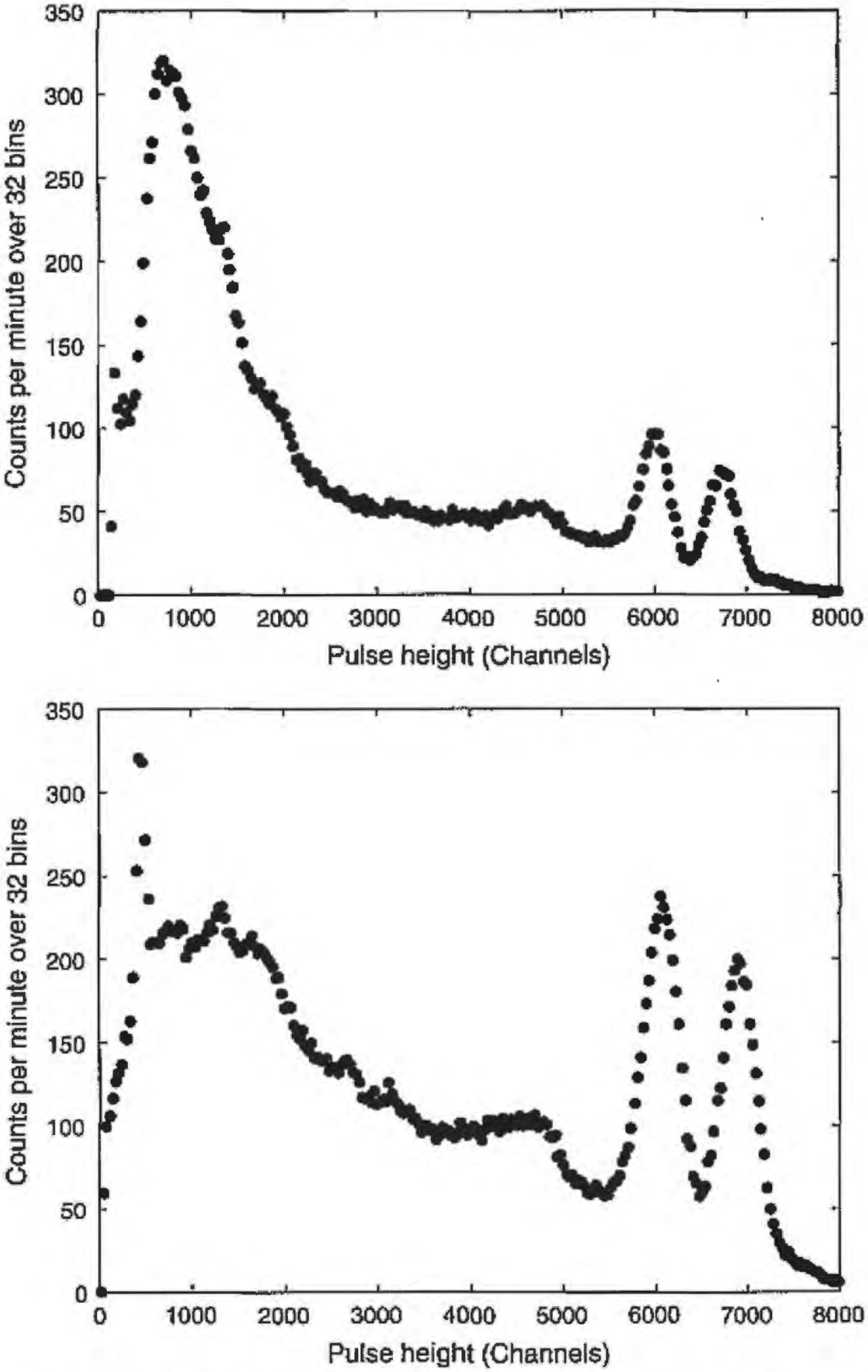


FIGURE 8.20 Pulse-height spectrum of ^{60}Co gamma rays obtained with a NaI crystal, along with the decay scheme of ^{60}Co. The upper spectrum was taken with a 2-in.-diameter crystal detector, while the bottom was taken with a 3-in. crystal. The ^{60}Co source was relatively weak (less than 1 μCi when these data were taken) and the source-to-detector distance was 10 cm. The decay scheme is also shown.

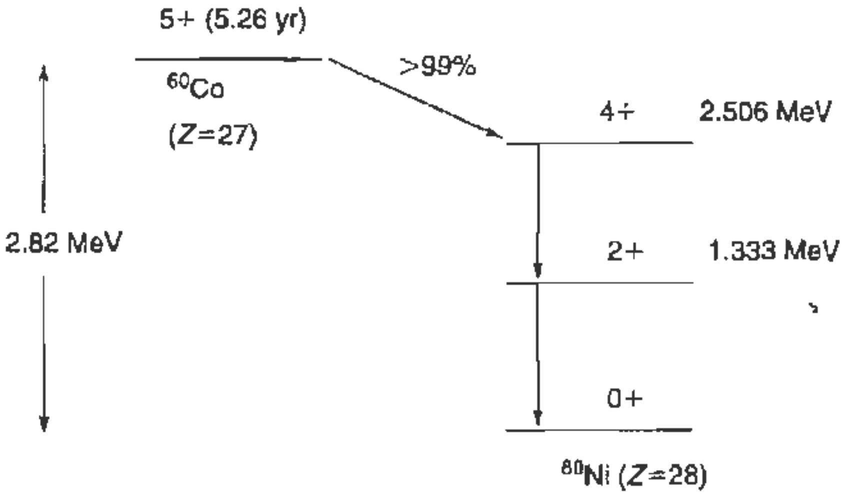


FIGURE 8.20 (*Continued*)

GPIB interface, in this case using a laptop computer. Spectra acquired in this way are histograms with $8192 = 2^{13}$ bins. (Adjacent bins were added together to reduce the statistical fluctuations from bin to bin. This is easy to do with the reshape command in MATLAB.) The conversion of bin number to photon energy depends on the combined gain of the photomultiplier and the amplifier, and must be calibrated with sources of known photon energy.

The following figures give the results obtained by a student. Figure 8.20 gives the spectrum of ^{60}Co and shows two distinct peaks, which we attribute to gamma rays emitted in the de-excitation of ^{60}Ni from its 2.505-MeV level to the 1.333-MeV level, and from that level to the ground state according to the decay scheme also shown in the figure. For comparison, we also show a spectrum taken with a 3-in.-diameter and 3-in.-wide crystal. As a measure of the energy resolution, we may consider the full-width of the peak at half-maximum, which is on the order of 480 channels, hence a resolution of $480/6000 \approx 8\%$. We also notice a significant background for pulse heights lower than that of the peaks, which is due to Compton-scattered gamma rays that subsequently escaped from the crystal. This background is much less severe for the larger crystal.

Figure 8.21 gives similar data for a sample of ^{137}Cs; here the 0.662-MeV gamma ray represents the de-excitation of ^{137}Ba. Again we notice some Compton background and an energy resolution on the order of 10%. Figures 8.22 and 8.23 give the pulse-height spectra from ^{22}Na and ^{133}Ba, respectively. For the ^{22}Na, the peak at 1.277 MeV arises from the de-excitation of ^{22}Ne; the larger peak at 0.511 MeV arises from annihilation radiation. Indeed, from the level diagram of ^{22}Na decay, we notice that

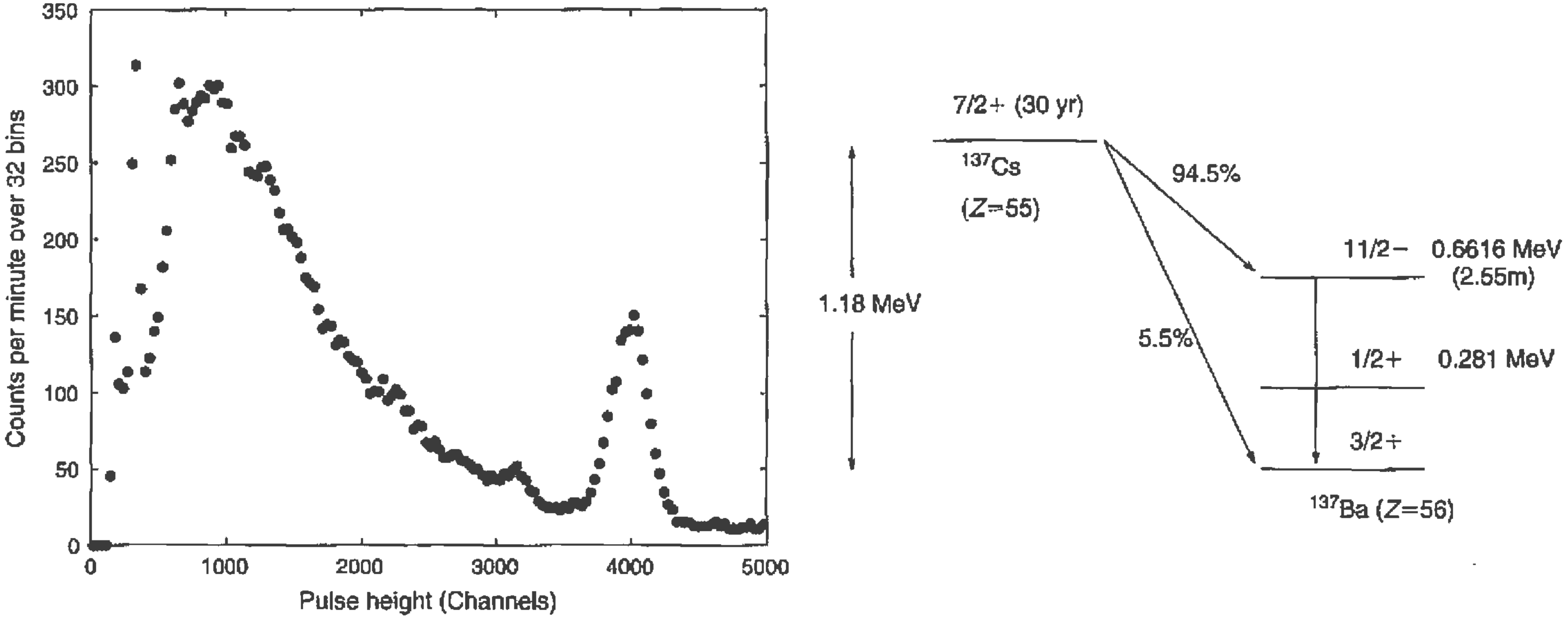


FIGURE 8.21 Pulse-height spectrum of ^{137}Cs gamma rays obtained with a NaI crystal, and the associated decay scheme.

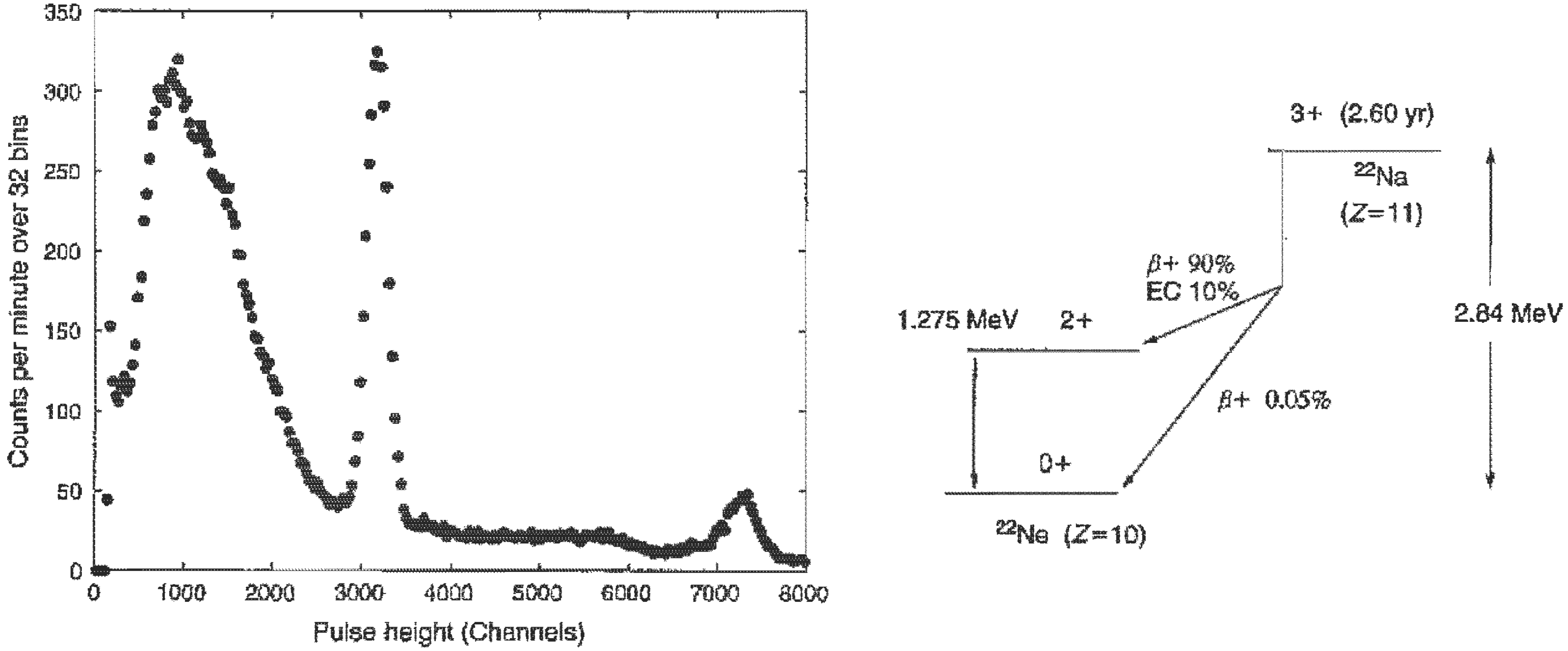


FIGURE 8.22 Pulse-height spectrum of ^{22}Na gamma rays obtained with a NaI crystal, and the associated decay scheme. Note that the 511-KeV line is due to positron annihilation.

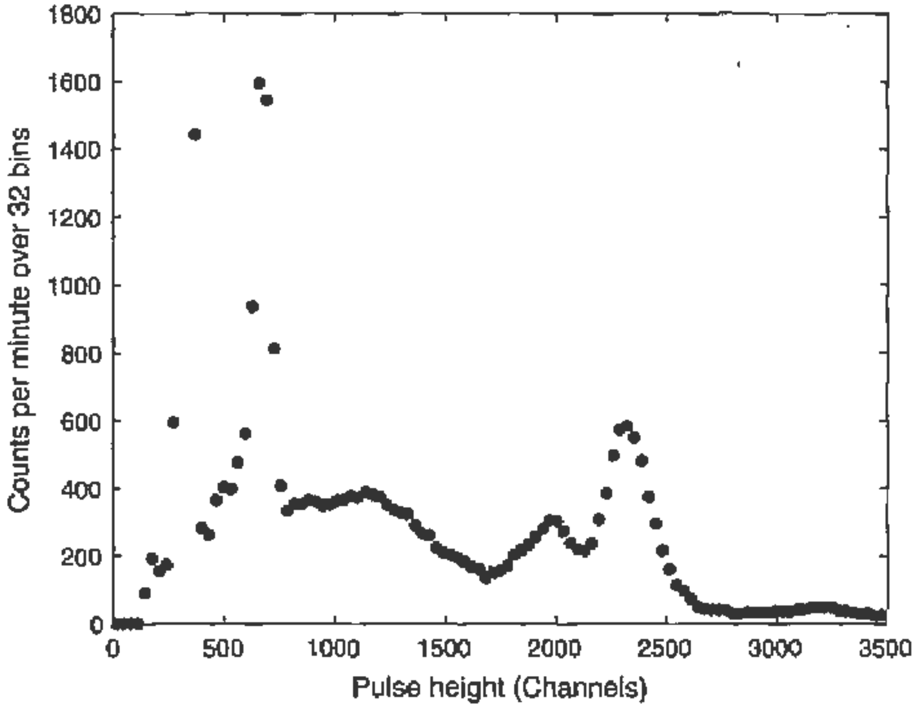


FIGURE 8.23 Pulse-height spectrum of ^{133}Ba gamma rays obtained with a NaI crystal. The decay scheme is complicated, but the most dominant γ rays at high energy are at 356 and 302 keV.

positrons are emitted; the positrons are usually stopped in the walls of the source container, or in the crystal face, and as they come close enough to an electron they annihilate into two gamma rays, each gamma ray sharing the energy of the electron–positron pair.[37] It is *one*[38] of these gamma rays that is then converted in the crystal and gives rise to the 0.511-MeV peak.

Finally, in Fig. 8.24 is given a plot of all the observed peaks against channel number, showing the linearity of pulse height against energy. (MATLAB provides a useful utility command, ginput, for interactively identifying the peak position in spectrum plots using the cursor on your computer.) In addition to the gamma rays, the nuclei investigated also emit beta rays, and one would expect to see the corresponding peaks in the pulse-height spectrum. This, however, is not true because the beta spectrum is continuous instead of being a sharp line as is the case with gamma-ray spectra; in addition, electrons may lose variable amounts of energy before reaching the scintillation crystal, so that unless special precautions are taken, the energy resolution is usually poor.

[37] See also the detailed discussion in Chapter 9.

[38] Note that they are emitted with a relative angle of 180°.

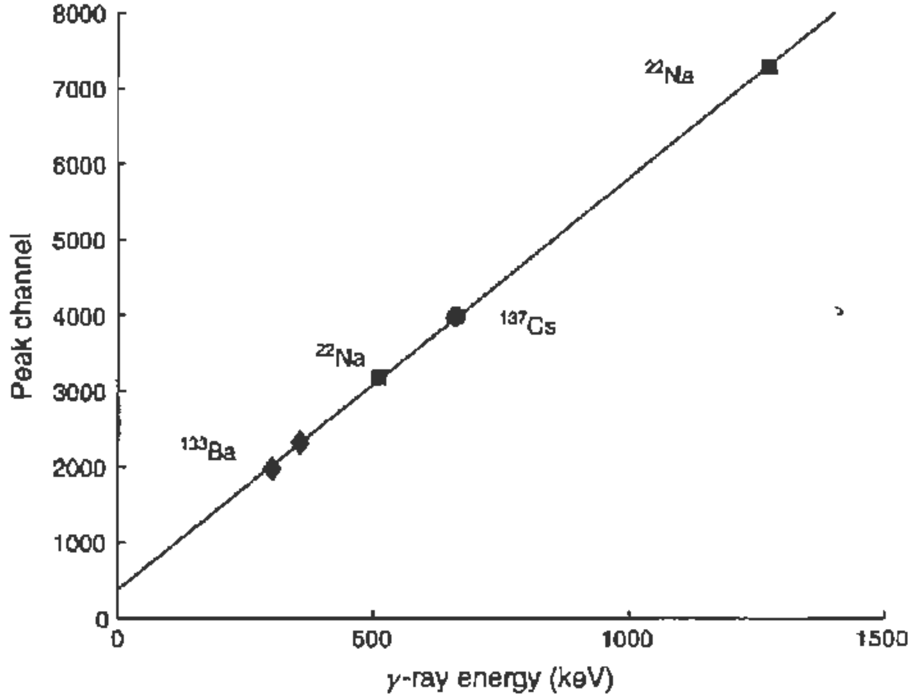


FIGURE 8.24 Plot of gamma-ray energy against the central channel of the photopeaks appearing in the spectra of Figs. 8.21 through 8.23. The detector response is obviously quite linear over this range. Note also that for a zero photon energy, there is a "pedestal" of a few hundred channels. This ensures that none of the spectrum is lost below the range of the multichannel analyzer.

In interpreting gamma-ray spectra some care must be taken since spurious peaks due to instrumental effects or physical effects do appear. First, there can be peaks arising from the emission of X-rays, following photoejection of K-shell electrons either in the source or in the shielding. Also, a peak may appear due to photons that backscatter (by 180°) in the photomultiplier window or elsewhere; then the Compton-scattered electron escapes, but the scattered photon becomes converted in the crystal. For ^{137}Cs with its 0.662-MeV gamma ray, the backscattering peak appears at 0.185 MeV and can be identified in a carefully measured spectrum.

Another spurious effect occurs when an incoming photon of energy E ejects a K-shell electron from the iodine of the crystal, but the emitted X-ray escapes without converting in the detector. The ejected photoelectron has an energy

$$E - E_K,$$

where E_K is the energy of the K shell of iodine, namely, 29 keV, and will give rise to a peak not coinciding with the true photopeak. This so-called "escape-peak" can be identified because it is located 29 keV below the photopeak; it is most pronounced in the pulse-height spectra of low-energy gamma rays.

The relative ratio of counts in the photopeak as compared to the counts in the Compton background depends on the crystal and source geometry and on the gamma-ray energy. Usually the relative counts in the photopeak give sufficient information, but when the absolute number of gamma rays is required, we must calculate the efficiency of the crystal for the particular geometry and gamma-ray energy. Extensive tables of efficiency for most combinations of the relevant parameters have been calculated.[39]

8.5. SOLID-STATE DETECTORS

8.5.1. General

We have seen how the gaseous ionization counters and the scintillation counters are widely used for the detection of radiation and charged particles. It is also possible to use semiconductor materials for the detection of charged particles, especially those of low energy; such detectors are appropriately referred to as "solid-state counters."[40]

In a general sense, we can think of this type of detector as a solid-state ionization chamber, having two basic advantages over a gas-filled ionization chamber:

(a) The energy required for the creation of an electron–ion pair is 3 eV (as compared to approximately 30 eV in a gas) so that stronger signals and better statistics can be achieved.

(b) The stopping power is approximately 10^3 times that of a gas-filled device (since the detector material is so much denser), and thus it becomes possible to stop, in the *detector*, particles with energies typical of nuclear interactions. Consequently a very large number of electron–ion pairs are formed, leading to very good energy resolution. A 1-MeV proton stopping in a solid-state detector will create 300,000 electron–ion pairs, while the same proton traversing a proportional counter of 2-cm thickness would only release approximately 30 pairs.

[39] See the *Encyclopedia of Physics*, Vol. 45, *Nuclear Instrumentation II*, p. 110.

[40] The scintillation counter is also a detector in the solid state!

In practice, however, it must be possible to collect the free charges (those created by the passage of the charged particle) before they recombine; this might be done, for example, by the application of an electric field in the detector material. This requirement is very difficult to meet with any of the ordinary crystals. Clearly, the material must have a high resistivity, since otherwise current will flow under the influence of the field, masking the effect of the pulse produced by the passage of the particle; on the other hand, in high-resistivity materials, the mobility of the free carriers is very low and the recombination probability high.

Even though some results have been obtained by using diamond as a detector, semiconductor materials come much closer to fulfilling the requirements mentioned above. Very pure material (an intrinsic semiconductor) is used to achieve the necessary high resistivity, on the order of 10^7 Ω-cm, and the detector is operated at low temperatures. Such devices are called "bulk semiconductor detectors."

A great improvement occurs when a semiconductor junction[41] is used as the detector volume; a device of this kind is called a barrier-layer detector. The junction is made by either of the following methods:

(a) Diffusing a high concentration of *donor* impurities on a p-type material, usually silicon, thus creating an n–p junction.

(b) Utilizing a thin p-type surface formed by oxidizing n-type silicon or germanium when it is exposed to air. This surface is so thin that it is usually coated with gold to provide a good electrical contact; thus we have a p–n junction.

In either case the operation is similar, but the junction is always reverse biased.

Below we will briefly discuss the diffused junction (n–p) type of detector; Fig. 8.25a is a reproduction of Fig. 2.20, and gives the configuration of the energy bands at an n–p junction, electrons being the majority carriers in the left, or n, region, and holes the majority carriers in the right, or p, region. Electrons may not move to the right, since the conduction band is at a higher (negative) potential, and holes may not move to the left, since the valence band is now at a higher (positive) potential; as a consequence there is some *repulsion* of *majority* carriers from the junction; Fig. 8.25b shows their density distribution. We note a "depletion zone" in the region marked $S-T$.

[41] Semiconductor junctions were discussed in 2.4.2, and the reader may find it useful to review that material.

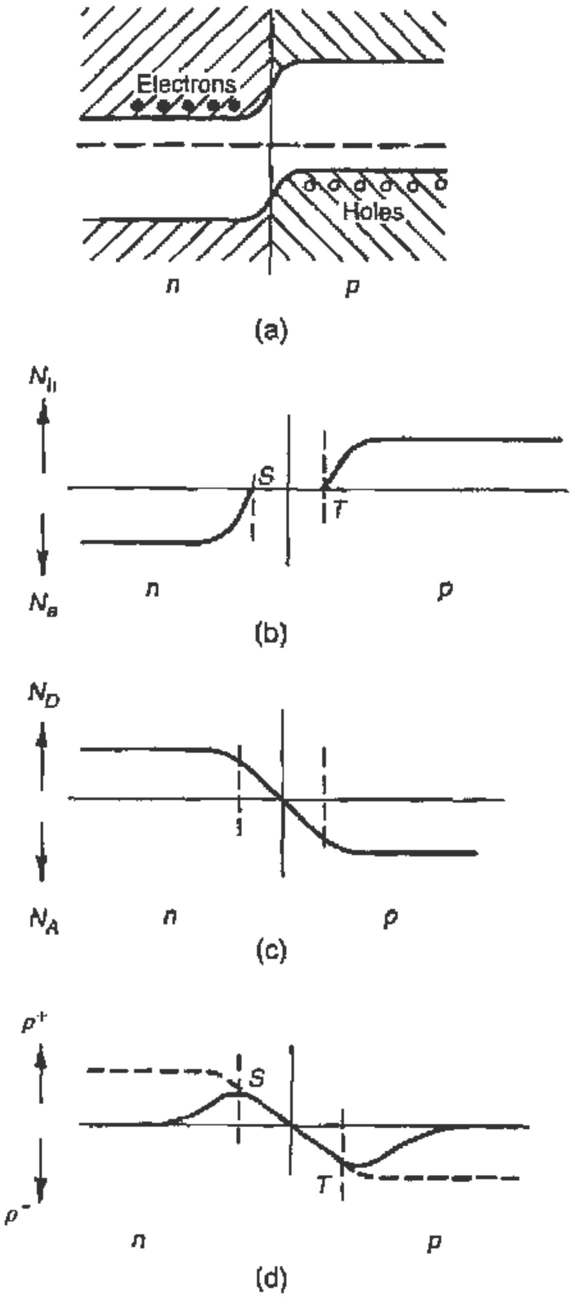


FIGURE 8.25 The *n–p* semiconductor junction. (a) Position of conduction and valence bands and of the Fermi level across the junction; note the majority carriers for each region. (b) Density distribution of majority carriers on the two sides of the junction. (c) Density distribution of impurity centers on the two sides of the junction. (d) Distribution of space charge on the two sides of the junction.

Next, Fig. 8.25c shows the density of impurity centers on the two sides of the junction; that is, these centers which may be *expected* to be ionized by the passage of a charged particle. To the left the donors have given electrons to the conduction band and are left positive; to the right the acceptors have acquired electrons from the valence band and are left negative. However, these impurity centers are neutralized by the majority carriers, so that the free (space) charge distribution is the sum of Figs. 8.25b and 8.25c, as shown in Fig. 8.25d.

Thus we see that space charge exists in the region of the junction, and as a consequence an electric field (the so-called barrier) exists as well, and extends over the depletion zone. If an electron–ion pair is created in the *depletion zone*, the electric field is such as to accelerate the negative charge toward the *n* region, where it will have high mobility (being a majority carrier); similarly, the hole will be accelerated toward the *p* region. Thus good collection efficiency is achieved.

Figure 8.26 shows the same junction under reverse bias, 8.26a being the same as Fig. 2.21. Figure 8.26b gives, as before, the density distribution of majority carriers, which are now further removed from the junction, and Fig. 8.26c is exactly the same as 8.25c, giving the density of impurity centers. Figure 8.26d, however, which gives the space-charge distribution, shows that the *ionized* impurity centers have reached saturation and extend beyond the junction. Thus, most of the applied bias voltage appears across the depletion zone, which now is much more extended; the limit to this increase in *sensitive detector depth* is set by the breakdown voltage of the semiconductor material itself.

In a diffused junction, such as used for a detector, the concentration of donors in the *n*-type material is much larger (about 10^3) than the concentration of acceptors in the *p*-type material. Since the total free charge must be the same on both sides of the junction, the space-charge distribution is asymmetric, as shown in Fig. 8.27b. Figure 8.27a gives some of the physical dimensions in a realistic diffused junction; we note that most of the "sensitive volume" is in the *p*-type material.

8.5.2. Practical Considerations in Solid-State Detectors

From the previous discussion we have seen how a semiconductor junction may provide the appropriate electric field within a solid so as to collect electron–hole pairs produced by the passage of a charged particle. Multiplication such as occurs in the proportional or Geiger counter never takes place in a solid, except under special conditions ("avalanche detectors"). To achieve good resolution in a solid-state detector one must *always* collect *all* the electron–hole pairs produced. Thus the sensitive volume of

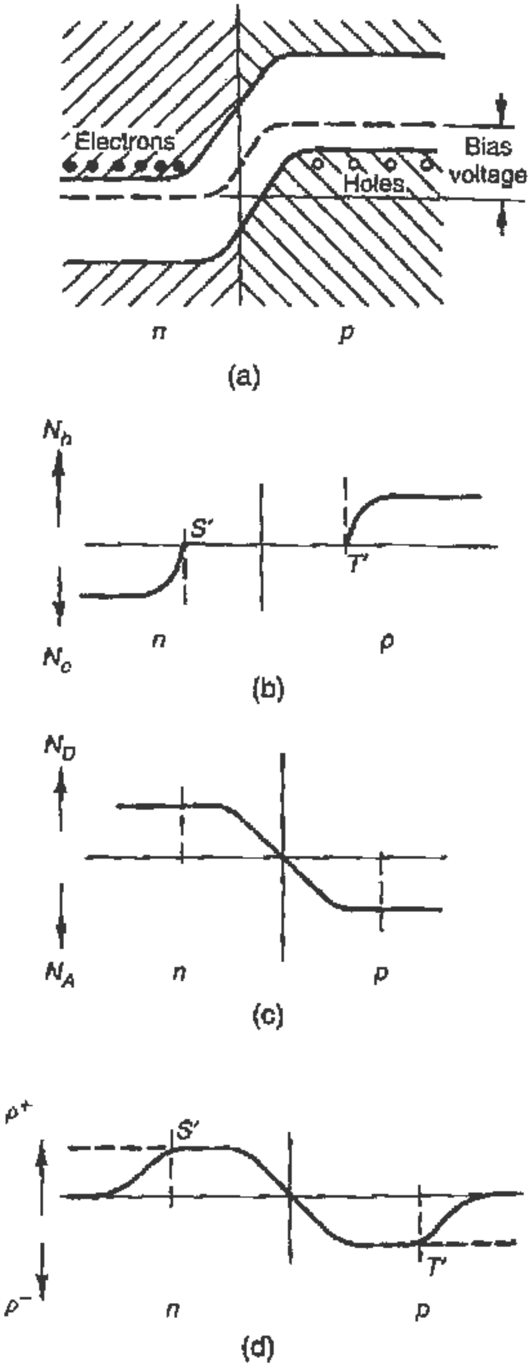


FIGURE 8.26 The *n–p* semiconductor junction under reverse bias. The plots in (a), (b), (c), and (d) pertain to the same distributions as described in the legend to Fig. 8.25 but under reverse bias. Note the increase of the "depletion zone," $S'T'$.

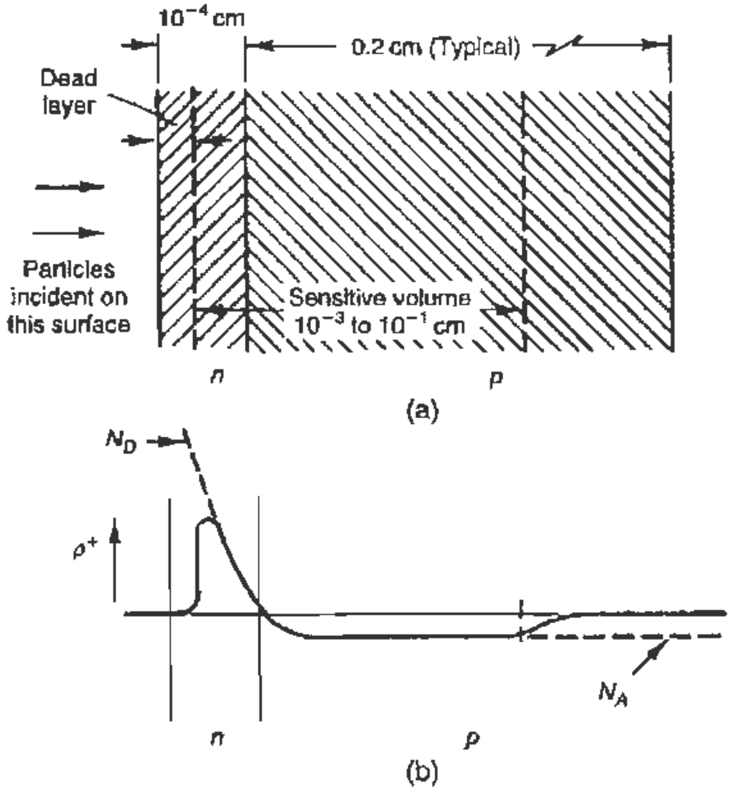

10⁻⁴ cm
0.2 cm (Typical)
Dead layer
Particles incident on this surface
Sensitive volume 10⁻³ to 10⁻¹ cm
n
p
(a)
N_D
ρ⁺
N_A
n
p
(b)

FIGURE 8.27 Arrangement of an n–p semiconductor junction for use in a solid-state detector. (a) Actual dimensions. (b) Distribution of the space charge.

the detector must be longer than the range of the particle detected; it is also desirable that the dead layer at the entrance side be as thin as possible.

Since detectors with sensitive volumes[42] of a length of 3 mm have been achieved, the use of solid-state detectors has been extended to particles of energies as high as 30 MeV. The resolution in energy is usually extremely good—that is, on the order of 0.25% for alpha particles (see also Fig. 8.31). The overall size of the detector is restricted to a few cubic centimeters, due to the available semiconductor crystals; on the other hand, small size and the absence of need for a photomultiplier are a great advantage.

It is also possible to use solid-state detectors, not as total absorption counters, but as dE/dx devices, in which case the p region is also made thin and no electrodes are placed in the path of the particle. Such detectors have been made to respond to high-energy (minimum ionizing) particles

[42] The sensitive volume or barrier depth can be obtained from a nomograph, as given by J. L. Blankenship, "Proceedings of the Seventh Scintillation Counter Symposium, Institute of Radio Engineers, NY," *Nucl. Sci.* 7, 190 (1960).

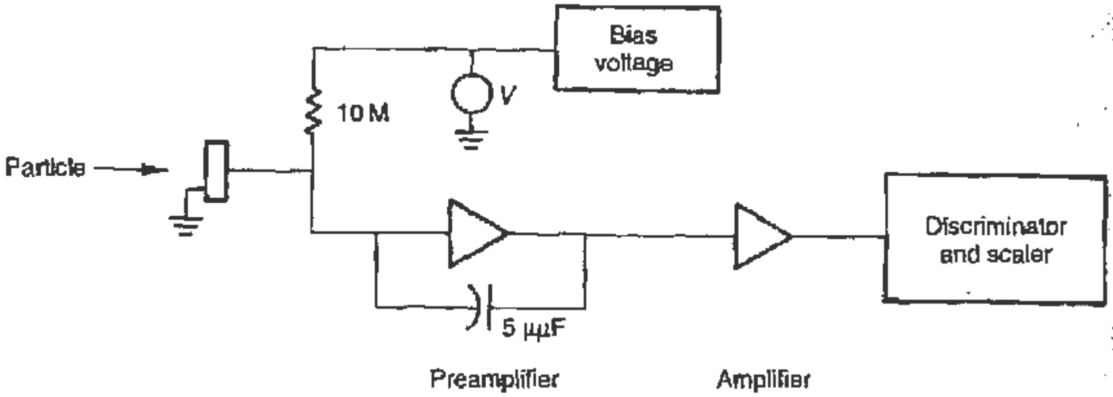


FIGURE 8.28 Typical setup for use with a solid-state detector including a feedback preamplifier.

as well. Semiconductor devices are also very useful for the detection of gamma rays. In general due to their small size, the ratio of counts in the photopeak as compared to background counts is smaller than that for a scintillation crystal; however, the resolution is excellent, reaching one part in a thousand.[43]

In practice, the construction of a solid-state detector is an art, and the attachment of electrodes to ensure good ohmic contacts may be quite difficult. When germanium is used, cooling to liquid nitrogen temperatures may be required, while silicon gives good resolution at ambient temperature. The output signals are small, the voltage being determined by the capacities of the junction and of the amplifier input; the former depends on the length of the depletion zone and the area of the detector. If we assume a typical capacity of 200 μμF, then for 1-MeV energy loss, the signal voltage is

$$V = \frac{Q}{C} = \frac{1.6 \times 10^{-19} \times (10^6/3)}{200 \times 10^{-12}} \approx 2.5 \times 10^{-4}\ \text{V}. \qquad (8.42)$$

It is necessary to use a charge-sensitive preamplifier because the capacity C depends on the applied bias; thus if voltage is directly measured, severe variations in gain occur when the bias is changed. Leakage current in the crystal and amplifier noise set the limits of the smallest detectable signals.

Most of the hardware for solid-state detectors as well as the detectors themselves are now commercially available[44]; Fig. 8.28 shows a typical setup with a feedback preamplifier. A surface-barrier silicon detector is used

[43] G. T. Ewan and A. J. Tavendale, *Can. J. Phys.* **42**, 2286 (1964).

[44] For example, from Oak Ridge Technical Enterprises, Oak Ridge, TN.

and operated at room temperature. Figure 8.31 gives the response obtained from polonium alpha particles of different energies (after attenuation in air). Another type of solid-state detector, called *p–i–n* (positive-intrinsic-negative material), consists of a layer of intrinsic crystal placed between *p*- and *n*-type material. It has the advantage of a much longer sensitive volume.

8.5.3. Range and Energy Loss of ^{210}Po Alpha Particles in Air

In Section 8.3.2 a description of the method of obtaining an estimate of the range (and hence energy) of ^{210}Po alpha particles in air, by means of a crude ionization chamber, has been given. With solid-state detectors, it is possible to improve on these measurements, as well as to study the rate of energy loss of the alpha particles as a function of their energy.

A collimated ^{210}Po source and the detector are both placed in an evacuated vessel at a fixed distance of 15 cm, as shown in Fig. 8.29. Then air is allowed into the vessel, and as a function of the pressure we measure:

(a) The number of particles counted in the detector, and

(b) The pulse-height distribution of the output signals, namely, the energy of the alpha particles when they reach the detector.

In measurement of type (a), the *same* number of alpha particles should be reaching the detector until the pressure is raised to the point where the amount of material (g/cm^2 of air) between source and detector is equal

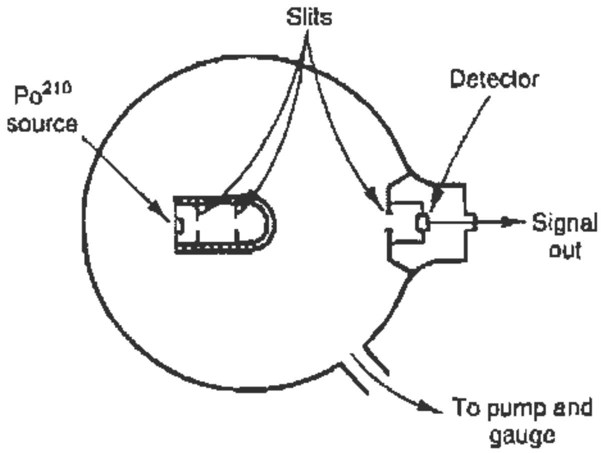


FIGURE 8.29 Arrangement for the measurement of the range in air of ^{210}Po alpha particles. Note mounting of the solid-state detector and source inside an evacuated chamber.

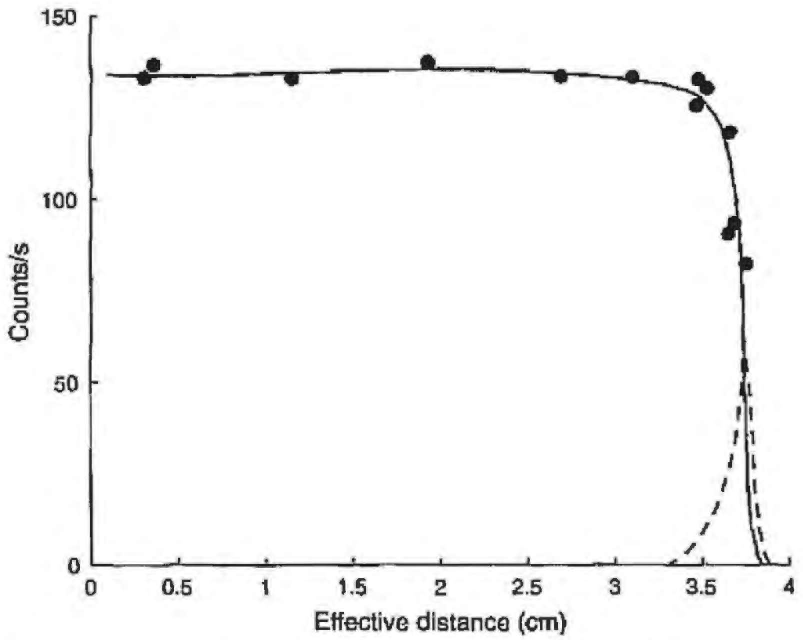


FIGURE 8.30 Data on the number of counts from a ^{210}Po alpha source reaching the solid state detector as a function of pressure in the experimental chamber. Note that the corresponding effective distance in centimeters of air at stp is indicated. The dashed curve is the derivative of the solid line; it indicates the "straggling" in the range of the alpha particles.

to the range of the alpha particles; beyond that pressure the counting rate should abruptly fall to 0. Note that since the relative position of source and detector is not altered, the solid angle $\Delta\Omega$ does not change, and the only variation arises from the increased multiple scattering; this, in turn, may result in some loss of particles from the beam.

These considerations are indeed borne out by the results obtained by a student and shown in Fig. 8.30. Here the ordinate to the left gives the counts per second while the abscissa gives the pressure of air in centimeters of mercury, or, equivalently, the effective distance of air at stp. The dashed curve to the right is the derivative with respect to distance of the counting curve and gives the range (and so-called range straggling) of ^{210}Po alpha particles. We obtain a mean range of

$$R = 3.72 \pm 0.06 \text{ cm}$$

and an extrapolated range

$$R = 3.82 \pm 0.06 \text{ cm},$$

which might indicate some systematic discrepancy from the accepted value for the extrapolated range of 3.93 cm.

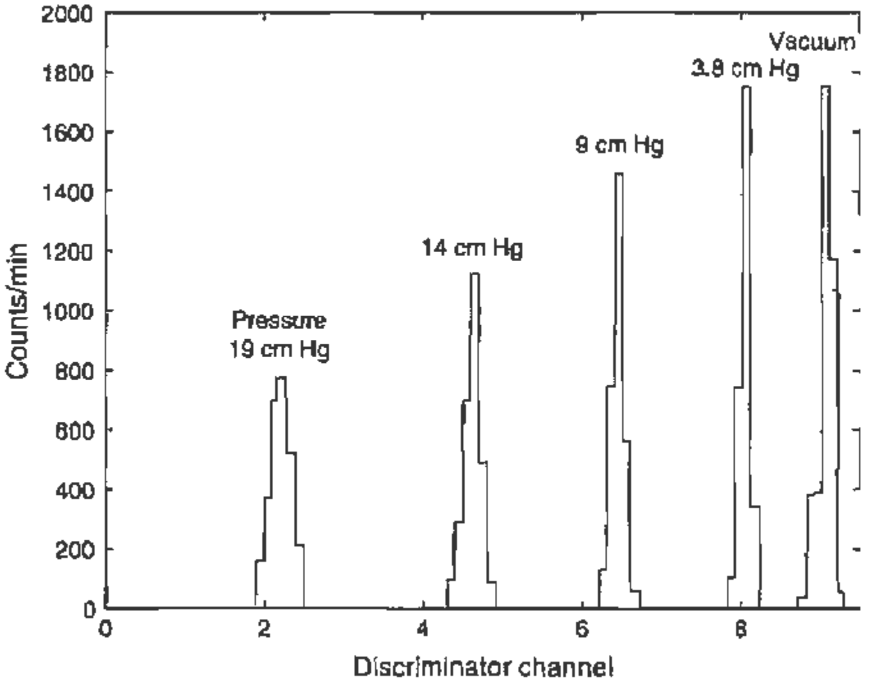


FIGURE 8.31 Distribution of output pulse height of the solid-state detector for five different pressures. Note the gradual decrease of the energy of the alpha particle.

Turning now to the measurements of type (b), Fig. 8.31 shows the distribution of the detector pulse heights as obtained with the single-channel discriminator (described in connection with the scintillation counter). Each peak corresponds to a different pressure, and we thus note that the alpha particles reach the detector with progressively less energy when they have traversed more grams per squared centimeter of air. We set the pulse height obtained in vacuum equal to the full energy of the ^{210}Po alpha particle, namely, 5.25 MeV, and use the linear characteristic of the solid-state detector to obtain the energy of the alphas as a function of material traversed. The results obtained by a student are given in Fig. 8.32 (solid curve). If the derivative of the energy curve is taken with respect to distance, we obtain the energy-loss curve, dE/dx, as a function of distance, as shown by the dashed curve in Fig. 8.32. Such a curve is called a Bragg curve, and shows a $1/E$ dependence[45] as predicted by Eq. (8.12); for the α particles KE $= \frac{1}{2}Mv^2$ and the influence of the logarithmic term of Eq. (8.12) is minimal. As the particle reaches the end of its range the energy loss dE/dx drops rapidly to 0.

[45] We might plot the dE/dx curve against energy by making use of the data of the energy curve to express the distance from the stopping point in energy units.

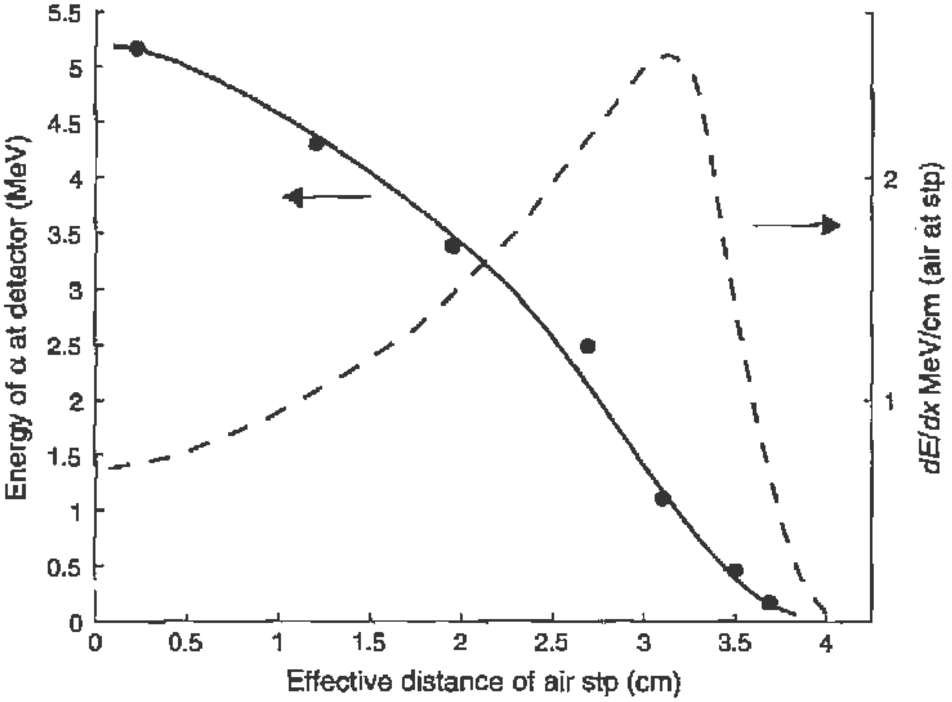


FIGURE 8.32 Plot of the residual energy of a polonium alpha particle when it reaches the detector as a function of air pressure (plotted, however, in terms of the equivalent amount of air (stp) traversed). These data are obtained from distributions such as those shown in Fig. 8.31. The dashed curve represents the derivative of the solid (energy) curve; thus, it gives the energy loss per unit length. It is called the "Bragg curve."

From the energy curve of Fig. 8.32, we note that in air at stp the polonium alpha particle produces at the end of its range approximately 67,000 electron–ion pairs per centimeter, whereas at its full energy it produces only 20,000 pairs per centimeter; these numbers were obtained by using an average loss of 36 eV for the production of one electron–ion pair in air.

8.6. NUCLEAR HALF-LIFE MEASUREMENTS

We will now discuss the measurements of the half-lives of some short-lived nuclear states. A thorough discussion of the time distribution in the decay of nuclear states is given in Section 10.5. Simply put, for a large sample of nuclei, the number of decays per second, the decay rate[46] R, will be

[46] For historical reasons, the standard unit for decay rate is the Curie $\equiv 3.7 \times 10^{10}$ decays per second. This is the number of decays per second in one gram of radium. The modern (SI) unit is the Bequerel, defined as one disintegration per second, so 1 Bq $= 1/(3.7 \times 10^{10})$ Ci. For more details, see Appendix D.

proportional to the number N of nuclei in the sample at any particular time. That is,

$$R = \frac{dN}{dt} = -\lambda N.$$

The proportionality constant is called $-\lambda$, the minus sign reflecting the fact that the decay causes the number of nuclei to decrease with time. This differential equation has a simple solution, namely

$$N(t) = N_0 e^{-\lambda t},$$

where N_0 is the number of nuclei present at $t = 0$. Obviously, λ characterizes the lifetime. The larger λ is, the faster the sample decays, and the shorter the lifetime is. There are two definitions we use for the lifetime. One is the *mean life*:

$$\tau = \frac{1}{\lambda}.$$

The other is more practically minded, and measures the time it takes for the sample to decay to $1/2$ its original number. This is called the *half-life*, and it is determined by solving $N(t) = N_0/2$ for t.

$$t_{1/2} = \frac{\ln 2}{\lambda} = 0.693\tau.$$

References *usually* quote the half-life, but not always. Be sure when you look up a lifetime, that you are getting the half-life or mean life. A good source of information on nuclear decay half-lives is the National Nuclear Data Center at Brookhaven National Laboratory and available at the Web site http://www.nndc.bnl.gov/nndc/nudat/radform.html.

Obviously, we must resort to some sort of trick to obtain a sample nuclei with short-lived states that can be measured. One trick we will use is the chemical separation of barium from cesium. However, we will also create new isotopes using a type of nuclear reaction called *neutron activiation*. In neutron activation, reactions with neutrons are used to create radioactive isotopes from stable nuclei. Neutrons are produced using a plutonium-beryllium (PuBe) source, whicb is safely packaged away so you cannot get near it, and allows the neutrons to irradiate samples inserted into the container. Plutonium decays by α-emission, that is,

$$^{239}\mathrm{Pu} \rightarrow {}^{235}\mathrm{U} + \alpha,$$

and the α particles react with the beryllium

$$\alpha + {}^{9}\mathrm{Be} \rightarrow {}^{12}\mathrm{C} + n,$$

releasing neutrons. These neutrons are slowed down by collisions with protons (in all the paraffin, a hydrocarbon, surrounding the source), making them available for other reactions. When you put an isotope in the neutron radiation "oven," make sure you "cook" it for at least a large fraction of one half-life. Otherwise, you may not get enough of a rate for you to measure.

Typically, the radiation from these sorts of sources are easily detected with a Geiger counter, and simple setups are available from a wide variety of vendors. These setups include simple interfaces to computers, as discussed in Section 3.9. The data presented here were taken with some different setups, including a multichannel scaler plug-in board for a desktop computer, and a Universal Laboratory Interface module with a computer running LoggerPro. All one really needs is an interface and software for counting pulses from the Geiger counter (or other detector and electronics) for a fixed period of time ("dwell time"), recording that number, and then counting again for the same fixed period of time, and so on. A graphical display of the data as it comes in is very useful, and generally part of any commercial package.

In what follows, we discuss the analysis of three radioactive isotopes with varying half-lives. A key point is the presence of some sort of "background" signal, in addition to the primary radioactive decay. (Such backgrounds are always present, at least at some level.) In the first example (^{116}In decay), the half-life is rather long, and a method for estimating the background level "by hand" and for incorporating its effect into the systematic error is outlined. In the case of ^{137m}Ba decay, a fitting technique that allows one to determine the background precisely and find the half-life with its corresponding random uncertainty is discussed. Finally, we discuss radioactive silver isotopes, which present a combined signal from two radioactive isotopes, each with relatively short half-lives.

8.6.1. Production and Decay of ^{116}In

You can produce ^{116}In using neutron capture on a piece of indium. Indium is a very common metal used for soldering compounds, and all of natural indium is the isotope ^{115}In. The decay scheme for ^{116}In to ^{116}Sn is shown in Fig. 8.33. Note that the ground state has a very short half-life, only 14 s.

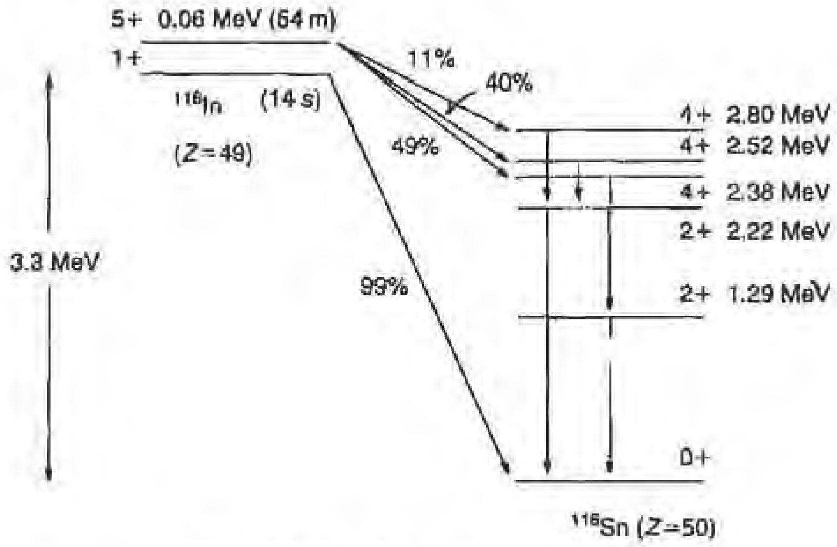


FIGURE 8.33 Decay scheme for ^{116}In.

You will be detecting β^- decay of the *excited* state, 60 keV above the ground state. The decays proceed mainly to a couple of states at around 2.3 MeV, and the available energy is 3.3 MeV, so the β^- typically have energies up to a megaelectronvolt or so. These are easy to detect in a Geiger counter.

Irradiate the piece of indium for an hour or so. Remove it and place it on the Geiger counter platform, close to the counter window. Take data for an hour or so, setting the multichannel scaling program to count for intervals of something like a minute.

It is probably a good idea to make a semilog plot of the data, and estimate the half-life by hand, just to make sure the result looks about right. To do a better job, you can easily fit the data to a decaying exponential. Just use the MATLAB function polyfit to fit the logarithm of the number of counts versus channel to a straight line. In fact, this is a case where you can accurately write the random errors of the points, since they are governed by a Poisson distribution. That is, if there are N counts in any one channel, then the random uncertainty in N is $\delta N = \sqrt{N}$, and the random uncertainty in the logarithm of N is $\delta \log N = 1/\sqrt{N}$.

A sample of data on indium decay is shown in Fig. 8.34. Each channel represents 30 s. The simple fit described above is shown by the *dashed* line. Note that the fit is not really very good. You can see that more clearly if you plot the difference between the fitted function and the data points.

In fact, this is not too surprising since you expect some background radiation from other radioactive isotopes in the piece of irradiated solder. You can try subtracting a constant value (representing the background counts)

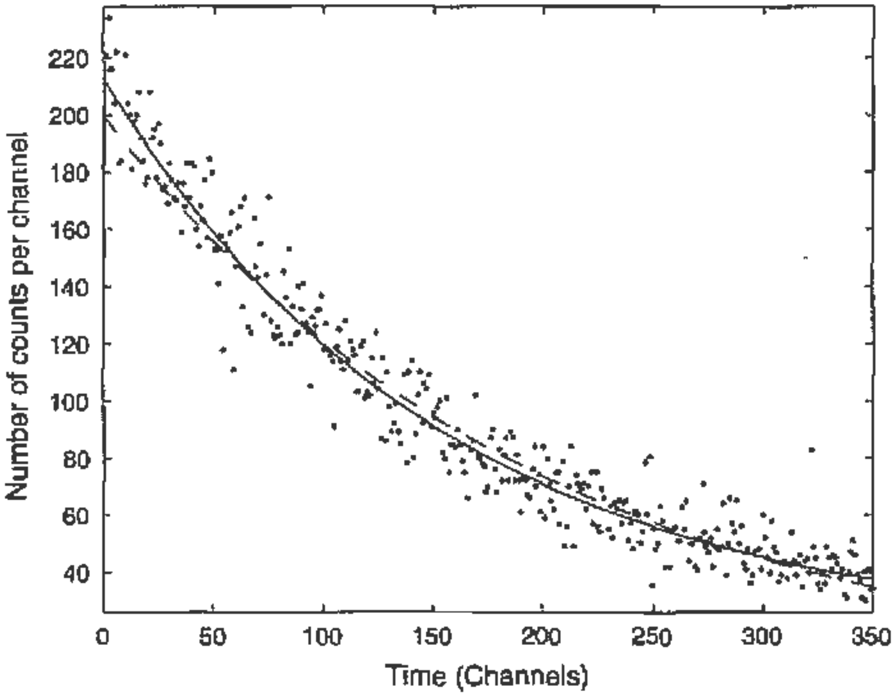


FIGURE 8.34 Data and fits for the decay of ^{116}In. The dashed line is fitted to a decaying exponential, while the solid line includes a constant background of 17 counts. The multichannel scaler recorded data every 30 s; that is, each channel represents 30 s.

from the data before you fit it, and see whether it looks better. By calculating the χ^2 function, you can even optimize the background term by minimizing χ^2.

The MATLAB program shown in Fig. 8.35 was used to do exactly this. After reading in the values of channel and counts, the user is asked for a number of background counts. Then this value is subtracted from the data, and care is taken to make sure the value is not less than 1. (Remember, you are going to take a logarithm.) Two fits are done, one that is unweighted (using polyfit) and one that is weighted according to the Poisson uncertainty in the points (using linreg). The results, including the χ^2, are printed and plotted. By trying various backgrounds, you find that the lowest χ^2 (i.e., the "best fit") is found for 17 background counts. You can even estimate your *systematic* uncertainty by looking at how much the lifetime varies as you move around in χ^2 near the minimum. This can be large if the minimum in χ^2 is shallow. For this particular data set, we find that

$$\tau = 160.7 \pm 2.0 \pm 10 \text{ channels},$$

```
% LOAD AND EXTRACT DATA POINTS
 load indium.dat
 chan=indium(:,1);
 data=indium(:,2);
%
% PREPARE DATA FOR FITTING LINE TO LOGARITHM
 bkgd=input('Background counts ');
 dnet=max(data-bkgd,1);
 ndof=length(data)-2;
 edata=sqrt(data);
 ldata=log(dnet);
 eldata=edata./dnet;
%
% UNWEIGHTED FIT
 coefa=polyfit(chan,ldata,1);
 fita=exp(polyval(coefa,chan));
 chisqa=sum(((dnet-fita)./edata).^2);
 fprintf('Unweighted fit:\n');
 fprintf(' tau=%6.3e\n',-1.0/coefa(1));
 fprintf(' chisquare/dof=%6.3f\n',chisqa/ndof);
%
% WEIGHTED FIT
 [coefb,ecoefb,lfitb]=linreg(chan,ldata,eldata);
 fitb=exp(lfitb);
 chisqb=sum(((dnet-fitb)./edata).^2);
 fprintf('Weighted fit:\n');
 fprintf(' tau=%6.3e',-1.0/coefb(2));
 fprintf(' uncert=%6.3e\n',ecoefb(2)/coefb(2)^2);
 fprintf(' chisquare/dof=%6.3f\n',chisqb/ndof);
```

FIGURE 8.35 A MATLAB program (i.e., m-file) used to fit indium data. The program asks the user for a number of background counts, then carries out the fit, and reports the results, including the χ^2. Although the background level can be fitted automatically using nonlinear fitting techniques, this program gives one a feeling for the sensitivity of the χ^2 to the background level.

where the first uncertainty is random and the second is systematic. Since each channel is 30 s, we determine that

$$t_{1/2} = \log 2 \times \tau \times \frac{1}{2}\mathrm{min/channel} = 55.7 \pm 0.7 \pm 3.5 \mathrm{~min},$$

which agrees well with the accepted value of 54 min. In fact, it seems we may have overestimated the systematic uncertainty.

Actually, this business of adjusting the background term to minimize χ^2 can be done automatically in MATLAB. That brings us into the world of nonlinear fitting, and we will do that next.

8.6.2. The Half-Life of ^{137m}Ba

Now we will measure the half-life of another short-lived isotope, ^{137m}Ba. The background is very clear in this case, and we will use that to go a step further in our data analysis techniques. This isotope does not need to be produced in the neutron oven.

Recall the decay scheme of ^{137}Cs in Fig. 8.21. The daughter nucleus, ^{137}Ba, is produced in its ground state only 5.4% of the time. The rest of the time it is made in the excited state, called ^{137m}Ba for "metastable," which decays by γ-ray emission, but with a relatively large half-life (for γ decay) of around 2.5 min. Of course, ^{137m}Ba is produced all the time, as the very long-lived ^{137}Cs decays, so you cannot isolate the ^{137m}Ba decay without somehow separating it from the ^{137}Cs.

You can make this separation because *chemically*, cesium is very different from barium. By passing a weak acid solution through a ^{137}Cs source, barium is captured and comes out in solution. Some cesium comes through as well, but most of the radioactivity of the solution is from ^{137m}Ba. Simple kits are available[47] for carrying out this chemical separation. It is best if you squeeze the drops through *slowly*, enough to fill the small metal holder in about 30 s. Then place the holder in the Geiger counter tray, and start the data acquisition program.

Realize that you are working with radioactivity and hydrochloric acid. **Do not be careless.** None of this is concentrated enough to be particularly dangerous, but you should take some simple precautions. Disposable gloves are located near the setup. It is also a good idea to wash your hands soon after you are finished.

You should choose a dwell time that allows you to get a relatively large number of points in each channel, but many channels over the expected decay time of a few minutes. You should be able to get several hundred counts per bin in the first bin or two, and a background of less than 20 counts per bin. (The background level will be clear after counting for a half-hour.) You might need a few tries to get all of this where you want it.

[47] For example, from TEL-Atomic, Inc., http://www.telatomic.com/.

You can use the program in Fig. 8.35 to fit the data and adjust the background counts, but that is tedious. In this case, since the background will be very clear, you can determine it precisely by averaging over the last many channels, and subtract that number from the data before fitting. However, MATLAB gives you the ability to fit things all at once.

What you need to do is minimize the χ^2 function numerically, and MATLAB gives you a numerical minimization function called fminsearch that can do this. You need to minimize χ^2 as a function of three variables, two for the exponential fit and one for the background value.

First, write a simple m-file called expcon.m, which calculates the function you are going to fit to the data:

```
function y=expcon(x,N0,tau,bkgd)
 y=N0*exp(-x/tau)+bkgd;
```

and then write another called fitexpcon.m, which calculates χ^2:

```
function chisqr=fitexpcon(pars,xdata,ydata,edata)
chisqr=sum(((ydata-expcon(xdata,pars(1),pars(2),
 pars(3)))./edata).^2);
```

Do not forget that for these data, the array of uncertainties edata is just the square root of the counts, i.e., edata=sqrt(ydata). (If any of the channels has zero counts, then set edata equal to unity.)

Play around with some values of pars(1,2,3) so that you have a good starting point. (Just plot the data points, and then overplot the function expcon until it looks kind of close.) Then type the command

```
fminsearch(@fitexpcon,pars,0,[],xdata,ydata,edata)
```

and you will get the best-fit values returned. (Check the help documentation for details of the arguments for fminsearch.)

Exactly this procedure was followed to fit the data shown in Fig. 8.36. The fit achieves a minimum χ^2 for a lifetime $\tau = 3.80$ min, corresponding to a half-life $t_{1/2} = 2.63$ min. The random uncertainty is determined, as shown on the right in Fig. 8.36, from the values of τ that increase the minimum χ^2 by one unit. These χ^2 "data" are fitted to a parabola, and we determine the uncertainty in τ to be ± 0.10 min. Consequently, we

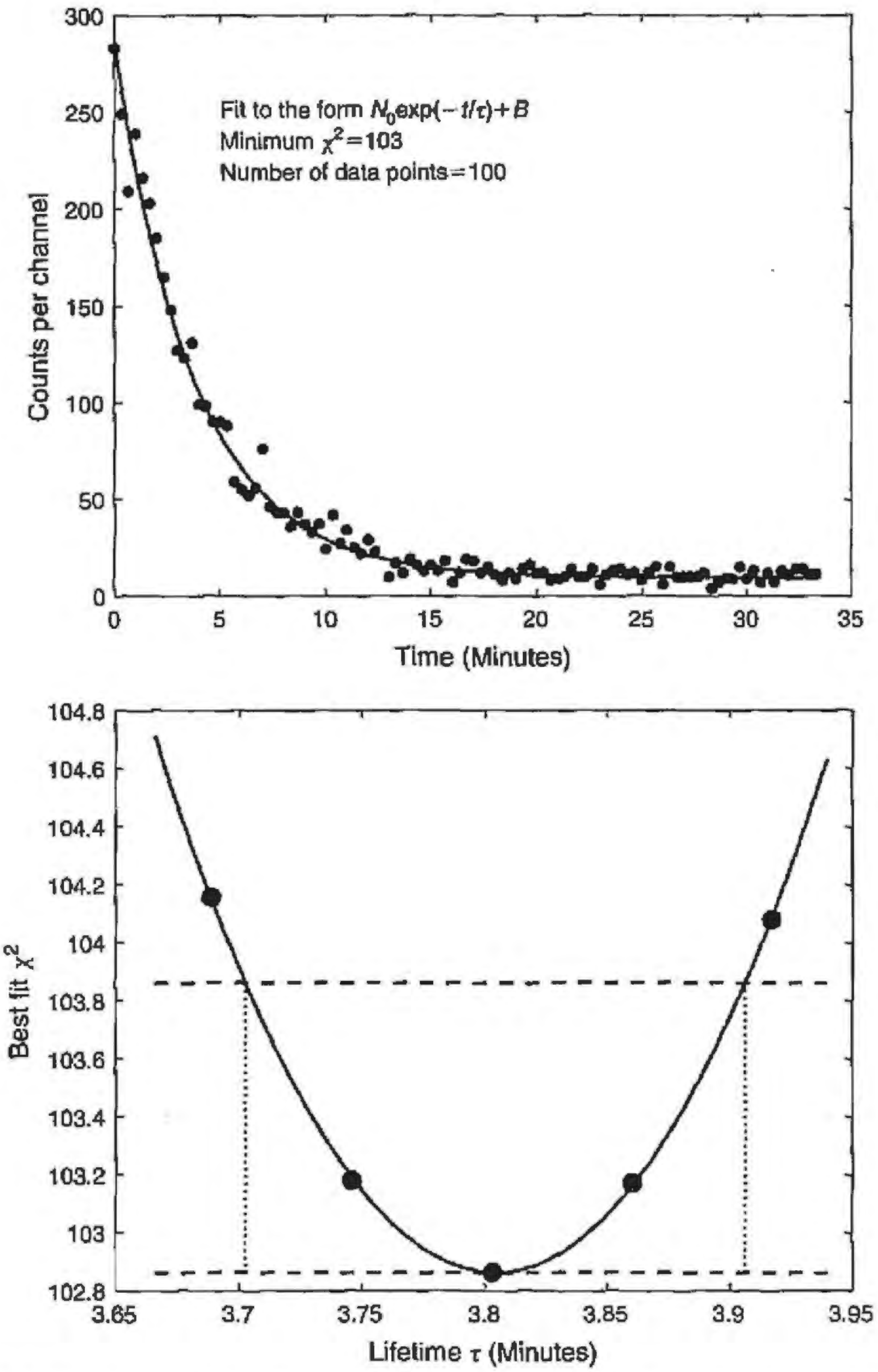


FIGURE 8.36 An example of a nonlinear fit. The data are from the decay of ^{137m}Ba, including some constant background. The MATLAB function fminsearch was used to make the fit. The plot on the top shows the best-fit curve, while the lower graph shows the χ^2 minima found by fixing the decay lifetime to various values. The random error in the lifetime is determined from the values that increase χ^2 by one unit.

find that

$$t_{1/2}(^{137m}\mathrm{Ba}) = 2.63 \pm 0.07 \text{ min}.$$

This is in good agreement with the accepted value of 2.55 min.

Note that the radioactivity you detect from ^{137m}Ba decay is γ radiation, which is not detected very efficiently by a Geiger counter. You might try using a NaI(Tl) detector instead, keying in on the particular γ-ray in question. This should greatly increase your counting statistics, as well as reduce the background.

8.6.3. Radioactive Silver Isotopes

Natural silver is pretty much evenly divided between two isotopes, ^{107}Ag and ^{109}Ag. Neutron activation captures a neutron equally well on these two isotopes, producing the two *radioactive* isotopes ^{108}Ag and ^{110}Ag. Both of these decay with a relatively high momentum β^- that is easy to detect, but one isotope has a half-life of 24.4 s and the other of 2.42 min. You might want to look up the decays to get more details.

Take a piece of pure silver foil and cook it in the neutron oven for at least 10 min. *Quickly* take it out, put it in the Geiger counter, and start the program. Do not forget that the lifetime of the shorter-lived isotope is only half a minute. It should be clear from the raw data that there are two lifetime components from the decay.

Representative data taken by students is shown in Fig. 8.37. The dwell time was set to 2.5 s, but in order to get better statistics in each channel, the MATLAB function reshape was used to add every four channels together. Error bars are added to the data points using the errorbar function. The points are fitted to a double exponential decay, completely analogous to the way we fitted a constant plus an exponential to the ^{137m}Ba data. The only difference is that the m-files for the fit function and for the χ^2 are changed slightly.

The best fit yields half-lives of 26.9 s and 3.53 min. The shorter half-life is in good agreement with the accepted value. The longer agrees much less well, but this is not surprising. No background term was included in the fit (leading to an overestimate of the half-life), and the statistical accuracy of the longer decay is clearly marginal. The ambitious student can explore these points using the techniques discussed for ^{116}In and ^{137m}Ba decay in the previous sections.

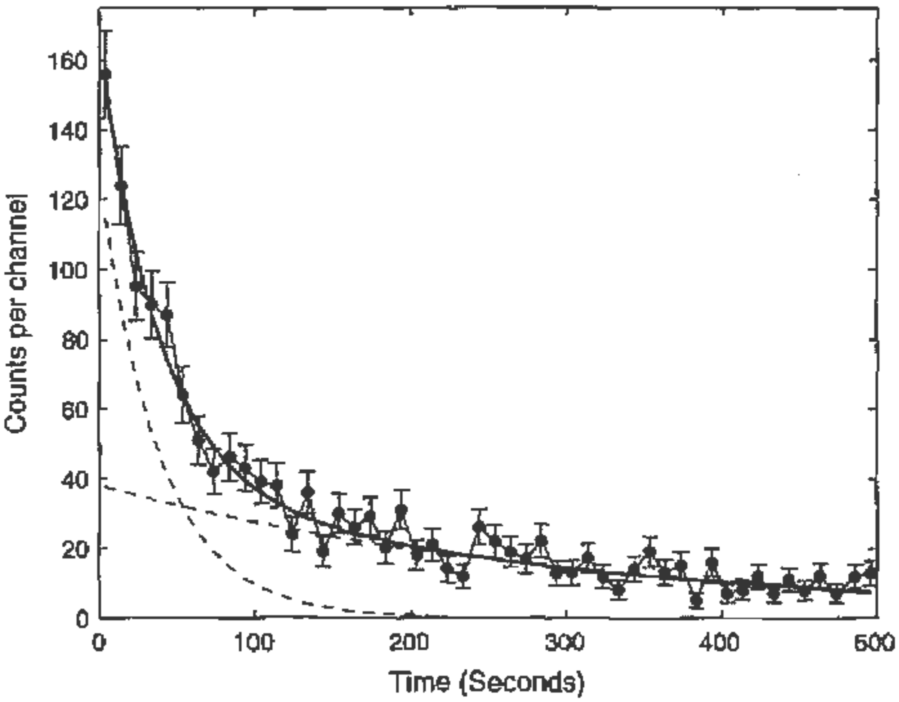


FIGURE 8.37 The decay of neutron-activated natural silver, fitted to the sum of two decaying exponential functions. The plot was made using the MATLAB function errorbar. In addition to the best-fit curve, we show the two individual exponentials separately.

8.7. REFERENCES

By necessity the discussion presented in this chapter is not complete. Below is a selective list of references (including those already mentioned in the footnotes to the chapter) that the reader may consult for additional information.

On interaction of radiation and particles with matter:

E. Fermi, *Nuclear Physics*, Univ. of Chicago Press, Chicago, 1950.
J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Wiley, New York, 1962.
W. Heitler, *The Quantum Theory of Radiation*, 3rd ed., Oxford Univ. Press, London, 1954.

On gaseous and scintillation detectors; neutron detectors:

W. J. Price, *Nuclear Radiation Detectors*, McGraw-Hill, New York, 1958.
B. Rossi and H. Staub, *Ionization Chambers and Counters*, McGraw-Hill, New York, 1949.
J. Sharpe, *Nuclear Radiation Detectors*, Methuen, London, 1955.
Encyclopedia of Physics, Vol. 45, *Nuclear Instrumentation II*, Springer-Verlag, Berlin, 1958.
J. B. Birks, *The Theory and Practice of Scintillation Counting*, Pergamon, New York, 1964.

On solid state detectors:

J. M. Taylor, *Semiconductor Particle Detectors*, Butterworth, London, 1963.

There are a number of good introductory textbooks on nuclear and particle physics. Some examples are:

K. S. Krane, *Introductory Nuclear Physics*, Wiley, New York, 1988. This is a good basic book with some discussion of experiments and experimental methods.

S. S. M. Wong, *Introductory Nuclear Physics*, 2nd ed., Wiley, New York, 1998. A bit higher level than Krane, but a thorough survey of the underlying physics of nuclei.

D. Griffiths, *Introduction to Elementary Particles*, Wiley, New York, 1987. An excellent undergraduate level discussion of particle physics.

D. H. Perkins, *Introduction to High Energy Physics*, 4th ed., Cambridge Univ. Press, Cambridge, UK, 2000. A modern, up-to-date version of a classic book.

Many of the details of detectors, materials, and the statistics of nuclear process, as well as an excellent summary of particle physics, can be found in:

Particle Data Group, Review of particle properties, *Eur. Phys. J. C* 15, 1–878 (2000).

CHAPTER 9

Scattering and Coincidence Experiments

9.1. INTRODUCTION

Ever since Rutherford performed his original experiments on the scattering of energetic alpha particles from atomic nuclei, scattering has become increasingly more powerful as a tool for investigating the forces between elementary particles. By now it is familiar to the reader that an electron, under the influence of the attractive electromagnetic force of the nucleus, may be found in a bound state. The classical analogue of this situation is the motion of the planets around the sun under the influence of the gravitational force; they describe elliptical orbits.

In general, a scattering experiment probes a system by sending a projectile "into" it, and then studying what "comes out" of it. Similarly, correlation or "coincidence" experiments can probe a system by looking at what comes out simultaneously in two or more directions. In this chapter, we will study some types of each of these measurements.

The experiments in this chapter make use of radioactive sources. We recommend that the reader review the material on radiation safety in Appendix D before undertaking these measurements.

The concept of "solid angle" is important for understanding the formalism dealing with cross sections. The solid angle is a three-dimensional generalization of the familiar planar angle $\Delta\theta$, which is the length of a circular arc Δs divided by the radius r of the circle, i.e., $\Delta\theta = \Delta s/r$. Solid angle $\Delta\Omega$ is the area ΔA of a piece of a spherical surface, divided by the square of the radius, i.e., $\Delta\Omega = \Delta A/r^2$. Planar angles are measured in radians and solid angles are measured in steradians. Just as a circle subtends a planar angle of 2π to any point included in the circle, a sphere subtends a solid angle of 4π to any included point.

Solid angle is a useful concept whenever we are dealing with some sort of detector intercepting radiation which spreads out in all directions from a source. Ionizing radiation and elementary particle detectors are just one example, but you would encounter the same thing in fields like optics or sonics.

To be explicit, let $d\mathbf{A}$ be a vector whose magnitude is an area dA in some plane, and whose direction is normal to that plane. Let $\mathbf{n}$ be a unit vector pointing toward the source, which is a distance r away. Then

$$d\Omega = \frac{\mathbf{n}\cdot d\mathbf{A}}{r^2} \equiv \frac{dA_\perp}{r^2}, \qquad (9.1)$$

where $dA_\perp$ is just the perpendicular component of the area. A spherical surface is most convenient since all surface elements are normal to the direction to the center. In spherical coordinates (r, θ, ϕ), where $0 \leq \theta \leq \pi$ is the polar angle and $0 \leq \phi \leq 2\pi$ is the azimuthal angle, a differential element of the surface has area

$$dA = \text{width} \times \text{height} = (r \sin\theta \, d\phi) \times (r \, d\theta) = r^2 \sin\theta \, d\theta \, d\phi,$$

so the infinitesimal solid angle is just

$$d\Omega = \sin\theta \, d\theta \, d\phi. \qquad (9.2)$$

You will encounter this equation many times in physics.

We can easily apply this to the common case of a "detector face," normal to the direction of the incident radiation, as shown in Fig. 9.1. Let the "detector face" be a circular area with radius R located a distance d from a source. There is perfect azimuthal symmetry, so we immediately integrate

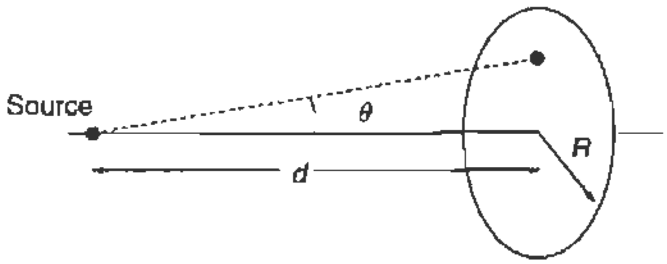


FIGURE 9.1 Calculating the solid angle of a circular face.

over ϕ to get

$$d\Omega = 2\pi \sin\theta\, d\theta$$

and integrate from $\theta = 0$ to $\theta_{\max} = \tan^{-1}(R/d)$ to get

$$\frac{\Delta\Omega}{4\pi} = \frac{1}{2}\int_{\theta=0}^{\theta_{\max}} \sin\theta\, d\theta,$$

where we have written the fraction of the total solid angle as $\Delta\Omega/4\pi$. This integral is done most easily by a change of variables to $\mu = \cos\theta$ with μ ranging from $\cos\theta_{\max} = d/\sqrt{d^2 + R^2}$ to 1. Since $d\mu = -\sin\theta\, d\theta$,

$$\frac{\Delta\Omega}{4\pi} = \int_{\cos\theta_{\max}}^{1} d\mu = \frac{1}{2}\left[1 - \frac{d}{(d^2 + R^2)^{1/2}}\right]. \tag{9.3}$$

For $d = 0$, $\Delta\Omega/4\pi = 1/2$, that is, the surface covers one entire hemisphere. For $d \to \infty$, expand Eq. (9.3) to first order in R/d to find $\Delta\Omega/4\pi = R^2/4d^2$ or $\Delta\Omega = (\pi R^2)/d^2$, which is just what you expect from the basic definition of solid angle.

9.2. COMPTON SCATTERING

9.2.1. Frequency Shift and Cross Section

This section deals with the scattering of electromagnetic radiation by free electrons. As mentioned in the introduction to this chapter, it is the scattering of electromagnetic radiation from various objects that makes it possible for us to "see" them. However, as the frequency of the radiation is increased beyond the visible region, the light quanta have energies comparable to, or larger than, the binding energy of the electrons in atoms, and the electrons can therefore be considered as free.

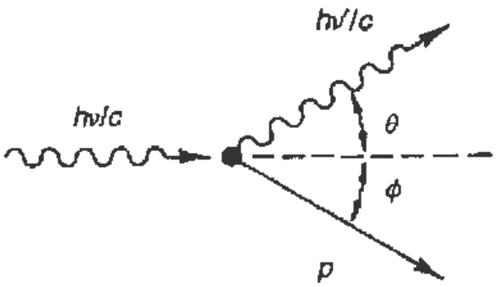


FIGURE 9.2 Compton scattering of a photon from a free electron.

In 1920 A. H. Compton investigated the scattering of monochromatic X-rays from various materials. He observed that after the scattering, the energy (frequency) of the X-rays had changed, and had always decreased. From the point of view of classical electromagnetic theory, this frequency shift cannot be explained,[1] since the frequency is a property of the incoming electromagnetic wave (field) and cannot be altered by the change of direction implied by the scattering. If, on the other hand, we think of the incoming radiation as being represented by a beam of photons, we need only consider the scattering of a quantum of energy $E = h\nu$ from a free electron; then, because of energy–momentum conservation, the scattered quantum has energy $E' = h\nu' < E$, in complete agreement with the experiments of Compton.

The frequency shift will depend on the angle of scattering and can be easily calculated from the kinematics. Consider an incoming photon of energy $E = h\nu$ and momentum $h\nu/c$ (Fig. 9.2) scattering from an electron (at rest) of mass m; p is the momentum of the electron after scattering, and $h\nu'$ and $h\nu'/c$ are the energy and momentum of the photon after the scattering. The three vectors $h\nu/c$, $h\nu'/c$, p must lie on the same plane, and energy conservation yields

$$h\nu + mc^2 = h\nu' + \sqrt{p^2c^2 + m^2c^4}. \tag{9.4}$$

From momentum conservation we obtain

$$h\nu = h\nu' \cos\theta + cp\cos\phi \tag{9.5}$$

$$0 = h\nu' \sin\theta - cp\sin\phi. \tag{9.6}$$

[1] See, for example, J. D. Jackson, *Classical Electrodynamics*, 3rd ed., p. 694, Wiley, New York, 1999.

Here θ is the photon scattering angle, and ϕ the electron recoil angle. To solve the above equations we transpose appropriately, square, and add Eq. (9.5) and Eq. (9.6) to obtain

$$h^2\nu^2 - 2h^2\nu\nu'\cos\theta + h^2\nu'^2 = c^2p^2.$$

By squaring Eq. (9.4) we obtain

$$h^2\nu^2 + h^2\nu'^2 - 2h^2\nu\nu' + 2hmc^2(\nu - \nu') = c^2p^2;$$

and substraction of the two above expressions yields

$$\frac{\nu - \nu'}{\nu\nu'} = \frac{h}{mc^2}(1 - \cos\theta). \tag{9.7}$$

We can recast Eq. (9.7) into two more familiar forms: (a) to give the shift in wavelength of the scattered X-ray beam

$$\Delta\lambda = \lambda' - \lambda = \frac{h}{mc}(1 - \cos\theta) \tag{9.8}$$

or (b) to give the energy of the scattered photon

$$E' = \frac{E}{1 + (E/mc^2)(1 - \cos\theta)}. \tag{9.9}$$

From Eq. (9.8) we see that the shift in wavelength, except for the angular dependence, is a constant, the Compton wavelength[2]

$$h/mc = 2.42 \times 10^{-10}\ \text{cm} = 0.0242\ \text{Å}.$$

For low-energy photons, with $\lambda \gg 0.02$ Å, the Compton shift is very small, whereas for high-energy photons with $\lambda \ll 0.02$ Å, the wavelength of the *scattered* radiation is always on the order of 0.02 Å, the Compton wavelength. These conclusions can equally well be obtained from Eq. (9.9), where the energy shift increases when E/mc^2 becomes large. For $E/mc^2 \gg 1$, E' is independent of E and on the order of $E' \approx mc^2$. Hence $\lambda' = c/\nu' = c/(E'/h) \sim c/(mc^2/h) = h/mc$ as stated before.

As an example, in this laboratory gamma rays from ^{137}Cs are scattered from an aluminum target; since $E = 0.662$ MeV, we have $E/mc^2 = 1.29$, so that backscattered gamma rays ($\theta = 180°$) will have $E' = E/3.6$,

[2] The mass of the electron m_e was used in evaluating h/mc; by using the mass of the pion, or another particle, we obtain the pion Compton wavelength, and so forth.

which is less than 30% of their original energy. It thus becomes quite easy to observe the Compton energy shift as compared to X-ray scattering, where, if we assume $\lambda = 2$ Å, $\Delta\lambda/\lambda = \Delta E/E = 0.01$.

In the original experiments Compton and his collaborators observed (especially for high Z materials) in addition to the frequency-shifted X-rays, scattered radiation *not shifted* in frequency. The unshifted X-rays are due to scattering from electrons that remained bound in the atom[3]: in this process the recoiling system is the entire atom, and we replace in Eq. (9.8) m by m_A (where $m_A \approx 2000 \times A \times m_e$), resulting in an undetectable wavelength shift, $\Delta\lambda' \approx 10^{-7}$ Å.

Next we are interested in the differential cross section for the scattering of the radiation from the electrons. Classically this is given by the Thomson cross section,[4] which can be easily derived: consider a plane wave propagating in the z direction with the $\mathbf{E}$ vector linearly polarized along the x direction. This is incident on an electron of mass m, as shown in Fig. 9.3. The electron will experience a force $F = eE = eE_0 \cos \omega t$, and its acceleration will be

$$\dot{v} = \frac{eE_0}{m} \cos \omega t.$$

According to Eq. (8.27), the power radiated by this accelerated electron will be (nonrelativistically, in SI units)

$$\frac{dP}{d\Omega} = \frac{1}{4\pi} \frac{e^2}{4\pi\epsilon_0} \frac{1}{c^3} \dot{v}^2 \sin^2 \Theta, \tag{9.10}$$

where Θ is the angle between the direction of observation and the $\mathbf{E}$ vector of the incoming wave. Using the expression for $\dot{v}$, we can write for Eq. (9.10) averaged over one cycle

$$\left\langle \frac{dP}{d\Omega} \right\rangle = \frac{1}{2} \left(\frac{e^2}{4\pi\epsilon_0 mc^2} \right)^2 \epsilon_0 E_0^2 c \sin^2 \Theta.$$

Finally, from the definition of the cross section (see Section 8.2.1.a) we have

$$\frac{d\sigma}{d\Omega} = \frac{\text{energy radiated}/(\text{unit time} - \text{unit solid angle})}{\text{incident energy}/(\text{unit area} - \text{unit time})}.$$

[3] A similar situation is discussed in the following section on the Mössbauer effect, where the nucleus remains bound in the lattice and the recoiling system is the entire crystal.

[4] See also Section 8.2.5.

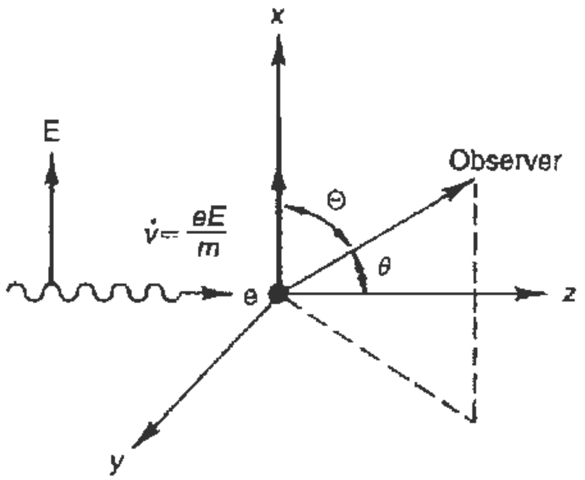


FIGURE 9.3 Classical picture of the scattering of electromagnetic radiation by an electron; this leads to the Thomson cross section.

Here the denominator is clearly given by the Poynting vector

$$\langle I \rangle = \frac{1}{2}\sqrt{\frac{\epsilon_0}{\mu_0}}E_0^2 = \frac{1}{2}\epsilon_0 c E_0^2.$$

Thus we obtain

$$\frac{d\sigma}{d\Omega} = \left(\frac{e^2}{4\pi\epsilon_0 mc^2}\right)^2 \sin^2\Theta, \tag{9.11}$$

where

$$\frac{e^2}{4\pi\epsilon_0 mc^2} = r_0$$

has dimensions of length, and is referred to as the "classical electron radius"

$$r_0 = 2.82 \times 10^{-13}\ \text{cm}.$$

Finally, we average over all possible directions of polarization of the incoming wave and use the angle θ measured from the direction of propagation of the incident wave to obtain

$$\frac{d\sigma}{d\Omega} = r_0^2\left(\frac{1+\cos^2\theta}{2}\right). \tag{9.12}$$

When integrated over all angles, Eq. (9.12) yields the Thomson cross section

$$\sigma_\mathrm{T} = \frac{8\pi}{3} r_0^2. \tag{9.13}$$

(This result was given without proof in Eq. (8.21).)

Several objections can be raised to the simple cross section given by Eq. (9.12) or Eq. (9.13): (a) it does not depend on frequency, a fact not supported by experiment; (b) the electron, even though free, is assumed not to recoil; (c) the treatment is nonrelativistic; and (d) quantum effects are not taken into account. Indeed, the correct quantum-mechanical calculation for Compton scattering yields the so called Klein–Nishina formula[5]

$$\frac{d\sigma}{d\Omega} = r_0^2 \frac{1+\cos^2\theta}{2} \frac{1}{[1+\gamma(1-\cos\theta)]^2} \times \left[1 + \frac{\gamma^2(1-\cos\theta)^2}{\left(1+\cos^2\theta\right)[1+\gamma(1-\cos\theta)]}\right], \tag{9.14}$$

where r_0 and θ were defined previously, and $\gamma = h\nu/mc^2$. The cross section has been averaged over incoming (and summed over outgoing) polarizations. By integrating Eq. (9.14), the total cross section can be obtained. We will not give the complete result here, but the asymptotic expressions have already been presented in Eq. (8.22).

A comparison of the Thomson (Eq. (9.12)) and Klein–Nishina cross sections, including the results obtained in this laboratory for $\gamma = 1.29$, is shown in Fig. 9.8. We remark that although the Thomson cross section is symmetric about 90°, the Klein–Nishina cross section is peaked forward strongly as γ increases. This is due to a great extent to kinematical factors associated with the Lorentz transformation from the center of mass to the laboratory; note that the center-of-mass velocity of the (indicent gamma ray + free electron) system is

$$\bar{v} = c\bar{\beta} = c\gamma/(1+\gamma),$$

where as before $\gamma = h\nu/mc^2$.

The experimental data are in perfect agreement with the results of Eqs. (9.9) and (9.14), which are among the most impressive and convincing

[5] See for instance F. Gross, *Relativistic Quantum Mechanics and Field Theory*, Section 10.5, Wiley, New York, 1993.

successes of quantum theory. In the following two sections we will describe the experimental verification of these predictions.

9.2.2. The Compton Scattering Experiment

As with any scattering experiment, the apparatus will consist of:

(a) The beam of incident particles, in this case photons,

(b) The target (containing the electrons from which the photons scatter), and

(c) The detector of the scattered photons.

The beam of photons is obtained by collimating the gamma radiation from a ^{137}Cs source. An intense source is required in order to get an appreciable counting rate for the scattered photons. As shown in Fig. 8.21 ^{137}Cs (^{137}Ba) emits a gamma ray of energy 0.662 MeV, and the detection techniques have been discussed in Chapter 8. Figure 8.21 also shows the pulse-height spectrum of the gamma radiation from ^{137}Cs, as obtained with standard equipment; the same detection equipment is used in this experiment with the only difference that heavy shielding is needed to prevent the detector from seeing the intense ^{137}Cs source directly.

A schematic of the apparatus is shown in Fig. 9.4. The lead pig A is fixed and holds the source, which can be introduced through the vertical hole (V). Another lead shield B contains the detector and can be rotated about the center, where the target is located. The lead assemblies are rather heavy (approximately 100 lb) and some provisions must be taken for adequate mounting.

For the source, a 7-mCi ^{137}Cs sample was used, which was properly encapsulated before being shipped to the laboratory. It should always be transported in a lead container, and when transferred into the lead pig A, it must be handled only by the attached string. The source holder (A) has a collimator (h) drilled horizontally, subtending a solid angle on the order of 0.03 sr. Of interest to us will be the density of the photon beam at the target, and the *expected* value is

$$\frac{3.7 \times 10^{10} \times 0.007}{4\pi} \frac{1}{r^2} = 1.3 \times 10^4 \text{ photons/cm}^2\text{-s},$$

where we use a source-to-detector distance $r = 40$ cm, for the data presented here.

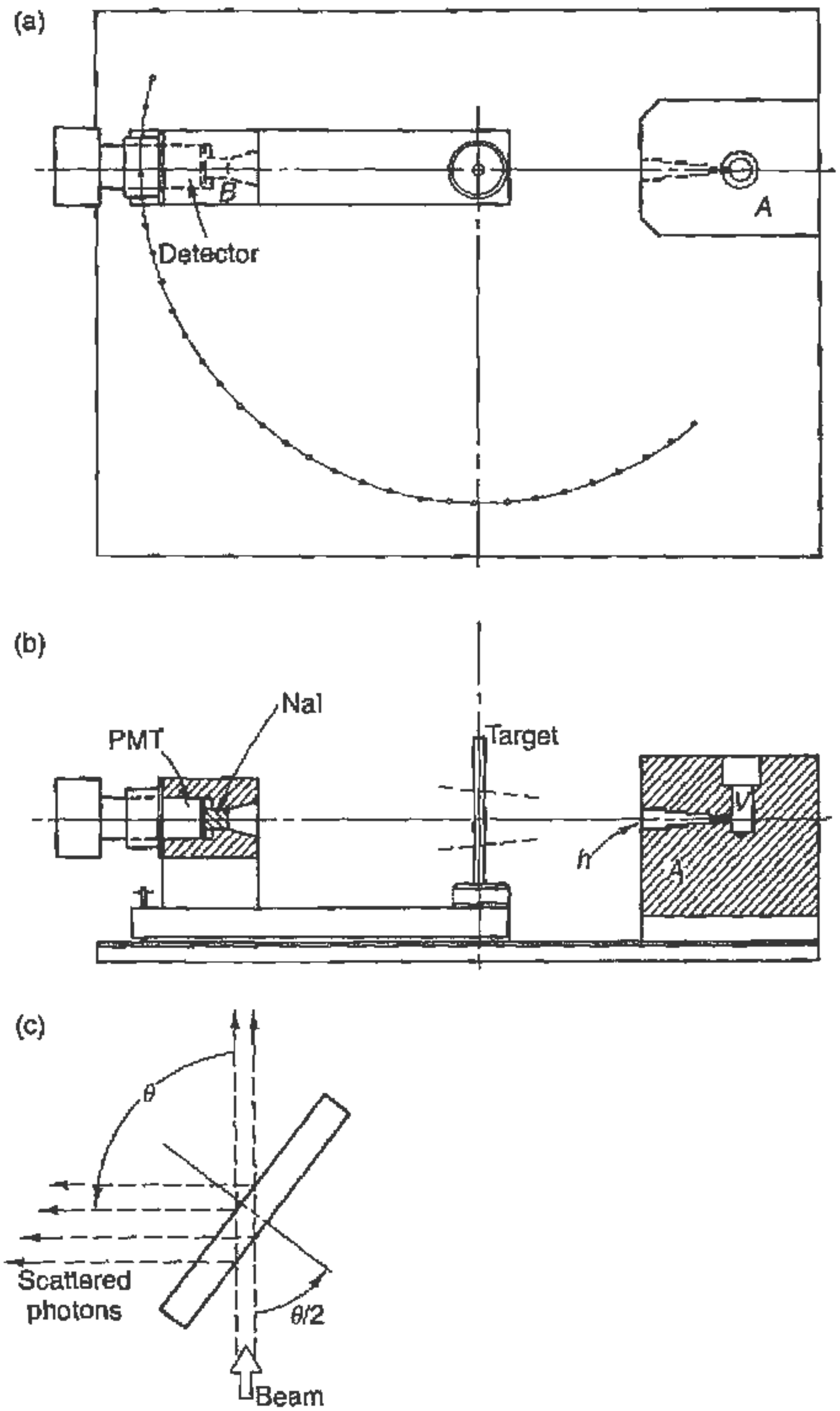


FIGURE 9.4 Schematic of an apparatus that can be used for measuring the Compton scattering of ^{137}Cs gamma rays from different targets: (a) top view and (b) elevation. The detector can be rotated relative to the beam direction, through a large angular range. Note that a less heavily shielded detector assembly is possible, but care needs to be taken so that the ^{137}Cs source is not directly visible to the detector at forward angles. (c) Use of a flat target when measuring Compton scattering at large angles. By such placement the scattered photons do not have to traverse very large amounts of the target material.

In contrast to the scattering of alpha particles, there is no need to enclose the beam and detector in vacuum or to use a very thin target. We know that gamma rays do not gradually lose energy when traversing matter as a charged particle does, but their interaction can be characterized by a mean free path. For the ^{137}Cs gamma ray we find that

$$\lambda = 4.7 \text{ cm in Al}, \qquad \lambda = 0.92 \text{ cm in Pb};$$

this corresponds to 10^4 cm of air, so that the interaction of the photon beam in the air of the apparatus (approximately 100 cm) is indeed negligible. Also, the target thickness can safely be a fraction of a mean free path before the probability for multiple interactions becomes considerable. Aluminum targets $\frac{1}{2}$ in. thick are quite adequate for this experiment.

Some special mention must be made of the geometrical shape of the target. We may use a flat target (such as an aluminum plate), in which event the cross section is obtained by considering the interaction of the total beam with the number of electrons per square centimeter of the target[6]; alternatively, we may use a target of circular cross section (such as a rod), in which event the cross section is obtained by considering the interaction of the beam density (photons per square centimeter) with the total number of electrons in the target.[7] When using a plate, it is advisable to rotate it so that it always bisects the angle between beam and detector, since otherwise the scattered photons may have to traverse a very large amount of material before leaving the target (see Fig. 9.4c). In that case, however, the amount of scattering material in the beam path varies as $1/\cos(\theta/2)$, and this correction must be applied to the yield of scattered particles. These effects are obviously eliminated when a target of circular cross section is used. In addition, the scattering point is better defined even if the beam is only poorly collimated. On the other hand, accurate evaluation of the flux density at the target is difficult. The results presented here were obtained by using a $\frac{3}{4}$ in.-diameter aluminum rod as the target.

An interesting refinement of the technique is made by observing the recoil electrons in time coincidence with the scattered photon. However, the kinetic energy of the recoil electron is

$$T_e = E - E' = E\frac{\gamma(1-\cos\theta')}{1+\gamma(1-\cos\theta)},$$

[6] See Fig. 8.1.

[7] See Fig. 8.1.

which at its maximum value ($\theta = 180°$) is

$$T(\text{electron}) = 0.662 \times (2.58/3.58) = 475 \text{ keV}.$$

The range of such an electron in aluminum is only 150 mg/cm^2 (see, for example, Feather's rule, Chapter 8, Eq. (8.15)), which corresponds to approximately 0.06 cm. Thus, the recoil electrons will, in almost all cases, stop in the target. On the other hand, if a plastic scintillator is used as the target, and is viewed with a photomultiplier, the recoil electrons do produce a signal that can be easily detected.

As mentioned before, the detection system consisted of a commercial NaI detector. The dimensions of the crystal were 3 in. diameter and 3 in. thick. Data was acquired with a multichannel analyzer, with a GPIB interface to a laptop computer. Figure 9.5 shows typical pulse-height spectra, taken at two different scattering angles (30° and 100°), and with the aluminum target rod both in and out of the beam, but with all running conditions otherwise identical. Each spectrum was acquired for 120 s. The difference between the target in and out spectra is also plotted.

By measuring the pulse-height distribution at various angles, we obtain the energy of the scattered photons as it is given by the position of the photopeak. This is most easily done by a simple Gaussian peak-fitting program to the photopeak as observed in the "background subtracted" spectra, for example, in the lower plots of Fig. 9.5. A rudimentary, but quite sufficient, Gaussian peak fit can be done in MATLAB by taking the logarithm of the net counts in the region of the photopeak, and fitting these to a second-order polynomial. To obtain the yield of scattered photons, we integrate the counts in the photopeak only and apply a correction for the "photofraction" or "peak-to-total ratio" as well as for crystal efficiency. These corrections depend on the crystal size and on the photon energy (which varies with angle). Figure 9.6 gives the peak-to-total ratio (for detectors at a specific distance from the photon source) and the detector efficiency, as a function of energy for several different NaI crystal dimensions.[8]

9.2.3. Results and Discussion

The results presented below were obtained by students using the apparatus described in the previous section.

[8] From *Efficiency Calculations for Selected Scintillators*, Bicron Corporation, available from the online library at http://www.bicron.com.

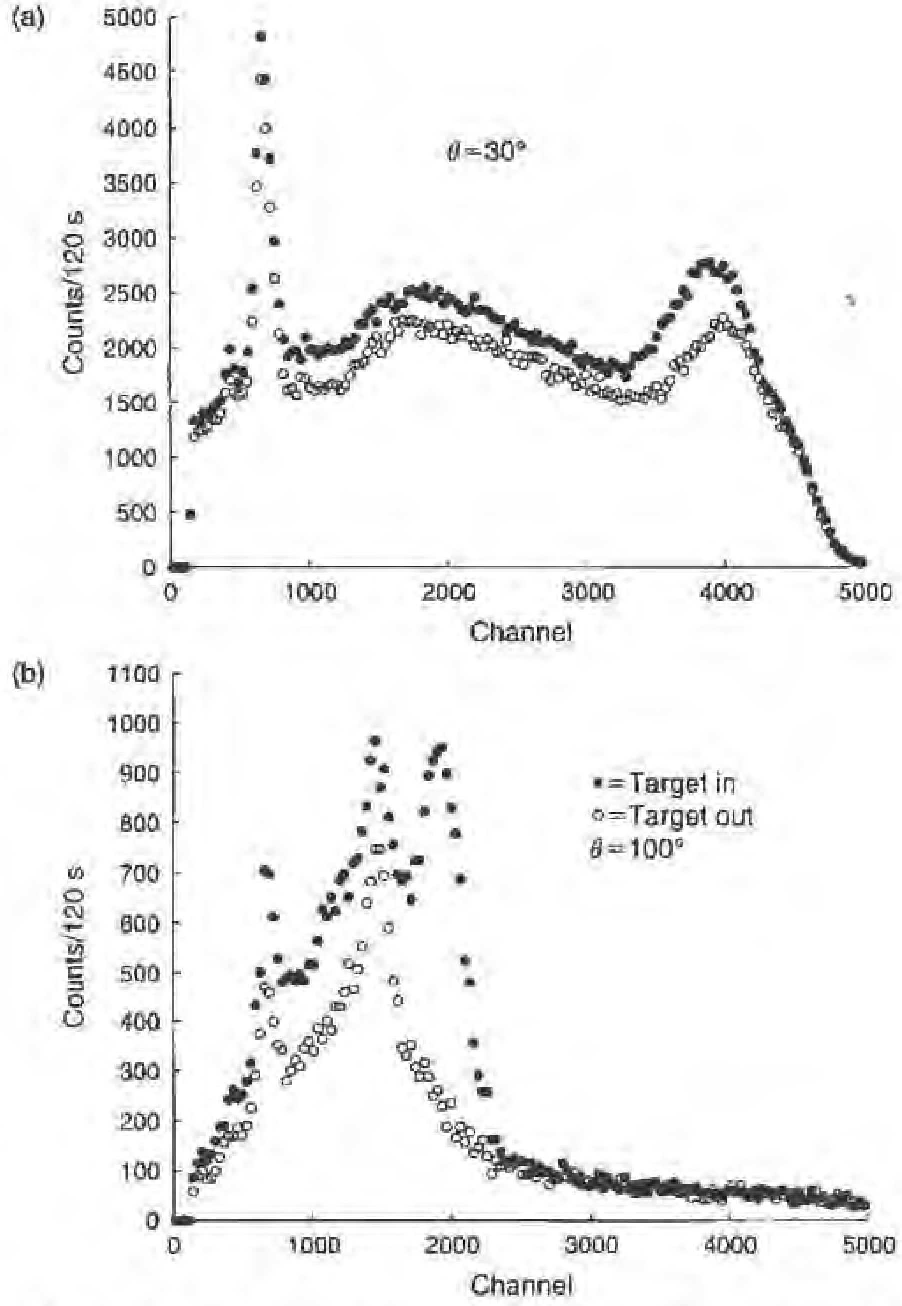


FIGURE 9.5 Pulse-height spectrum gamma rays in the Compton scattering apparatus. The plots (a), (b) show data acquired for 120 s both with the target rod in (solid points) and out (open circles) of the beam. At $\theta = 30°$, the detector intercepts some fraction of the primary beam, and the rate is considerably larger than at $\theta = 100°$. In addition, there are large signals due to K-shell X-rays and Compton backscattering in the lead shielding at both scattering angles. However, in each case, these background signals subtract cleanly away, leaving a pure Compton scattering signal from the aluminum target. The subtracted plots are shown in (c) and (d).

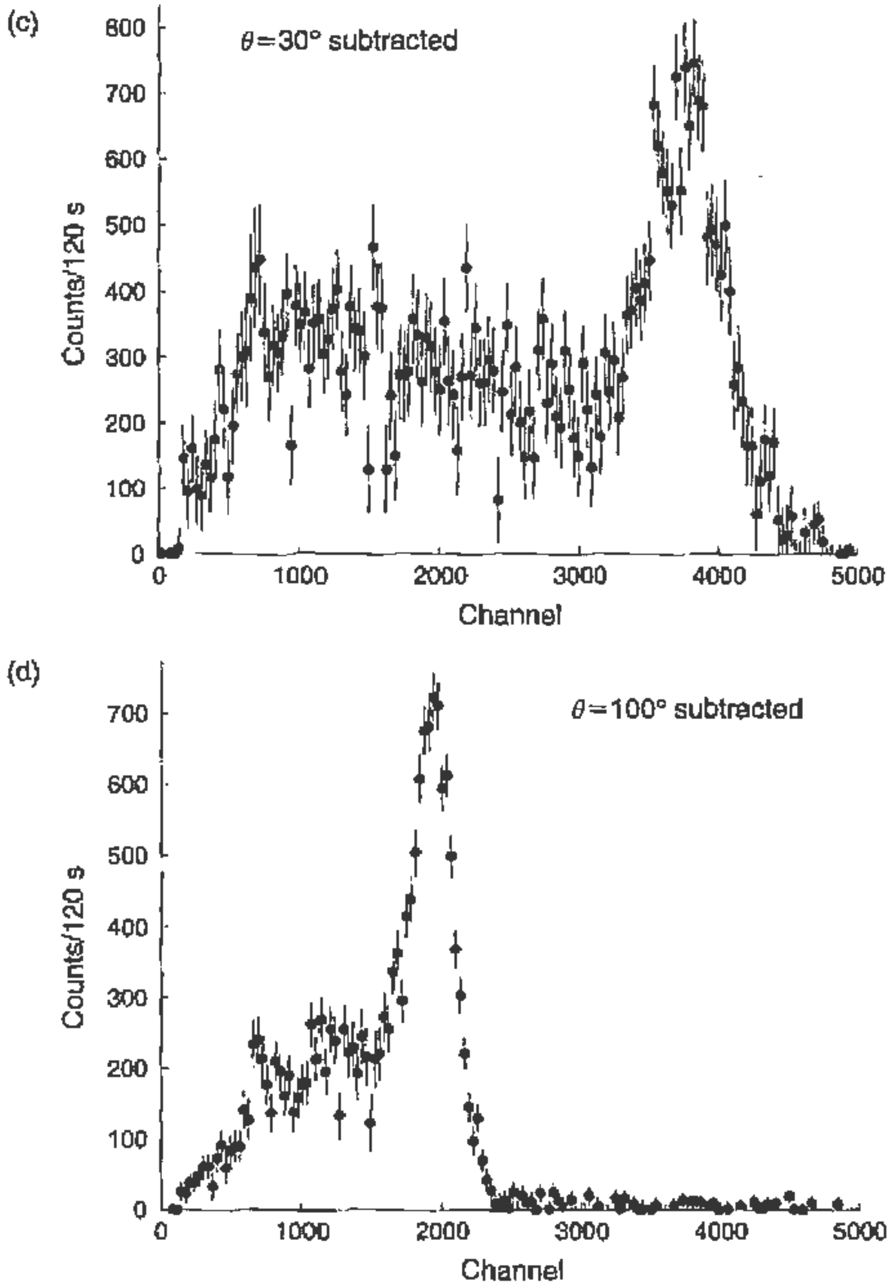


FIGURE 9.5 *Continued*

Before beginning measurements of Compton scattering, it is worthwhile to measure the beam profile of the ^{137}Cs source. This is best done by collimating the detector and putting it at a large distance from the source, so as to keep the count rate relatively low. (A number of difficulties arise at a high count rate, including severe dead time corrections and gain shifts, but these are negligible if the total rate is less than several kilohertz.) Then by moving the detector through different angles, one can map out the "shape" of the

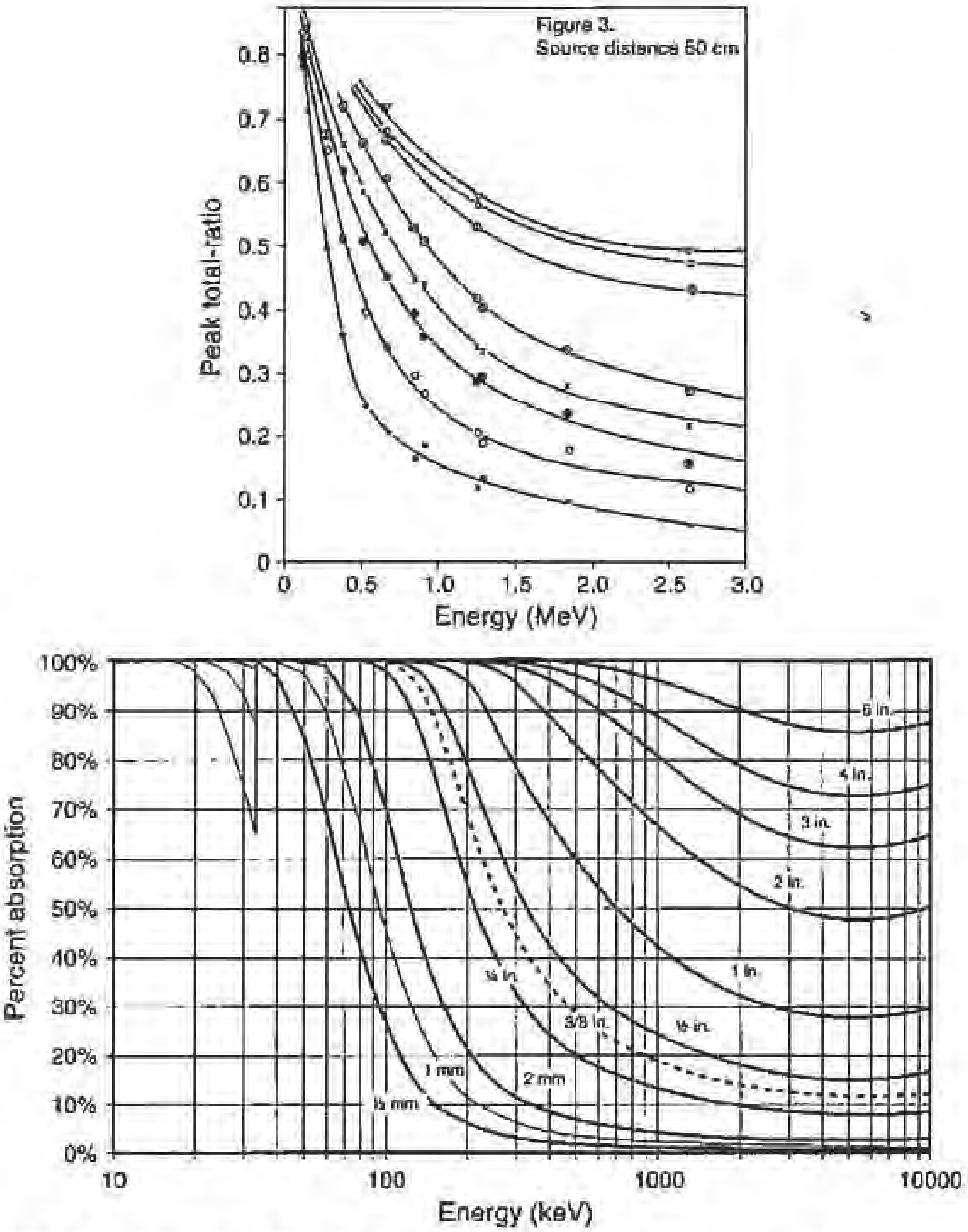


FIGURE 9.6 Detection efficiency plots for NaI crystals of various dimensions, from http://www.bicron.com. Shown are the peak-to-total ratio and the intrinsic absorption efficiency, all as a function of energy for various crystal dimensions.

photon beam. For our measurements here, however, we will simply assume the calculated beam flux for a measurement of the differential cross section.

Compton scattering data are taken by accumulating pulse-height spectra at various angles, both with the target in and out, for fixed periods of times. *In order to minimize the effects of gain drifts, and other changes over longer times, it is best to take the "in" and "out" spectra immediately*

TABLE 9.1 Summary of Compton Scattering Data

Angle (°)	Peak channel	Counts (in)	Counts (out)	E' (MeV)	Peak/total ratio	Efficiency	$d\sigma/d\Omega$ (10^{-27} cm^2/sr)
20	4300	528,161	508,714	0.614	0.47	0.865	55.2
30	3732	97,663	81,121	0.564	0.50	0.890	42.9
40	3384	29,856	14,566	0.508	0.53	0.930	35.8
60	2810	16,382	5062	0.402	0.57	0.960	23.9
80	2258	16,268	6251	0.320	0.65	0.990	18.0
100	1922	17,482	7632	0.263	0.72	0.999	15.8

Note. Each spectrum was acquired for 120 s.

one after the other. (For example, see Fig. 9.5.) Data taken by students are summarized in Table 9.1. In this table, E' is the photon energy as calculated from Eq. (9.9), and is used to look up the peak-to-total ratio and the intrinsic efficiency from Fig. 9.6.

Radioactive sources are used to calibrate the analyzer channel in terms of photon energy. (See Fig. 8.24 and the associated text. It is advisable to carry out a calibration both before and after taking Compton scattering data, in order to check for gain shifts.) In this experiment, it was determined that

$$\text{Energy} = 0.1527 \times \text{ Channel} - 34.96.$$

Then, using the photopeak values summarized in Table 9.1, we determine the scattered photon energy E'. In Fig. 9.7, we plot the inverse of the measured photon energy, $1/E'$, against $(1 - \cos\theta)$. According to Eq. (9.9), a straight line should be obtained, since

$$\frac{1}{E'} - \frac{1}{E} = \frac{1}{mc^2}(1 - \cos\theta).$$

This is indeed the result, and the slope of the line gives $1/mc^2$ with an intercept at $1/E$. From a least-squares fit we obtain

$$mc^2 = 505 \pm 12 \text{ keV}$$

in very good agreement with the known value of the electron mass. We thus conclude that Eq. (9.9) is very well verified and that our explanation of the Compton frequency shift is firmly supported by these data.

We next turn to the evaluation of the differential cross section. As explained before, we integrate the counts under the photopeak. The results

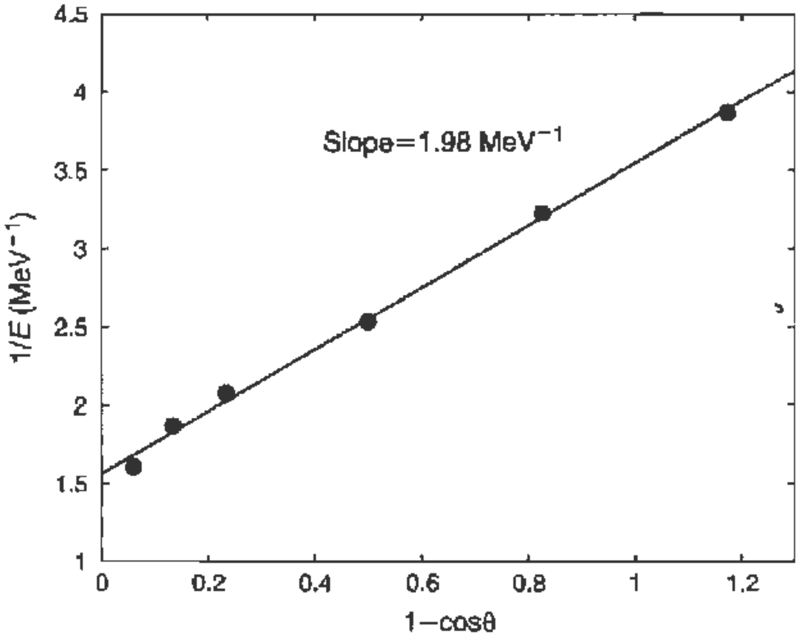


FIGURE 9.7 The results obtained for the energy (frequency shift) of the Compton scattered gamma rays. Note that $1/E$ is plotted against $(1-\cos\theta)$, leading to a linear dependence. The slope of the line gives the mass of the electron.

are also summarized in Table 9.1. To obtain the cross section we note that

$$\frac{d\sigma}{d\Omega} = \frac{\text{yield}}{(d\Omega)N I_0}.$$

The detector solid angle is given by

$$d\Omega = \frac{\text{crystal area}}{r^2} = 6.4 \times 10^{-2}\ \text{sr},$$

where r is the distance from the target to the detector. For the total number of electrons in the target, we have

$$N = \pi \left(\frac{d}{2}\right)^2 h\rho \frac{N_0}{A} Z,$$

where[9]

$$d = \text{diameter of target} = \tfrac{3}{4}\ \text{in.} = 1.91\ \text{cm}$$
$$h = \text{height of target} = 4\ \text{cm}$$

[9] The height of the target is obtained by estimating the length of target intercepted by the beam.

$$\begin{aligned}\rho &= \text{density of aluminum} = 2.7\ \text{gm/cm}^3\\ N_0 &= \text{Avogadro's number} = 6\times 10^{23}\\ A &= \text{atomic weight of aluminum} = 27\\ Z &= \text{atomic number of aluminum} = 13,\end{aligned}$$

thus

$$N = 8.9\times 10^{24}\ \text{electrons.}$$

For I_0, the flux density at the target, we use the previously obtained value

$$I_0 = 1.3\times 10^4\ \text{photons/cm}^2\text{-s},$$

and the data acquisition time for each spectrum is 120 s, so that finally

$$\begin{aligned}\frac{d\sigma}{d\Omega} &= \frac{\text{corrected yield}}{\left(6.4\times 10^{-2}\right)\times\left(8.9\times 10^{24}\right)\times\left(1.3\times 10^4\right)\times(120)}\\ &= \frac{\text{corrected yield}}{8.89\times 10^{29}}.\end{aligned}$$

The values of the differential cross section obtained in this fashion are given in Table 9.1, and are also plotted in Fig. 9.8. The solid line in Fig. 9.8

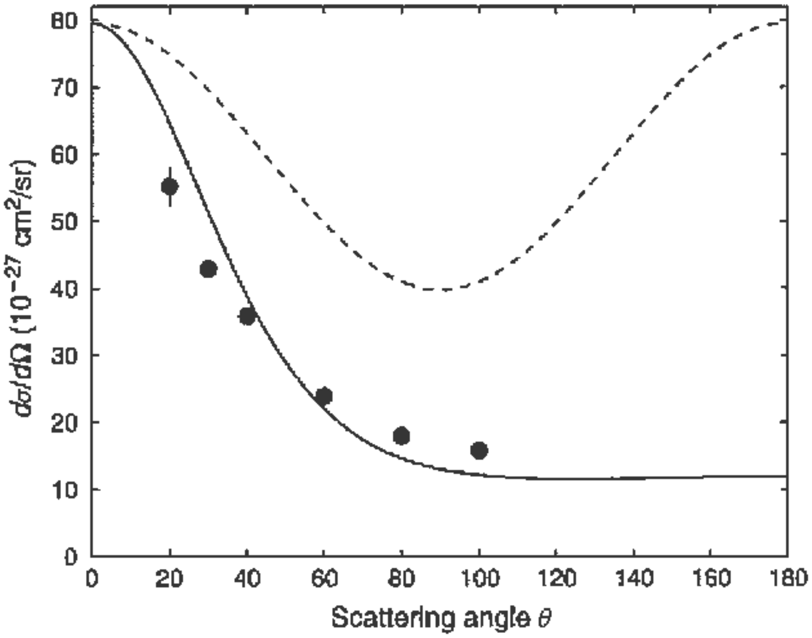


FIGURE 9.8 The results obtained for the scattering cross section of ^{137}Cs gamma rays as a function of angle. The solid line is the prediction of the Klein–Nishina formula for that particular energy; the dotted line is the Thomson cross section.

gives the theoretical values for $d\sigma/d\Omega$ derived from the Klein–Nishina formula (Eq. (9.14)) for $\gamma = 1.29$, while the dashed curve represents the Thomson cross section.

The agreement of the *angular dependence* of the experimental points with the theoretical curve is indeed quite good and clearly indicates the inadequacy of the Thomson cross section for the description of the scattering of high-energy photons, while confirming the Klein–Nishina formula. On the other hand the *absolute value* of the experimental cross section is subject to some uncertainty due to the way in which the flux density I_0 and total number of electrons N were estimated. Nevertheless, the agreement is good.

9.3. MÖSSBAUER EFFECT

9.3.1. General Considerations

In the Compton scattering experiment, we could visualize the scattering process as if it were a collision of two billiard balls in which the incoming photon maintained its identity but suffered a change in momentum and energy. The phenomenon of scattering can, however, also be visualized as the absorption by the target of quanta of the incoming beam, with the subsequent re-emission of these quanta; this was the model we used in the derivation of the Thomson scattering cross section in Section 9.2.

Since we know that emission of quanta of energy $h(\nu_\beta - \nu_\alpha)$ in the visible spectrum is due to transitions of atoms from a state of $\beta \to \alpha$ we must also expect that when quanta of this energy $h(\nu_\beta - \nu_\alpha)$ are incident on an atomic system in state α, they may be strongly absorbed, with the consequent raising of the atom from state α to state β. Evidence for such strong absorption is obtained by detecting radiation of frequency $(\nu_\beta - \nu_\alpha)$ emitted from the absorber in all directions; it is due to the atoms that, having absorbed a quantum from the beam, were raised to state β and then underwent a spontaneous transition back to state α, emitting the quantum $h(\nu_\beta - \nu_\alpha)$, but with equal probability into all directions. Such radiation is called "resonance radiation" and was first observed by R. W. Wood in sodium vapor in 1904. A schematic of the apparatus is shown in Fig. 9.9. An absorption cell was illuminated by sodium light, and at right angles to the incident beam the sodium D lines were observed.

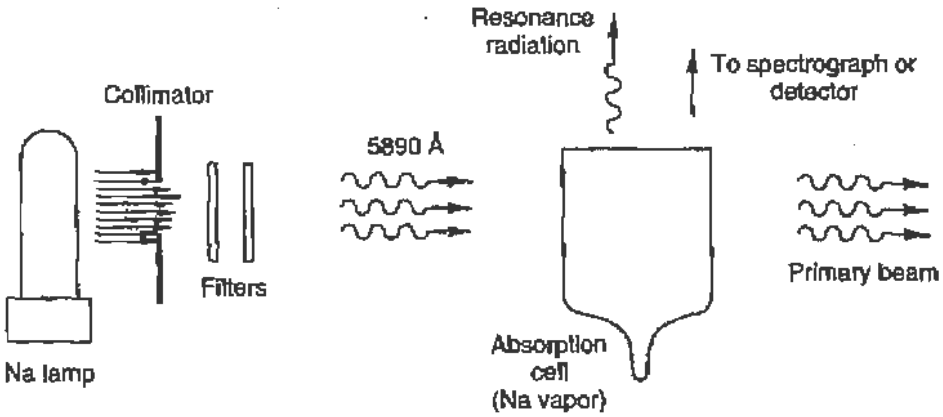


FIGURE 9.9 The arrangement of an optical (atomic) resonance radiation experiment. Here the sodium D lines are incident on a cell containing sodium vapor; it is then possible to observe, at right angles to the incident beam, the appearance of the D lines.

Let us note two facts: (1) Since the atom must be in state α when the radiation is incident, α is usually the ground state of the atom.[10] (2) The incident radiation must be exactly of the correct energy $h(\nu_\beta - \nu_\alpha)$ corresponding to the separation of levels α and β.

If we now try to observe in a similar manner resonance radiation, using a nuclear gamma ray (instead of the sodium D lines), we will obtain a negative result. This is a simple consequence of energy and momentum conservation, which produces a negligible effect in the case of an atomic line. To understand this, consider a system R originally at rest; R undergoes a transition from $\beta \rightarrow \alpha$, where the energy difference between states α and β is

$$E_\beta - E_\alpha = h\nu. \tag{9.15}$$

As a result of the transition, a quantum is emitted, which will carry away energy $h\nu_e$ and momentum $h\nu_e/c$; ν_e is to be determined. From Fig. 9.10a we see that to conserve momentum, the emitting system R must recoil with momentum $h\nu_e/c$; therefore it will have energy (nonrelativistically)

$$E_R = \frac{(h\nu_e)^2}{2mc^2}. \tag{9.16}$$

[10] The available intensities of visible radiation, the absorption cross section, and the density of the absorbers are all such that most of the atoms in the cell must be able to absorb (and re-emit) radiation in order to yield observable results. In very special cases, a metastable state, to which a large fraction of the atoms can be transferred (by some other means), can serve as state α.

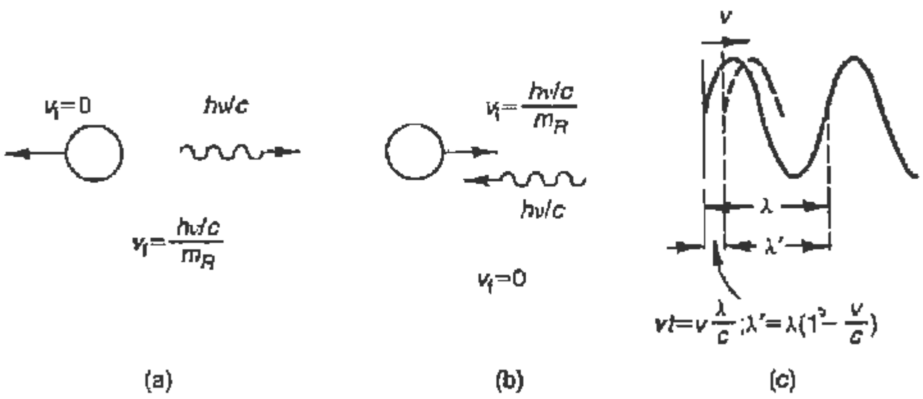


FIGURE 9.10 The effect of momentum conservation (recoil effects) in the emission and absorption of nuclear gamma rays. (a) A system R originally at rest emits a gamma ray $h\nu$; it must recoil with a velocity $v_f = (h\nu/c)/m_R$. (b) A system R moving originally with a velocity $v_i = (h\nu/c)/m_R$ absorbs a gamma ray $h\nu$; after the absorption the system will be at rest. (c) Derivation of the first-order Doppler shift for an observer moving with velocity v.

To balance energy, we must have

$$E_R + h\nu_e = h\nu,$$

leading to

$$h\nu_e = h\nu(1 - x + 2x^2 + \cdots), \tag{9.17}$$

where $x = h\nu/2mc^2$ will generally be small.

Similarly for a system R' originally at rest in order to be raised from level $\alpha \to \beta$, where $E_\beta - E_\alpha = h\nu$, it must absorb a quantum of energy

$$h\nu_c = h\nu(1 + x - 2x^2 + \cdots). \tag{9.18}$$

If the emitted quanta were strictly monochromatic, then it is clearly not possible for a free system R to absorb a quantum $h\nu_e$ emitted by a similar free system R', since $h\nu_a \neq h\nu_e$ (Fig. 9.11a).

We know, however, that spectral lines have a certain width[11] $\Delta\nu$; in Fig. 9.11b the emission and absorption lines are shown appropriately centered about $h\nu_c$ and $h\nu_a$, but with a width $\Delta\nu$. If then the two line shapes overlap, it is possible to have resonant absorption.

[11] The minimum or "natural width" of a line is determined from the lifetime τ of the transition $\beta \to \alpha$; from the uncertainty principle $\Delta E \Delta t \approx \hbar$, and thus $\Delta\nu \approx 1/\tau$. Other contributions include the "Doppler broadening" due to the thermal motion of the atom or nucleus, collisions, external perturbations or imperfections in a crystal lattice.

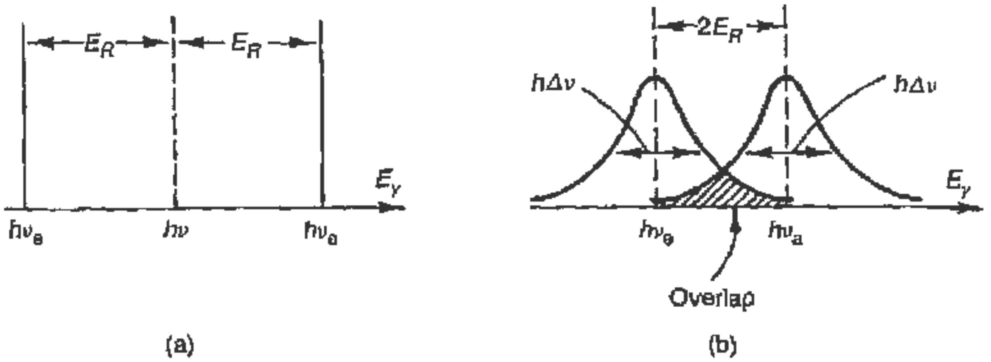


FIGURE 9.11 Indication of the energy shift of an emitted or absorbed gamma ray due to the recoil of the nucleus. (a) The situation when the line width is very narrow in comparison to the recoil energy; no resonant absorption can then take place under normal conditions. (b) The situation when the line width is on the same order as the recoil energy; note that resonance absorption can now take place and it will be proportional to the convolution of the two line shapes.

This is true for atomic systems: here $h\nu \approx 2$ eV, and for hydrogen $mc^2 \approx 10^9$ eV; thus $x \approx 10^{-9}$. The width of atomic spectra lines, however, is on the order of $\Delta\nu/\nu \approx 10^{-6}$. Thus

$$\left(\frac{\Delta\nu}{\nu} \approx 10^{-6}\right) \gg \left(\frac{h\nu}{2mc^2} \approx 10^{-9}\right).$$

For nuclear gamma rays, $h\nu \approx 10^4$–10^6 eV; also, in general, nuclear lifetimes are longer than those for atomic systems, so that

$$\frac{\Delta\nu}{\nu} \approx 10^{-10} - 10^{-15}.$$

Thus we see, in contrast to the situation for atomic systems,[12] that

$$\left(\frac{\Delta\nu}{\nu} \approx 10^{-10}\right) \ll \left(\frac{h\nu}{2mc^2} \approx 10^{-7}\right),$$

making resonance radiation impossible.

In the preceding discussion we assumed the that the emitting and absorbing nuclei were at rest. We could, on the other hand, think of imparting to the absorbing nucleus (by some means) enough velocity in a direction opposite to that of the quantum (Fig. 9.10b) so as to satisfy Eqs. (9.17) and (9.18).

[12] For example if $\tau \approx 10^{-9}$ s, then $\Delta E \approx 6 \times 10^{-7}$ eV. Further, nuclear gamma rays are subject to broadening influences much less than atomic lines.

For example, if $h\nu \approx 10^4$ eV, and the nucleus has $A \approx 100$, and we wish that

$$\frac{h\nu}{c} = mv \qquad h\nu c = (mc^2)v \tag{9.19}$$

we find for the velocity

$$v = \frac{3 \times 10^{10} \times 10^4}{100 \times 10^9} = 3 \times 10^3 \text{ cm/s.}$$

Such velocities can be obtained in the laboratory by placing the samples on the rim of a centrifuge and orienting the incoming beam toward one of the tangents. It then becomes possible to observe nuclear resonant absorption.

Nuclear resonant absorption would also occur if both the emitter and absorber were so massive that momentum could be balanced with negligible energy being given to the recoiling system, that is, if the denominator m in Eq. (9.16) became infinite. Indeed, R. Mössbauer showed in 1958 that for atoms bound in a crystal lattice, a nucleus does not recoil individually[13] but the momentum of the nuclear gamma ray is shared hy the entire crystal. This can be understood if we consider that the binding energies of the atoms in a lattice site are on the order of 10 eV, whereas the recoil energies, given by Eq. (9.16), are always less than 1 eV.

Since, however, the nucleus is now part of a larger quantum-mechanical system, there exists the possibility that the energy available from the deexcitation of the nucleus $\beta \rightarrow \alpha$ might not all be given to the gamma ray, but might be shared between the gamma ray and the lattice, in the form of vibrational energy. Lattice vibrations—the so-called emission of *phonons*—are a quantized process, and the lowest energy phonon that a single nucleus can emit has

$$E = kT,$$

where $T = \Theta_D$ is a characteristic temperature for the crystal, the *Debye temperature.* Thus, if the recoil energy of the free nucleus, as given by Eq. (9.16), is $E_R < k\Theta_D$, it is not possible for the lattice to become excited into a vibrational mode, and the total energy of the transition is taken by

[13] It is customary to say that "the nucleus does not *always* recoil individually," in order to account for the instances where the nucleus *transfers energy* to the lattice as explained in the following paragraph.

the gamma ray. The probability of recoilless emission of the gamma ray is then given by

$$f = \exp\left(-\frac{3}{2}\frac{E_R}{k\Theta_D}\right). \tag{9.20}$$

Equation (9.20) holds at absolute zero, and for finite temperatures we may use

$$f = \exp\left(-\frac{\langle x^2 \rangle}{\bar{\lambda}^2}\right). \tag{9.21}$$

Here $1/\bar{\lambda}^2 = (2\pi\nu/c)^2$ is the square of the wave number of the emitted gamma ray and $\langle x^2 \rangle$ is the mean square deviation of the atoms from their equilibrium position and is proportional to T. As an example, for the 14.4-keV line of ^{57}Fe,

$$E_R = 0.002\ \mathrm{cV} \qquad \text{and} \qquad \Theta_D = 490\ \mathrm{K};$$

hence

$$f = e^{-0.08} = 92\%.$$

We therefore see that in certain materials (^{57}Fe being the most suitable) the Mössbauer conditions are met; recoilless emission and absorption can take place, and consequently nuclear *resonance radiation* can be observed.

It has been explained earlier (Eq. (9.19)) that we could compensate for the recoil of the nucleus by moving the absorber in a direction opposite to the incoming gamma ray (so as to make the total momentum of the nucleus-plus-gamma-ray system zero). It follows then that if the absorption is recoilless, such motion of the absorber would *destroy* the resonance condition. In recoilless emission (absorption) the gamma ray has energy $E_\gamma = h\nu_0$ in the system, which is at rest with respect to the nucleus; if the nucleus is moving in the laboratory with a velocity v in the direction of the gamma ray, the laboratory energy of the gamma ray E'_γ is given by a Lorentz transformation

$$E'_\gamma = \frac{1}{\sqrt{1-\beta^2}}(E_\gamma + vp_\gamma) = E_\gamma \frac{1+\beta}{\sqrt{1-\beta^2}},$$

where $\beta = v/c$. For $\beta \ll 1$ we obtain to first order

$$\Delta E = E'_\gamma - E_\gamma = \beta E_\gamma \qquad \text{or} \qquad \frac{\Delta E}{E} = \beta = \frac{v}{c},$$

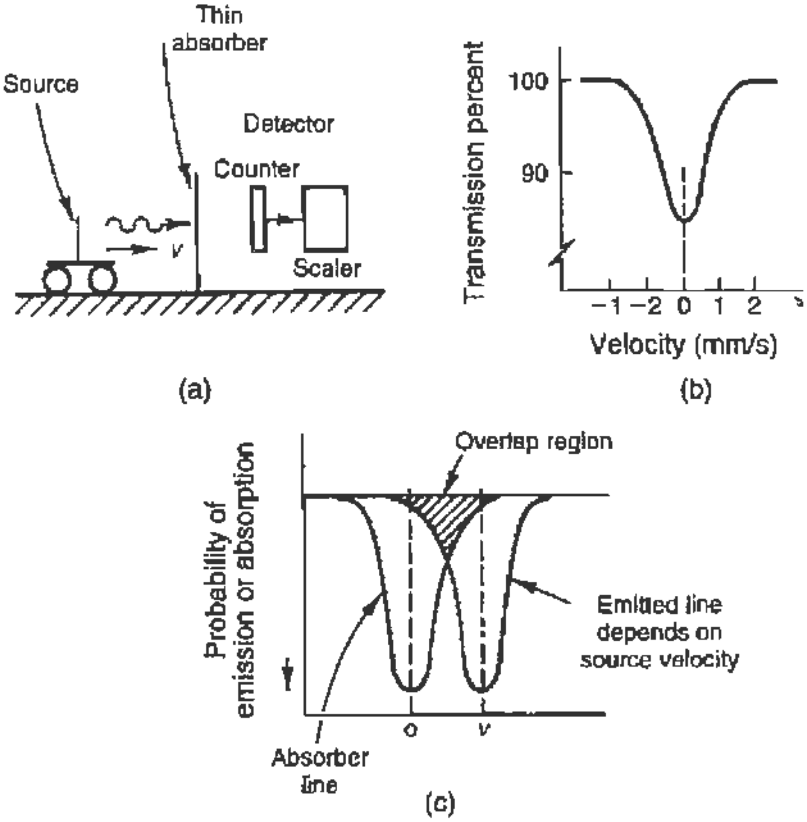


FIGURE 9.12 The Mössbauer resonant absorption experiment. (a) Diagrammatic view of the equipment. (b) The probability for transmission of a gamma ray as a function of the source (or absorber) velocity when no hyperfine structure is present. (c) The width of the transmission curve is a combination of the shape of both the source and absorber lines.

which, written as $\Delta\nu/\nu = v/c$, is the first-order Doppler shift of a wave emitted (absorbed) by a moving observer (Fig. 9.10c). To obtain a quantitative estimate we consider again the 14.4-keV line of ^{57}Fe, which has a lifetime $\tau \sim 10^{-7}$ s and hence $\Delta\nu/\nu = 4.5 \times 10^{-13}$. Thus, velocities on the order of $v = c(\Delta\nu/\nu) \sim 1.5 \times 10^{-2}$ cm/s will be sufficient to destroy the resonant absorption. Such velocities are easy to achieve and control in the laboratory. We therefore measure the transmission of the 14.4-keV gamma ray through an ^{57}Fe absorber as a function of its velocity. Alternatively we can leave the absorber stationary and move the source.

A possible experimental arrangement, indicated in Fig. 9.12a, consists of an ^{57}Fe source, an ^{57}Fe absorber that can be moved at a constant velocity,[14] and a detector for the 14.4-keV gamma rays; we measure the rate of transmitted gamma rays. At zero velocity the transmission is low because of

[14] The velocity, however, is varied in the course of the experiment.

resonant absorption; as the velocity of the absorber is increased, however, the resonance is destroyed and the transmission increases, leading to a typical curve as shown in Fig. 9.12b. We may think of the incoming gamma ray as scanning over the absorption line as a function of the velocity, and therefore the observed absorption is a measure of the convolution of the two lines as shown in Fig. 9.12c. In this way we "trace out" the natural line width for this nuclear gamma ray, and measure energy deviations of one part in 10^{13} ($v \approx 0.06$ mm/s). This represents a highly precise measurement and this is why the Mössbauer effect is an important tool in many physics applications.

9.3.2. The Apparatus and Some Experimental Considerations

In this laboratory the Mössbauer effect was observed using the 14.4-keV gamma ray of ^{57}Fe, which follows the decay, by electron capture, of ^{57}Co (see Fig. 9.13). Basically the apparatus required for the experiment consists of (Fig. 9.12) (1) the source (with or without appropriate collimation), (2) the absorber and a mechanism for moving the absorber or the source at constant speed, and (3) the detector for the 14.4-keV gamma ray. From Fig. 9.13 we note that the 14.4-keV line of interest will be accompanied by a 122-keV gamma ray as well as by a weaker 136-keV line. There is also a strong background present from the 6.5-keV X-ray of ^{57}Co, which follows the electron capture from the K shell. The source used was 1 mCi of ^{57}Co plated and annealed onto an ordinary iron backing.[15]

The detector is chosen so as to provide good efficiency and discrimination for the 14.4-keV gamma ray. A xenon–methane proportional counter, followed by a single-channel discriminator, was used. In Fig. 9.14, curve (a) gives the pulse-height spectrum of the gamma rays emitted by the source, while curve (b) gives the same spectrum after the gamma rays have traversed a 0.001 in. absorber. The shaded area represents the "window" selected on the discriminator, so that only gamma rays within these energy limits were recorded by the scaler.

The absorber in this case is usually a thin steel foil, but it should not exceed 0.001 in., since nonresonant scattering increases so much as to smear out the 14.4-keV line. Further, natural iron contains only 2.17% of ^{57}Fe, so that poor signal-to-noise ratios result. It is possible, however,

[15] Purchased from Nuclear Science and Engineering Co., P.O. Box 1091, Pittsburgh, PA.

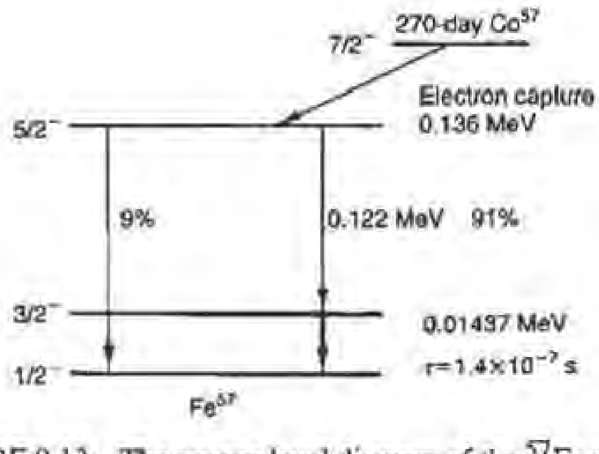


FIGURE 9.13 The energy-level diagram of the ^{57}Fe nucleus.

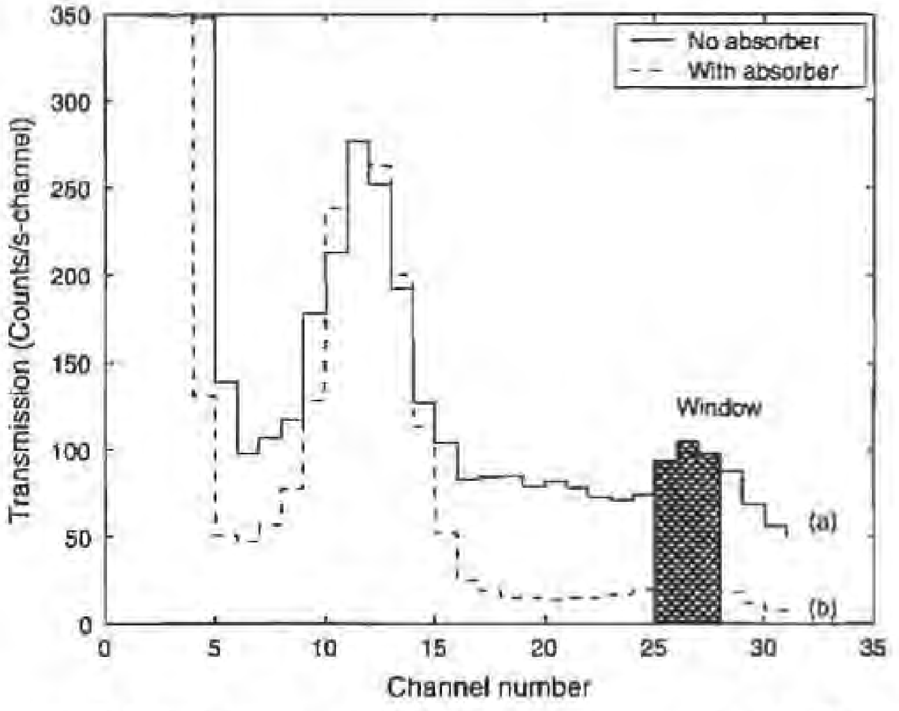


FIGURE 9.14 Pulse-height spectrum of the low-energy gamma rays of ^{57}Fe as obtained with a proportional counter. The solid curve has been taken without the absorber in place, whereas the dashed one has been taken with the absorber in place. The shaded region indicates the discriminator window used for observing the Mössbauer effect.

to obtain absorber samples enriched in ^{57}Fe, and in the present experiment, such a foil (of 1 cm^2 area) was used; the ^{57}Fe concentration was 91.2% and the thickness 1.9 mg/cm^2 (approximately 0.0001 in.).

The motion of the absorber can be achieved either by purely mechanical arrangements, or by a transducer of some type. Examples in the former

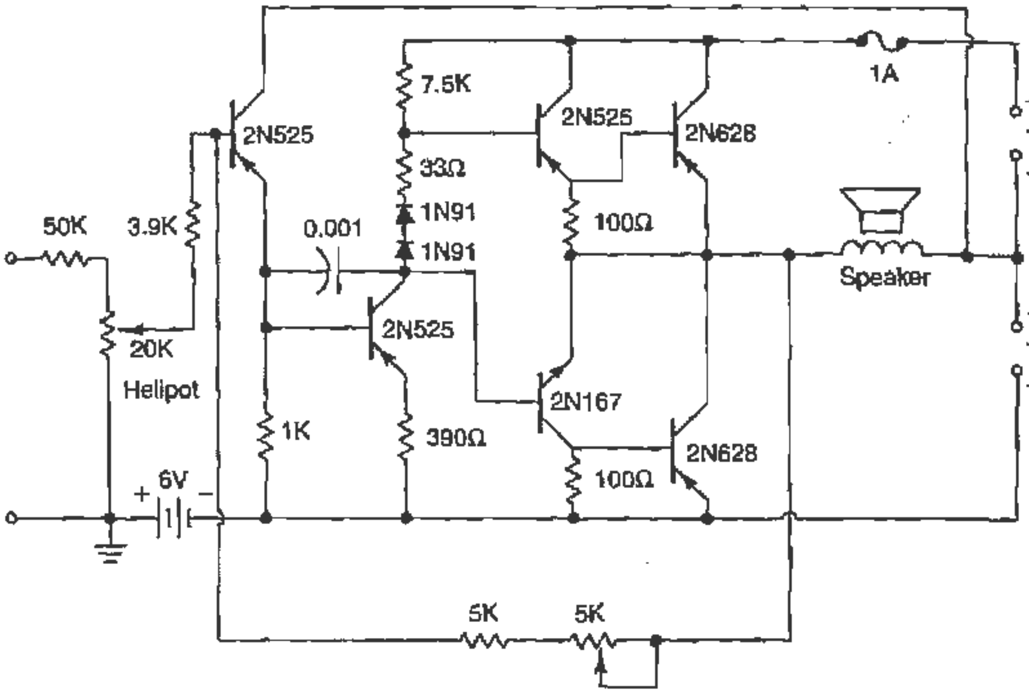


FIGURE 9.15 An amplifier circuit capable of driving a speaker coil for use in the Mössbauer experiment.

category are a plunger driven by an appropriately shaped cam (logarithmic spiral $r = k\theta$) or the rim of a wheel rotating about an axis that is not normal to the surface of the wheel. In all cases of mechanical motion, special attention must be paid to decoupling the vibrations of the driving motor from the absorber.

For the present experiment, a device of the latter category was chosen, namely, a loudspeaker driven by a sawtooth current (see Fig. 9.15). The source was mounted on the core of the speaker and the absorber was kept stationary. The driving waveform was obtained from the horizontal sweep of an oscilloscope after amplification.

To calibrate the speaker, a micrometer screw was mounted in a special manner above the speaker. By listening, the experimenter could discern when the screw touched the speaker, giving results to within ±0.003 cm out of a maximum travel of 0.2 cm. Assuming that the speaker is linear with current, the calibration shown in Fig. 9.16 was obtained. The small variation in solid angle with the change of source–detector distance does not affect the results obtained. It is also advisable to gate the scalers so as to count only during the linear part of the motion (and in the desired direction).

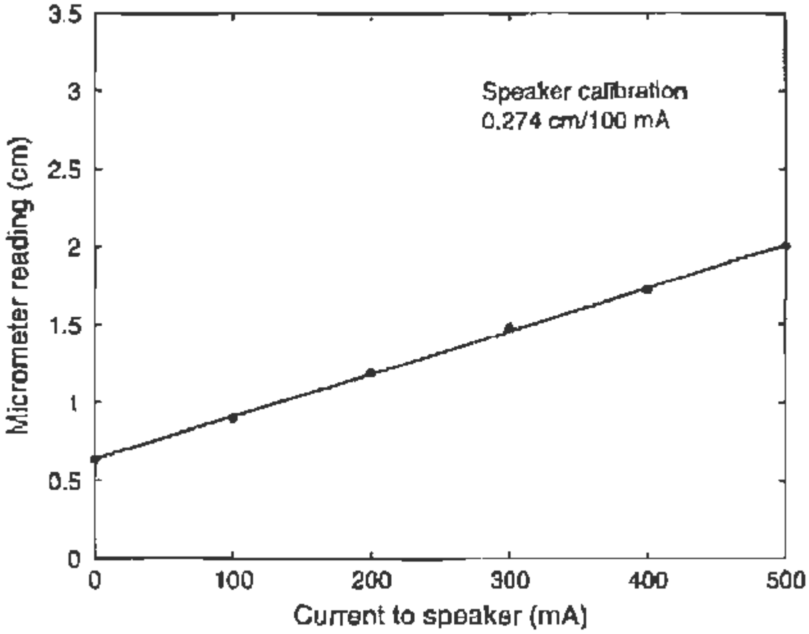


FIGURE 9.16 Velocity calibration of the speaker used to provide the motion of the source in the Mössbauer experiment.

9.3.3. Results and Discussion

In Fig. 9.17 the results obtained by a student are given; the abscissa gives the velocity of the source in millimeters per second, and the ordinate, the counting rate at the detector. It is clear that maximum absorption occurs at zero velocity, in accordance with the hypothesis of recoilless emission (and absorption) of the gamma ray and the conclusions reached in the previous sections.

The full-width at half-maximum for the zero-velocity peak as obtained from Fig. 9.17 is $\Gamma_{\text{app}} = 0.70$ mm/s. If the two curves shown in Fig. 9.12c are assumed to have a Lorentzian shape, then the apparent width Γ_{app} can be related to the true line width Γ through

$$\Gamma_{\text{app}}/\,\Gamma = 2.00 + \text{term correcting for absorber thickness.}$$

Thus we find that

$$\Gamma(14.4\ \text{keV}) \approx 0.30\ \text{mm/s}$$

and

$$\frac{\Delta\nu}{\nu} = \frac{\Gamma}{c} \approx 10^{-12},$$

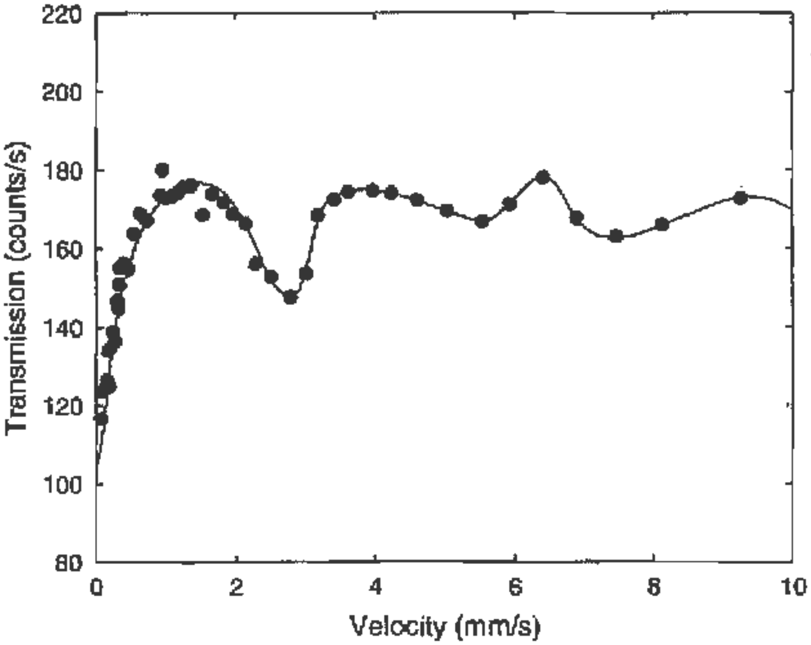


FIGURE 9.17 Results obtained for the Mössbauer effect of ^{57}Fe using a ^{57}Co source on ordinary iron backing, and an enriched ^{57}Fe absorber.

which is in fair agreement[16] with the accepted value of $\Delta\nu/\nu = 3 \times 10^{-13}$.

It is clear that in Fig. 9.17, apart from the zero-velocity peak, there also appear subsidiary peaks at $v = 2.5$, 5.5, and possibly also 7.5 mm/s. What is the origin of these peaks, so reminiscent of the hyperfine structure of atomic spectral lines?

Indeed this structure of the Mössbauer line is greatly dependent on the type of host material in which the absorber (or source) nuclei are embedded. In natural iron, there exist strong magnetic fields at the site of the nuclei; as a result, the nuclear energy levels are split, giving rise to a "Zeeman effect" for the nucleus.[17] Figure 9.18b shows the splitting of the $\frac{3}{2}$ excited state

[16] Most of the discrepancy can be traced to the considerable thickness of the absorber. The probability for interaction is given by

$$P = \sigma_0 f a (N_0/A) t,$$

where t = absorber thickness $\approx 2 \times 10^{-3}$g/cm^2, $N_0/A = 6 \times 10^{23}/57 \approx 10^{22}$/g, σ_0 = the Mössbauer absorption cross section = 1.5×10^{-18} cm^2, f = probability for recoilless absorption, approximately 1, and a = concentration of the resonantly absorbing nuclei in the sample, approximately 1. Hence, for the present case, $P \approx 30$!

[17] See Section 6.2 for a detailed discussion of the Zeeman effect.

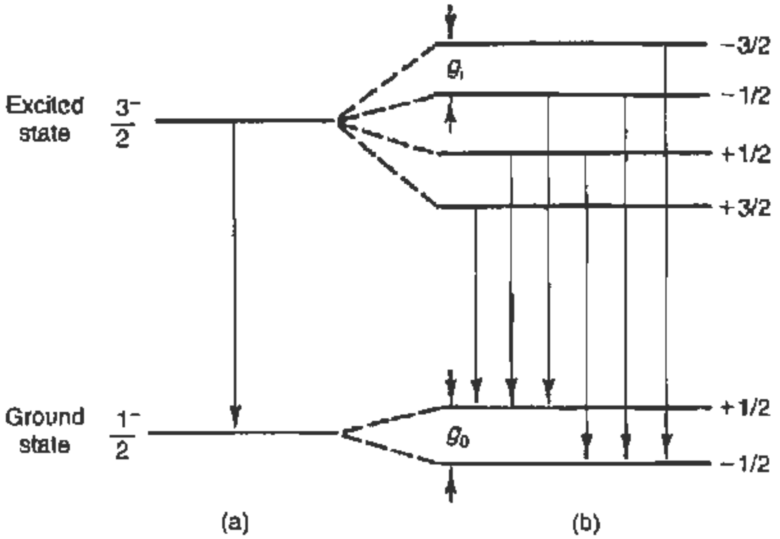


FIGURE 9.18 Hyperfine structure splitting of the nuclear energy levels of ^{57}Fe. (a) When stainless steel is used, the levels are not split. (b) In ordinary iron, however, both levels are split, giving rise to a hyperfine structure with six components.

and the $\frac{1}{2}$ ground state of ^{57}Fe, and consequently the 14.4-keV line has six hyperfine structure components. Figure 9.18a shows the same levels for stainless steel, where no splitting occurs.

If both the source and absorber are not split, then clearly only a single peak will be observed, as in Fig. 9.12b. If the source is not split, but the absorber is, then as a function of velocity we will "scan" with the single line over the hyperfine structure pattern of the absorber. In this case there is no absorption at zero velocity (see Fig. 9.19a). Finally, if both the source and absorber are split, a complicated pattern emerges, depending on the degree of overlap of the individual components as the two hyperfine structure patterns are shifted one over the other; however, maximum absorption occurs at zero velocity (see Fig. 9.19b).

In the experiment that yielded the data of Fig. 9.17, both the source and the absorber were split, so that a pattern of the type shown in Fig. 9.19b was obtained. Table 9.2 gives the relative intensities and known positions of the peaks as well as the positions obtainable from the results of Fig. 9.17.

The apparent discrepancies in the known and observed positions are due in part to a small velocity calibration error. Materials like stainless steel, potassium ferrocyanide, sources made by diffusing ^{57}Co into chromium metal, do not exhibit structure in the 14.4-keV line and give simple patterns. In Table 9.3 we summarize some of the numerical values pertinent to the Mössbauer effect in ^{57}Fe.

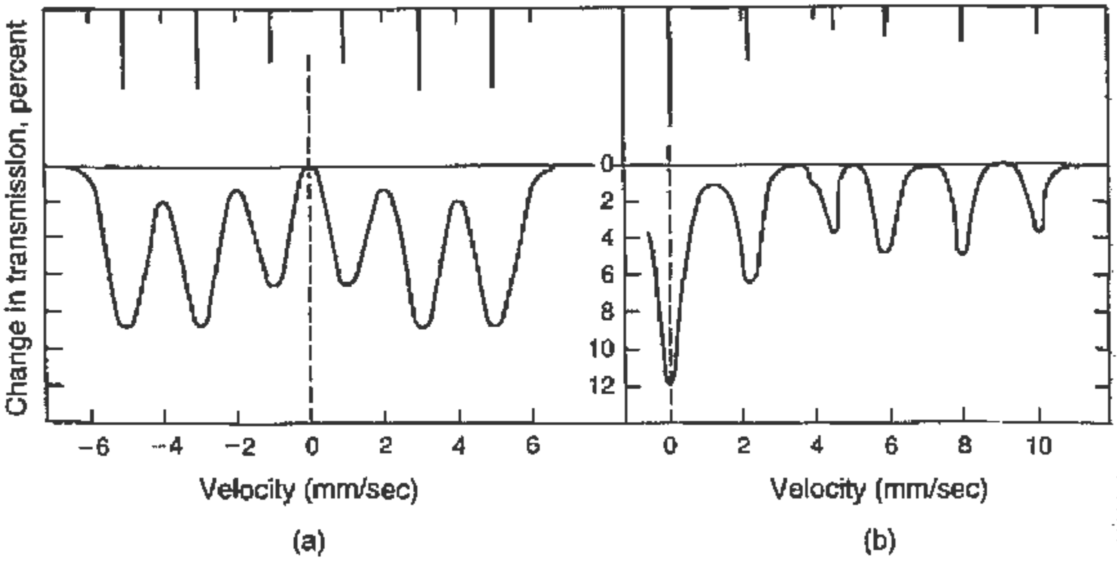


FIGURE 9.19 The expected pattern of the Mössbauer line when splitting of the levels takes place. (a) Either the source or absorber is split; note that the Mössbauer line is split into six components and no absorption takes place for zero velocity. (b) When both source and absorber are split a complicated pattern results with maximum absorption at zero velocity.

TABLE 9.2 Position and Amplitude of Mössbauer Peaks in ^{57}Fe, Including the Experimental Results

Peak	Amplitude	Position (mm/s)	Observed position (mm/s)
0	7	0	0
1	4	2.2	2.75
2–3	1.5	4.3	5.5
4	2.5	6	7.6 (?)
5	3	8	—
6	2	10	—

TABLE 9.3 Some Numerical Values Pertinent to the ^{57}Fe Mössbauer Line

Transition energy	$E_\gamma = 14.4 \times 10^3$ eV
Internal conversion coefficient	$\alpha = e/\gamma = 15$
Lifetime	$t = 1.4 \times 10^{-7}$ s
Relative width	$\Delta\nu/\nu = 3 \times 10^{-13}$
Recoil energy of free nucleus	$E_\mathrm{R} = 0.19 \times 10^{-2}$ eV
Debye temperature (Mössbauer)	$\Theta_\mathrm{D} = 490$ K
Probability for recoilless transition at room temperature	$f = 0.80$
Cross section for resonant absorption	$\sigma_0 = 15 \times 10^{-19}$ cm^2
Natural abundance of ^{57}Fe	2.17%

A very complete description of the Mössbauer effect, including reprints of the most important papers, will be found in H. Frauenfelder's *The Mössbauer Effect* (W. A. Benjamin, New York, 1962); this reference should be fully adequate until the student finds it necessary to consult the current literature.

9.4. DETECTION OF COSMIC RAYS

9.4.1. Flux, Composition, and Detection of Cosmic Rays

The earth is continuously bombarded by a flux of high-energy particles that originate outside of the solar system. These are mainly protons but the primary cosmic ray flux also contains a fraction of light nuclei. When these particles reach the earth's atmosphere they cause nuclear interactions so that at sea level we observe only the final products of the nuclear cascade.[18]

The interaction of the primary protons with the oxygen and nitrogen nuclei of the atmosphere results in the production of secondaries including unstable particles such as $\pi^{\pm}$ mesons, $K^{\pm}$ mesons, and others. These in turn decay by the weak interaction into lighter particles, including muons, electrons, and neutrinos. Electrons and high-energy γ-rays also interact rapidly, giving rise to electromagnetic showers as discussed in Section 8.2.6. Since the earth's atmosphere is equivalent to ten nuclear interaction lengths, all strongly interacting particles are absorbed before reaching sea level.[19] What is observed (at sea level) is a "hard component" consisting of $\mu^{\pm}$ (muons) and a "soft component" consisting of $e^{\pm}$ and low-energy γ-rays.

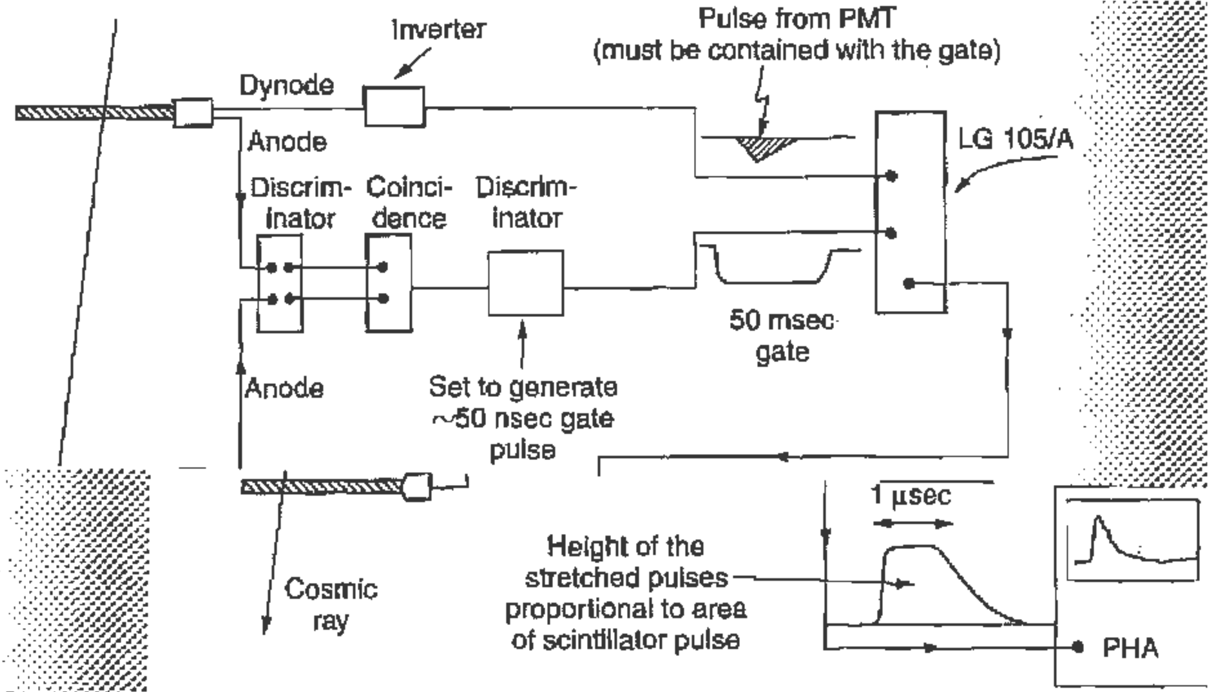
The total flux per unit solid angle around the vertical, crossing unit horizontal area is

$$1.1 \times 10^2/\text{m}^2\text{-sr-s}, \tag{9.22a}$$

where 75% of the flux is in the hard component. The angular distribution is approximately $\cos^2\theta$ (with $\theta = 0$ at the zenith). It is also useful to know

[18] A good reference on cosmic rays in general can be found online from the Particle Data Group at http://pdg.lbl.gov in the "Reviews" section under "Astrophysics and Cosmology."

[19] Some of these unstable particles were first observed and studied at high mountain altitudes or by baloon-borne detectors. Today such subnuclear particles are produced profusely by particle accelerators, but cosmic rays are still used for the study of the very highest energies.


Inverter
Pulse from PMT
(must be contained with the gate)
Dynode
Anode
LG 105/A
Discrim-inator
Coinci-dence
Discrim-inator
50 msec gate
Anode
Set to generate ~50 nsec gate pulse
1 µsec
Cosmic ray
Height of the stretched pulses proportional to area of scintillator pulse
PHA

isions for
) are also

(9.22b)

igh side

ıcidence
ove the
ıe coun-
angular
lvantage
f a third
:, but is
ɔbserves
he same
counter.

ıave been

FIGURE 9.20 Typical layout of a cosmic ray telescope and electronics. Pro\ measuring the pulse height in one of the counters (not discussed in the text shown.

that the total flux crossing unit horizontal area is

$$2.4 \times 10^2/\text{m}^2\text{-s}.$$

The mean energy of the muons is 2 GeV and falls off on the h as E^{-2}.

Cosmic ray muons can be easily detected by measuring the coi rate between two scintillation counters placed vertically one al other as shown in Fig. 9.20. By increasing the distance between t' ters one can restrict the solid angle acceptance and also study the distribution of the flux. Plastic scintillation counters have the a of large area so that the counting rate can be several per second. counter is placed in coincidence with the two-counter telescop located physically in a different location (as in Fig. 9.21) one still coincidences.[20] These occur because *several* cosmic rays arrive at time over the area covered by the telescope and the third "roving"

[20] These are true coincidences after any accidental effects (Section 9.5.1) subtracted.

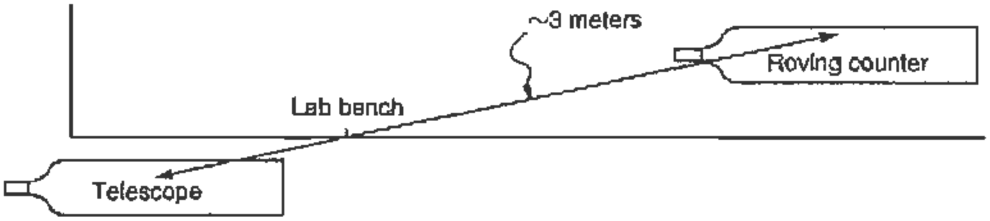


FIGURE 9.21 Arrangement of counters for measuring cosmic ray air showers (top view).

Namely a "shower" of cosmic rays occurred. One finds that the rate for such showers is 1/300 of the telescope rate, given a typical counter area of 0.2 m^2 and a displacement of 3 m.

We will describe an experiment in which cosmic ray muons are also detected by simply using a 5-gal tank of liquid scintillator, viewed by a 2-in. photomultiplier tube. Muons traversing the tank give a large signal so that it is possible to use the singles rate, without the need to form coincidences. However, the PMT high voltage and the discriminator must be set carefully. The dimensions of the tank are $d = 28$ cm diameter and $h = 35$ cm height from which we can estimate an *effective* horizontal area of $2 \times [\pi(d/2)^2] \sim 0.12\,\mathrm{m}^2$. The singles rate is of order 25/s, in reasonable agreement with Eq. (9.22b).

9.4.2. Time of Arrival of Random Events

The arrival of cosmic rays is a random process,[21] so we expect it to follow the distributions discussed in Chapter 10. In particular when the expected number of events in a given time interval is small, the observed number should obey the Poisson distribution. Let r be the average event rate, namely the average number of events per unit time. Then the probability of observing n events in the time interval t is

$$P(n,t) = \frac{(rt)^n e^{-rt}}{n!}. \tag{9.23}$$

From Eq. (9.23) we recover the differential probability for an event ($n = 1$) to occur in the differential interval dt. Since ($dt \to 0$), Eq. (9.23) reads

$$dP = P(1, dt) = r\,dt. \tag{9.24a}$$

[21] This is of course also true for the decay of radioactive nuclei.

Similarly the probability that no events ($n = 0$) occur in the interval t is

$$P(0, t) = e^{-rt}. \tag{9.24b}$$

We can test this proposition by measuring the distribution of the time *between* the arrival of adjacent events. A time interval t between events is defined in this case by requiring *no event* for the interval t and an event at the time t (in the differential dt). Thus the distribution is given by the product of Eqs. (9.24a), (9.24b), which we write in the form[22]

$$q_1(t) = \left.\frac{dP}{dt}\right|_{m=1} = re^{-rt}. \tag{9.25}$$

It is interesting that the above distribution is exponential; namely, short time intervals t between adjacent events are much more probable[23] than longer ones.

The result for the case $m = 1$ can be generalized for the time interval between every second event ($m = 2$), every mth event, etc. The derivation is given in Section 10.5.3 (see Eq. (10.75)), and we obtain

$$q_m(t) = r\,\frac{(rt)^{m-1}e^{-rt}}{(m-1)!}. \tag{9.26}$$

As m increases, the distribution tends to a gaussian centered at $t = mr$. Of course one could also test Eq. (9.23) directly by measuring how often one, two, etc., events are found within a fixed interval t. However, measuring the distribution of the time intervals between event arrivals, as done here, is by far more practical and efficient.

Data are acquired by recording the time of arrival of every muon in a computer file. Since the mean time between counts is ~40 ms, a precision of 0.1 ms is sufficient and can be easily provided by the computer clock. The file can then be analyzed by sorting the time intervals between adjacent pulses ($m = 1$) in time bins of 0.8 ms width. The same data are next analyzed by sorting the intervals for different values of m in correspondingly longer time bins.

Results obtained by a student for $m = 1$ are shown in Fig. 9.22, for $m = 3$ in Fig. 9.23, and for $m = 100$ in Fig. 9.24. One notes how the distribution becomes narrower as m increases. Namely the interval between every

[22] We imply that the second count arrives after the first one with a delay between t and $t + dt$.

[23] This justifies the old proverb that one calamity is always followed by a second one. See W. Bothe, *Phys. Zeit.* **37**, 520 (1936).

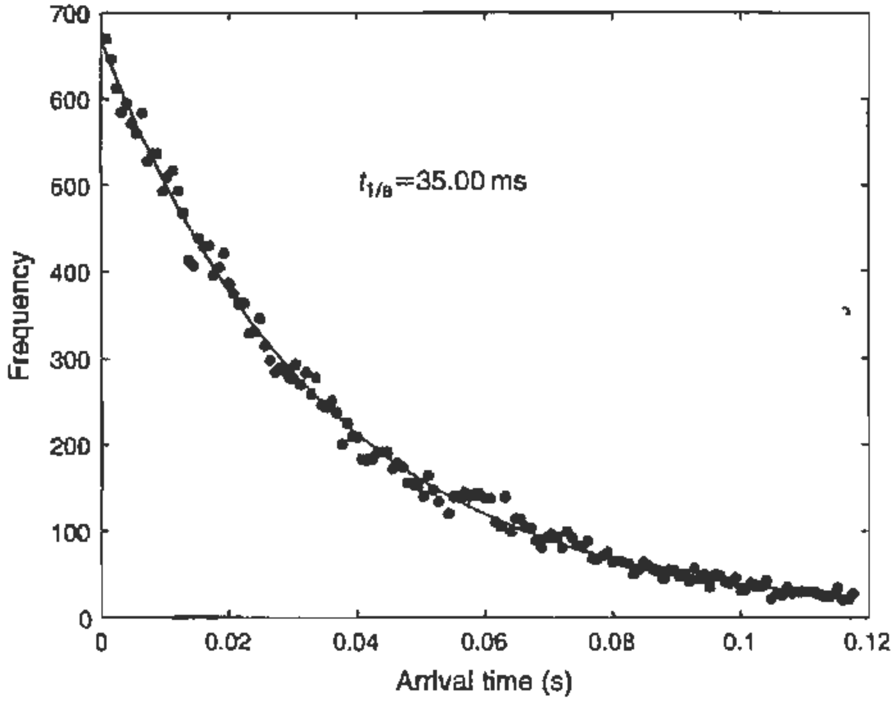


FIGURE 9.22 Distribution of the time between the arrival of two cosmic ray counts. The fit is the Poisson distribution for $m = 1$.

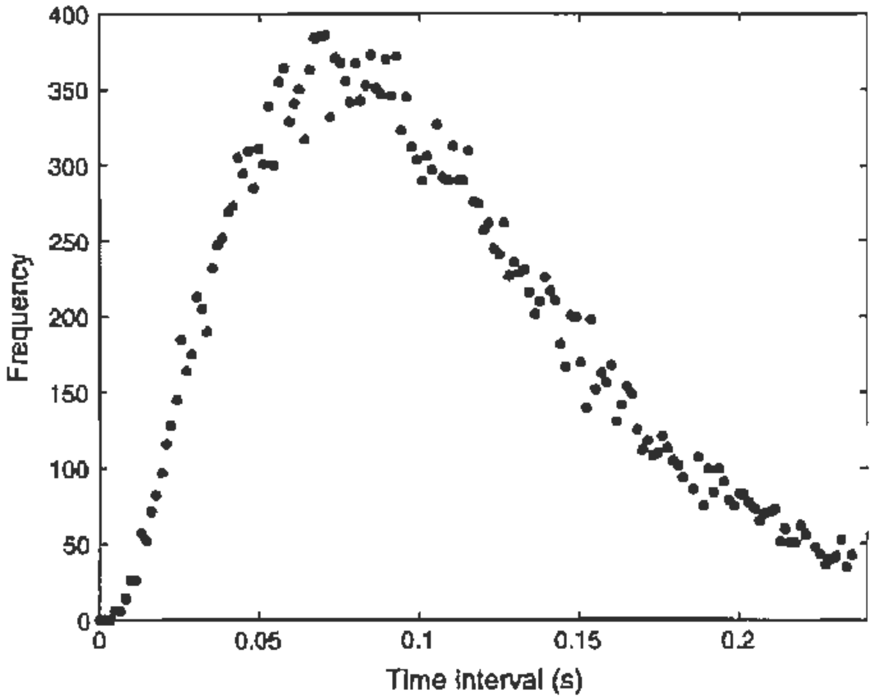


FIGURE 9.23 As described in the legend to Fig. 9.22 but for $m = 3$.

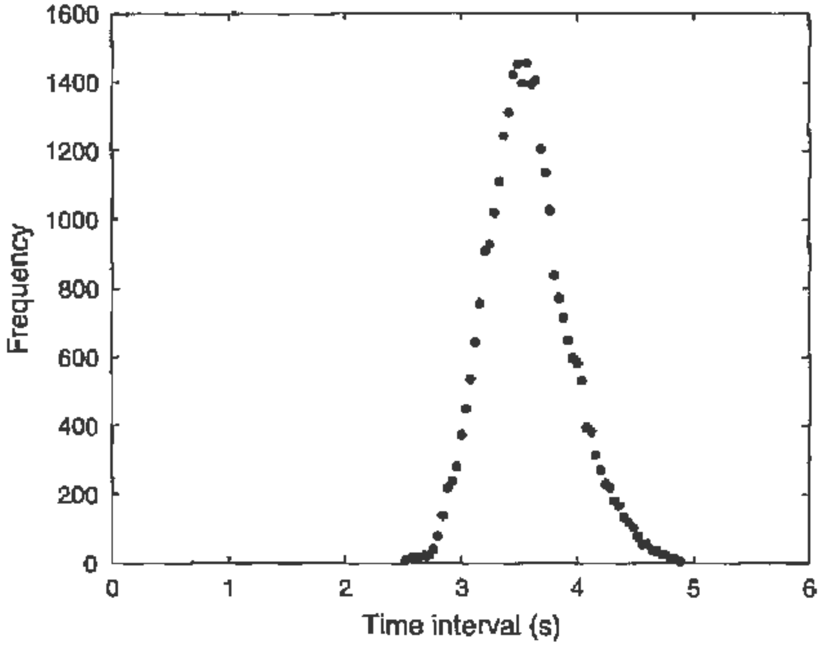


FIGURE 9.24 As described in the legend to Fig. 9.22 but for $m = 100$. Note that the distribution is centered at a mean time $t \approx 3.56$ s, where $t = (m - 1)/r \approx 100/r$.

100 events is much more "stable" (relative to its mean value) than between every second event. As can be seen from Eq. (9.24) the distributions for $m > 1$ have a maximum ($dq_m/dt = 0$) at

$$t = \frac{m-1}{r}. \tag{9.27}$$

Thus, from the location of the peak in the distribution we can obtain the average rate. We find that for the data shown in Figs. 9.23, and 9.24

$$\begin{aligned} m &= 3, & t_{\max} &= 0.073 \text{ s}, & r &= 27.5 \text{ /s} \\ m &= 100, & &= 3.56 \text{ s}, & &= 27.7 \text{ /s}. \end{aligned}$$

Furthermore, a fit to the exponential for $m = 1$ (see Fig. 9.22) yields $t_{1/e} = 3.50 \times 10^{-2}$ s, or $r = 28.6$/s in agreement with the average rate.

9.4.3. Measurement of the Mean Life of the Muon

The muon is not stable but decays into an electron, a neutrino, and an antineutrino:

$$\begin{aligned} \mu^+ &\to e^+ + \nu_e + \overline{\nu}_\mu \\ \mu^- &\to e^- + \overline{\nu}_e + \nu_\mu. \end{aligned} \tag{9.28}$$

The mean life, or lifetime, (i.e., the inverse of the decay rate) for this process is of order 2.2 μs, and thus the decay is easily detectable for muons at rest. The neutrinos are not observable but the electron (or positron)[24] is energetic enough to give a clear signal of the decay. The mass of the muon is

$$m_\mu = 105.65\ \text{MeV}/c^2,$$

approximately 200 times the electron mass. The maximum energy of the electron occurs when the two neutrinos recoil against it as shown in Fig. 9.25a. This corresponds, in the rest frame of the muon, to

$$E_e(\text{max}) \simeq \frac{1}{2}\, m_\mu c^2 = 53\ \text{MeV}.$$

The energy spectrum of the electrons from muon decay is shown in Fig. 9.25b.

The long lifetime for muon decay indicates that the decay does not proceed through the strong (nuclear) interaction but rather through the weak interaction responsible for the "β-decay" of nuclei. However, the process of Eq. (9.28) is very important because it involves only leptons (no strongly interacting particles participate) and thus can be used unambiguously to

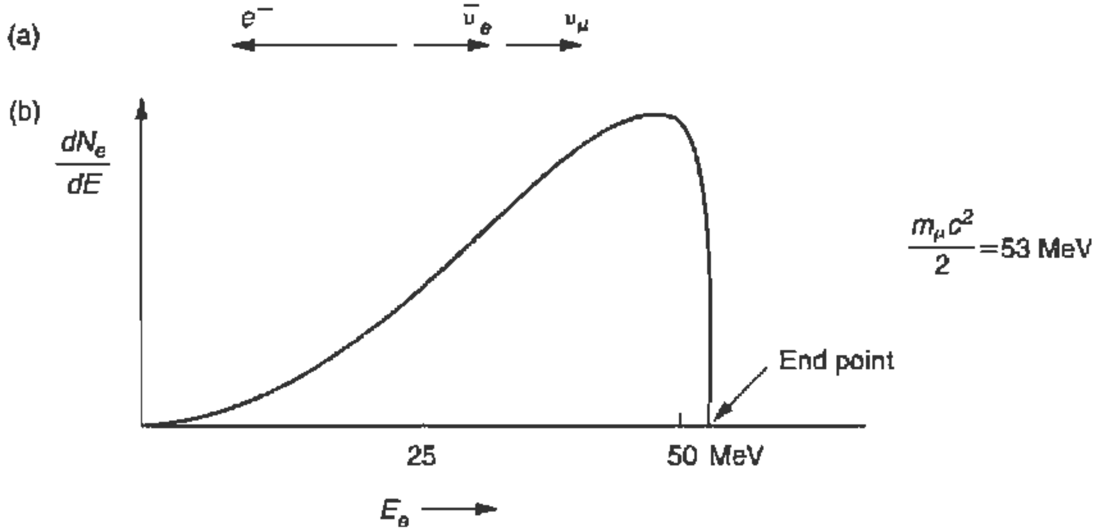


FIGURE 9.25 (a) Configuration of the particles in μ-decay for obtaining the maximum electron energy. (b) The energy spectrum of the electrons from μ-decay.

[24] To save words we will speak only of the electron even though we mean either e^- or e^+.

calculate the Fermi weak interaction constant G_F. The mean life of the muon is given by

$$\frac{1}{\tau} = \frac{1}{\hbar}\frac{G_F^2}{(\hbar c)^6}\frac{(m_\mu c^2)^5}{192\pi^3}. \tag{9.29}$$

Precise measurements of the muon mean life yield

$$\tau_\mu = (2.19703 \pm 0.00004) \times 10^{-6} \text{ s} \tag{9.30a}$$

and through Eq. (9.29) the value of the Fermi constant[25]

$$\frac{G_F}{(\hbar c)^3} = 1.1664 \times 10^{-5} \text{ GeV}^{-2}. \tag{9.30b}$$

We will measure the decay of muons that have come to rest in the liquid scintillator tank. Muons lose approximately 2 MeV of energy for each gram per squared centimeter of material that they traverse. Thus we expect that muons entering the 35-cm-high liquid scintillator tank with energy $E_\mu \lesssim 50$ MeV will stop in the tank. The fraction of muons that do stop is of order 0.3% of the flux going through the tank. Thus, the stopping rate is $R_S = 0.077$/s, or 4.6 muons/min. This is adequate to obtain good statistical accuracy for the mean life value in a reasonable time interval.

The experimental arrangement is shown in Fig. 9.26. When a muon enters the tank the PMT gives a signal, which is amplified and then discriminated. This pulse is used to start a "time-to-amplitude converter" (TAC) circuit. If the muon stops in the tank, then the decay electron will give a second signal within a time interval of a few mean lives. The second pulse is used to stop the TAC, and the time interval between start and stop is directly read out. The 60-ns delay in the start signal is to make sure that no pulse will be on the stop line when the start arrives. Commercial electronic modules can be used to achieve the logic indicated in Fig. 9.26. A GT200 computer card, designed by Professor D. Hartill of Cornell University,[26] performed the TAC functions and stored the data in a file in the computer memory. If no stop arrives within $\Delta t = 25$ μs, which corresponds to ~10

[25] Note that in contrast to the fine structure constant $\alpha = e^2/\hbar c$, which is dimensionless, $G_F/(\hbar c)^3$ has dimensions of inverse energy squared. In fact $G_F/(\hbar c)^3$ has the approximate value of $1/(M_W c^2)^2$, where M_W is the mass of the vector bosons that mediate the weak interactions.

[26] One can now purchase commercial versions of TAC cards to perform the required functions.

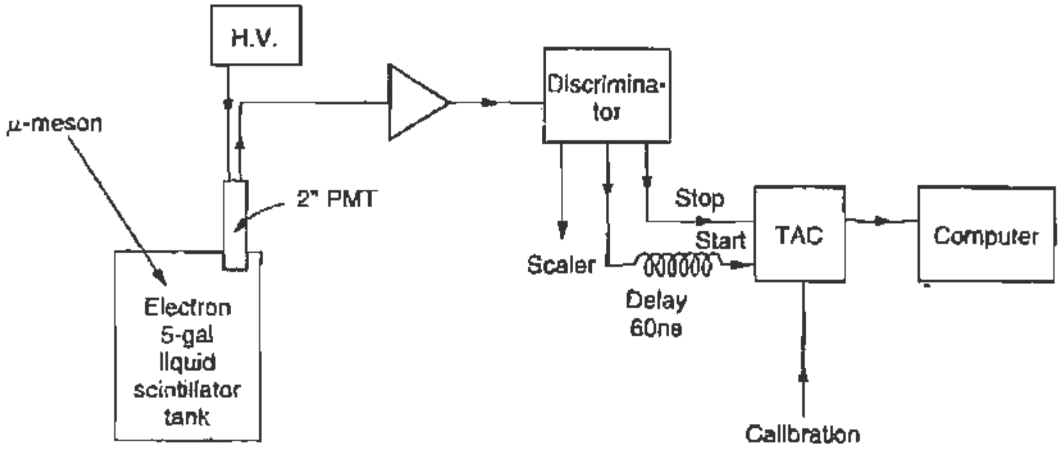


FIGURE 9.26 Block diagram of the electronics for measuring muon decay.

mean lives, the TAC is reset and the start pulse ignored. To calibrate the TAC one applies a fixed frequency (oscillator) signal to the discriminator input.

If the singles rate is too high, then the stop pulse may not be due to the decay of the muon that started the TAC, but to a different muon entering the tank. We call such events "accidental stops," and we can estimate their rates as follows. The singles rate is $r = 25/\text{s}$, so that using the Poisson distribution of Eq. (9.23) for $n = 2$ and $t = \Delta t = 25$ μs we find for the accidental rate

$$R_a = \frac{P_a(n = 2,\ \Delta t)}{\Delta t} = 7.8 \times 10^{-3}\ \text{s}^{-1}. \qquad (9.31)$$

This is ten times smaller than the stopping rate R_S, and does not affect the determination of the mean life as discussed later.

Data obtained by a student are shown in Fig. 9.27. The data were accumulated over five days and yielded $N_S = 32{,}000$ stops in 6921 min. The very early events, $t < 0.25$ μs, were discarded, leaving a sample of 30,069 events displayed in 100 bins each 0.25 μs wide. The data for $t \lesssim 5$ μs show an exponential drop-off, as expected, and in this region are well fitted by

$$N(t) = N_0 e^{-t/\tau} \qquad 0.25 < t < 5\ \mu\text{s}.$$

In contrast the data for late times, $t > 15$ μs, are flat and are well fitted by a constant

$$N(t) = C \qquad 5 < t < 25\ \mu\text{s}.$$

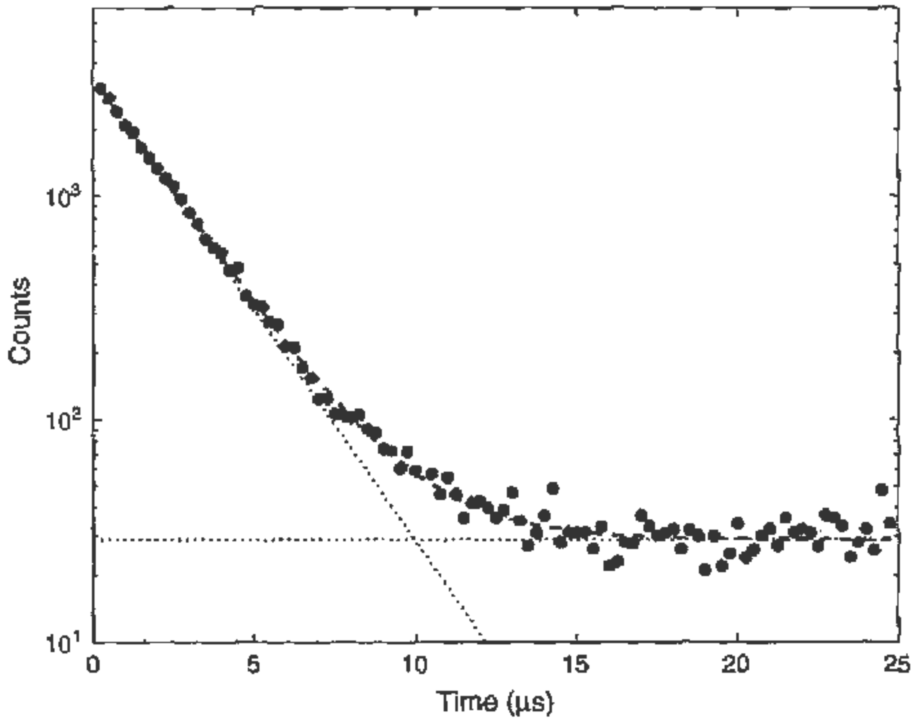


FIGURE 9.27 Data for 30,000 muon stops. The bin size is 0.25 μs, and the fit to the data including an exponential decay and a constant background term are shown.

A combined fit[27] of the form

$$N(t) = N_0 e^{-t/\tau} + C \tag{9.32}$$

yields $\tau = 2.088 \pm 0.016\,\mu\text{s}$, $N_0 = 3410$, and $C = 28.8$, and an excellent $\chi^2 = 0.909$ per degree of freedom. The contributions of the two terms of the fit are also indicated by the dashed lines in the graph.

We briefly discuss the background level. Since there are 100 channels, the total accidental count is $N_\text{a} = 2880$, and thus the accidental rate is $R_\text{a} = N_\text{a}/6921\ \text{min} = 6.9 \times 10^{-3}/\text{s}$ in agreement with our estimate of Eq. (9.31). One recognizes that the background does not affect the measurement until

$$N_0 e^{-t/\tau} \sim C.$$

This occurs when $t/\tau \sim 4.71$, which allows for a fairly long "lever arm" to determine τ_μ.

Our value for the mean life is in close agreement with the accepted value as given in Eq. (9.30a). The agreement is even closer because the

[27] See Section 8.6.2.

measured value for τ_μ must be corrected for the following effect. When negative muons stop in matter, there is a finite probability that the μ^- will be absorbed by a proton in the nucleus, leading to a "capture" reaction:

$$\mu^- + Z \rightarrow (Z-1)^* + \nu_\mu.$$

Thus the effective mean life is shortened and given by

$$\frac{1}{\tau_e} = \frac{1}{\tau_\mu} + \frac{1}{\tau_c},$$

where $1/\tau_\mu$ and $1/\tau_c$ are the rates for decay and capture, respectively. As a result the observed mean life is shorter; for mineral oil (the capture occurs mainly on carbon nuclei) and for the μ^-/μ^+ composition of cosmic rays this correction is approximately 4%. Therefore, the corrected measured value in this experiment is

$$\tau_\mu = 2.172 \pm 0.017\ \mu\text{s}. \tag{9.33}$$

The error shown in Eq. (9.33) is only statistical and does not include systematic effects, in particular any uncertainty in the TAC calibration.

9.5. γ–γ ANGULAR CORRELATION MEASUREMENTS

9.5.1. General Considerations

We will now discuss the measurement of the correlation in angle between two gamma rays that were emitted simultaneously from the same source. The origin of these gamma rays is frequently the cascaded decay of a nucleus, as in the case of ^{60}Ni (^{60}Co) already discussed in Chapter 8. (See Fig. 8.20.) We reproduce in Fig. 9.28 the decay scheme of this nucleus and note that the 1.333-MeV gamma ray follows the 1.172-MeV gamma ray, the lifetime of the intermediate state being only about 10^{-12} s, so that for all practical purposes the two gamma rays are coincident.

The fact that these two gamma rays are correlated in angle can be understood from the following general argument: the first gamma ray will have an angular distribution with respect to the spin axis of the nucleus; thus its observation at a fixed angle $\theta = 0$ conveys information about the probability of finding the spin at some angle ψ with respect to the

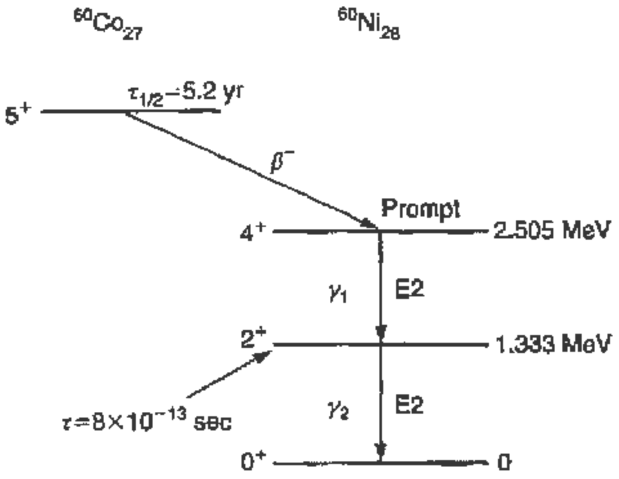


FIGURE 9.28 Nuclear decay scheme of ^{60}Co by beta decay to ^{60}Ni and subsequent deexcitation of the ^{60}Ni nucleus to its ground state by the emission of two cascaded gamma rays.

direction $\theta = 0$. Now the second gamma ray also has some angular distribution about the spin axis that now is known to be at ψ. Thus the probability that the second gamma ray will be emitted at an angle θ can be found; this is called the angular correlation function $C(\theta)$. The time coincidence signal assures us that the two gamma rays have indeed come from the same nucleus and, therefore, are the two gamma rays of interest. A discussion of this correlation between cascaded gamma rays is presented in Section 9.5.4.

In ^{22}Na, the angular correlation arises from a much simpler mechanism. ^{22}Na is a positron emitter as is shown from the decay scheme of Fig. 8.22. The positrons are slowed down in a thin copper sheet with which we surround the source; the slow positrons are captured by the electrons of the copper to form positronium, which decays by the annihilation of the positron and electron into two gamma rays. The energy of these gamma rays is precisely 0.511 MeV, and since the center of mass of positronium was at rest, the two gamma rays must be directed in exactly opposite directions in order to conserve momentum as shown in Fig. 9.29.

Thus the angular correlation theoretically is given by

$$C(\theta) = \delta(\pi - \theta),$$

and the observed finite width is due to the resolution of the apparatus; obviously the two gamma rays are simultaneous. Since the ^{22}Na angular correlation is so sharp, it is frequently used for calibration purposes.

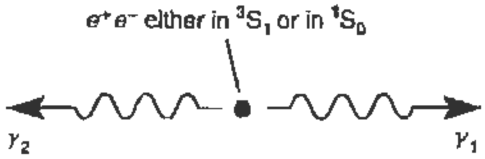


FIGURE 9.29 Capture of a positron by an electron to form positronium and the subsequent annihilation of the positron–electron pair into two gamma rays.

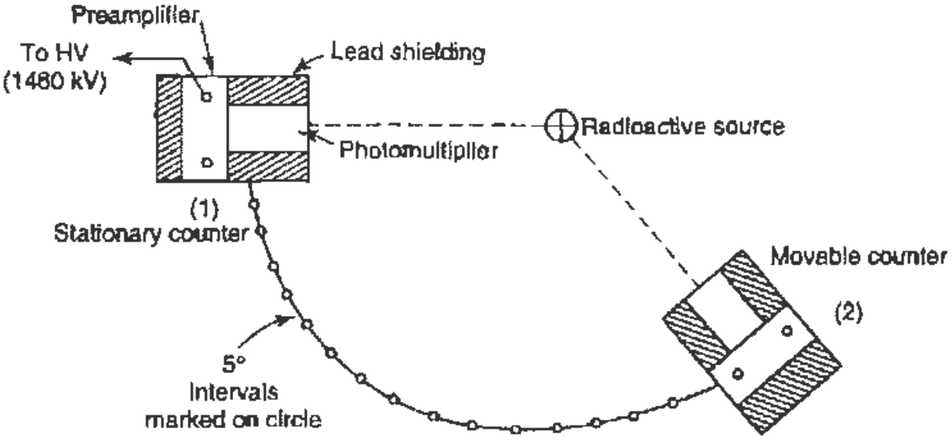


FIGURE 9.30 Apparatus that can be used for angular correlation measurements. Two scintillation crystals mounted on photomultipliers are protected by appropriate lead shielding. One counter assembly is fixed, whereas the other can be rotated about the position of the source.

Angular correlations may also be observed between beta and gamma rays, alpha and gamma rays, etc. This technique has proved very fruitful for the analysis of nuclear decay schemes and the assignment of spin and parity to excited nuclear levels.

We will describe a measurement of the gamma–gamma angular correlation of ^{22}Na and of ^{60}Co. The apparatus shown in Fig. 9.30 was used; it consists of two similar gamma-ray detectors placed at equal distances from the source; one detector is fixed and the other is free to rotate around the source, varying the angle θ between the detected gamma rays. The detector outputs are fed to a coincidence circuit, and the rate of coincident counts $C(\theta)$ is measured and compared with the theoretical correlation function.

It is important to measure $C(\theta)$ with the best possible resolution if the data are to be fitted with a polynomial in $\cos\theta$ of high order. It can be shown that $C(\theta)$ must be a polynomial in *even* powers of $\cos\theta$, the highest power being $2k$, where $k \leq I_b, l_1, l_2$ where I_b is the spin of the final nuclear state and l_1, l_2 is the angular momentum (multipole) of the emitted gamma rays. Frequently the experimental measurement may be restricted to the measurement of the anisotropy of the coincident gamma rays, that is,

$$\alpha = \frac{C(180°) - C(90°)}{C(90°)}.$$

The limiting factors in these experiments are two: (a) The coincidence rate must be high enough to allow statistically significant data to be accumulated in a reasonable time interval. To increase the coincidence rate a stronger source may be used, the solid angle may be increased, or the efficiency of the detector may be improved (if it has not been maximized already). (b) The accidental rate must be kept well below the coincidence rate; again it depends on source strength and solid angle, but also on the resolving time. Let $\Delta\Omega_1$ and $\Delta\Omega_2$ be the solid angles subtended at the source by detectors (1) and (2), and let ϵ_1 and ϵ_2 be their respective efficiencies. Then the "singles" counting rates are

$$\begin{aligned} R_1 &= N\Delta\Omega_1\epsilon_1 \\ R_2 &= N\Delta\Omega_2\epsilon_2, \end{aligned} \tag{9.34}$$

where N is the number of disintegrations per unit time of the source. If the two gamma rays are *uncorrelated* (or if the correlation is small, as happens mostly in nuclear decay), the coincidence rate[28] is

$$R_C = N\Delta\Omega_1\Delta\Omega_2\epsilon_1\epsilon_2. \tag{9.35}$$

For most experimental arrangements $\Delta\Omega_1 = \Delta\Omega_2$ and $\epsilon_1 = \epsilon_2$, so that we find for the accidental rate R_A,

$$R_A = R_1 R_2 \Delta t \tag{9.36}$$

$$= N^2\Delta\Omega^2\epsilon^2\Delta t, \tag{9.37}$$

and for the ratio of the accidentals to true coincidences,

$$\frac{R_A}{R_C} = N\Delta t. \tag{9.38}$$

We wish to keep this ratio small on the order of or smaller than 0.1. From Eq. (9.38) we see how important it is for correlation experiments to have a short resolving time; with $\Delta t = 10$ ns, a source as strong as 0.5 mCi may be used. We also note that the detector efficiency should be high,

[28] The efficiency of the coincidence *circuit* has been set to $\epsilon_c = 1$ as it should be.

since it enters Eq. (9.35) quadratically; however, the solid angle cannot be increased arbitrarily because this will destroy the angular resolution and wash out the correlation $C(\theta)$.

9.5.2. The Apparatus

The apparatus has been shown in Fig. 9.30, and we give here some additional details. The reader should, however, refer to Section 8.4.2, in connection with the instrumentation and techniques of gamma-ray detection. The detectors were NaI crystals 1 in. in diameter and 1 in. thick, mounted on RCA 6655 photomultipliers. Each was located 8 in. from the source. Both crystals are protected from scattered radiation with lead shielding, and the movable detector can be rotated about the center in 5° intervals.

The block diagram of the electronics is shown in Fig. 9.31, where the individual units are available from a number of vendors.[29] The units are interconnected with 50-Ω coaxial cable. Manuals accompanying the amplifier, discriminator, and coincidence modules should be consulted, especially to achieve the smallest possible resolution time. In the ensuing discussion we will assume that the circuits have been properly adjusted.

One of the outputs from each discriminator is fed to the coincidence module and a second output to a scaler capable of a peak rate of 10^5/s. The coincidence output is also fed to a scaler. In this way the "singles" in each channel and the "doubles" are counted. The delay between the two inputs to the coincidence circuit may be easily adjusted by inserting appropriate cable lengths between the discriminator and coincidence in one *or* the other of the channels. One foot of typical 50-Ω coaxial cable corresponds to a transit time of about 1.5 ns.

Some care is required in order to properly set the discriminator bias levels and photomultiplier high voltage. First the system is checked out with a pulser, to adjust the setting and functioning of the scaler drivers and scalers. Next the actual signals are fed into the circuits and the discriminator outputs "looked at" on an oscilloscope to ascertain that the pulses are "clean" and uniform. The high voltage is set by taking a plateau curve, which will not be completely flat but nevertheless should show a clear knee. If the system is

[29] For example, Canberra Industries (http://www.canberra.com/) and Ortec (http://www.orteconline.com/) both give details of similar setups, including cross references to their own product line.

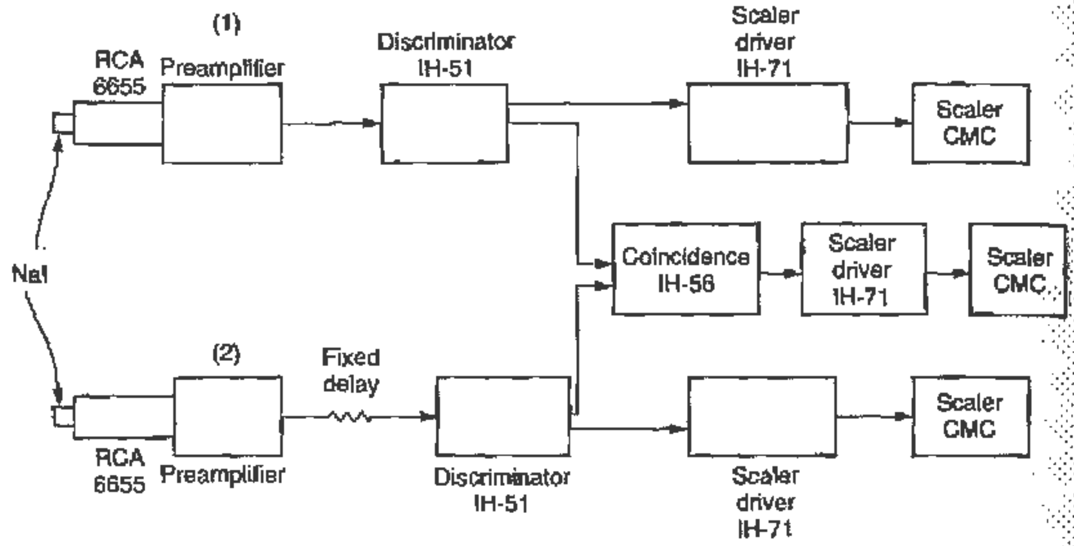


FIGURE 9.31 Block diagram of the electronics used for angular correlation measurements.

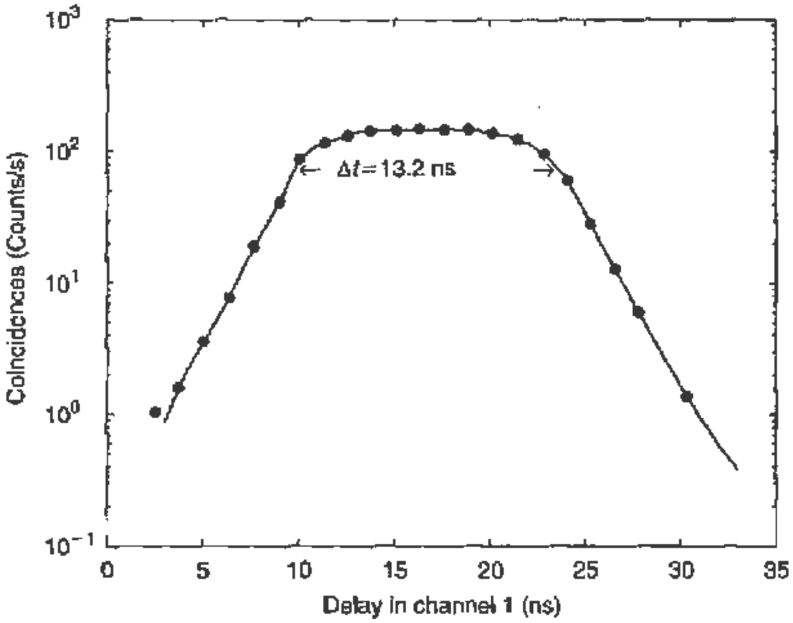


FIGURE 9.32 A delay curve for coincidences from a ^{22}Na source. Note that the resolving time is on the order of 13 ns and that the accidental rate is lower than the true coincidence rate by a factor of at least 1000. The curve through the points is a simple spline interpolation and is only meant to guide the eye.

working properly, the "singles" rates R_1 and R_2 in the two channels should be (almost) equal.

It is possible to measure the resolving time of the coincidence circuit either by taking a "delay curve" (see Fig. 9.32) or by making use

TABLE 9.4 Determination of Resolving Time from Accidental Coincidences

Counts/s			Δt (s)
Channel (1)	Channel (2)	Coincidence	($\Delta t = C/R_1R_2$)
2151	2056	0.061	13.8×10^{-9}
5920	6262	0.528	14.2×10^{-9}
14,662	13,481	2.912	14.7×10^{-9}
31,207	35,443	14.217	12.8×10^{-9}

of Eq. (9.36). When the latter method is used, the two counters are separated by a very large distance and a separate source is placed in front of each. In view of the geometrical arrangement and the fact that an additional delay of 60 ft is placed in one of the channels, all the coincidence counts are accidentals. By varying the distance between the source and the respective counter, the results given in Table 9.4 were obtained; the counting time was on the order of 10 min at each point. We note that the resolving time so obtained (column 4) is quite consistent despite the fact that the accidental coincidence rate increased by a factor of about 2000 between measurements; this resolving time is also consistent with the width of the two input signals (which were about 6 ns wide) and the data of Fig. 9.32.

The above results as well as those to be presented in the following two sections were obtained by students.

9.5.3. The γ–γ Correlation of ^{22}Na

A 100-μCi ^{22}Na source, wrapped with a 0.001-in. brass foil is placed at the center of the apparatus. The dimensions of the source are kept at a minimum, and it is positioned as accurately as possible. Since the solid angle is

$$\Delta\Omega = [\pi \times (0.5)^2]/(8)^2 \approx 4\pi \times 10^{-3}$$

where the dimensions are in inches (see previous section). Assuming a detector efficiency $\epsilon_1 \sim \epsilon_2 \sim 0.3$, the expected rate for "singles" is

$$R_1 \sim R_2 = \frac{3.7 \times 10^{10} \times 10^{-4}}{4\pi} \times (4\pi \times 10^{-3}) \times 0.3 \approx 1000 \text{ counts/s}.$$

Since the two gamma rays are completely correlated when the two counters are at $\theta = 180°$, the expected coincidence rate at this angle is

$$C(\theta) = n\Delta\Omega\epsilon^2 = R\epsilon = 300 \text{ counts/s}. \tag{9.39}$$

The observed rates are on this order of magnitude. However, the 1.277-MeV gamma ray also contributes to the single rate; on the other hand, the finite size of the source and errors in geometrical alignment reduce the coincidence rate from the calculated value.

We first wish to check whether the coincidence circuit is correctly "timed"—that is, whether the appropriate delay has been inserted so as to make truly coincident signals arrive at the circuit at the same time. To this effect the movable counter is rotated to 180° and the counting rate is obtained as a function of the variable delay introduced into channel (1); for convenience, a fixed delay of 12 ft of cable has been introduced into channel (2). The data so obtained have already been given in Fig. 9.32 on a semilog plot, which is the more appropriate representation for a delay curve.

We note that (a) indeed, the peak counting rate occurs when a 16-ns delay is inserted in channel (1) as expected; (b) in the peak region, the delay curve is flat over at least 6 ns; this indicates good efficiency and consequently that small time jitters will not result in changes in the counting rate (provided the delay is set at the center of the curve); (c) the width of the curve at half-maximum, which gives the resolving time of the circuit, is 13.2 ns, in excellent agreement with the values found in Section 9.5.2 and what is expected from the width of the input signals; (d) the accidental rate is very low; by inserting 40 ft of delay it is found to be 0.048 ± 0.005 counts/s, yielding a ratio

$$\frac{\text{signal}}{\text{noise}} = \frac{150}{0.05} \sim 3 \times 10^3, \tag{9.40}$$

which is more than adequate.

The considerable *slope* of the ascending and descending parts of the delay curve is due to the time jitter of the input signals associated with their low peak amplitude. The stability of the system can be judged from the fluctuations of the coincidence rate in the flat region as well as from the fluctuation of the singles rates given in Table 9.5.

We are now ready to obtain data on the angular correlation of ^{22}Na. The movable counter is rotated in appropriate steps to either side of 180°, and the doubles and singles rates are recorded. The resulting data are shown in Fig. 9.33, and in Table 9.5 some representative points are listed.

TABLE 9.5 Representative Data on the γ–γ Correlation of ^{22}Na

θ	Counts/s			Coincidences
(°)	Stationary counter	Movable counter	Coincidence	(Counts/s-degree)
90	3011	3086	1.5 ± 0.1	0.21
150	2996	3071	1.5 ± 0.2	0.23
160	3013	3090	1.7 ± 0.2	0.23
170	2994	3064	3.5 ± 0.2	0.49
175	3011	3114	66.8 ± 1.0	9.2
178	2992	3189	148.0 ± 1.5	20.6
180	2995	3035	159.0 ± 1.6	22.1
182	3014	3178	124.0 ± 1.2	17.2
185	2991	3069	50.2 ± 1.0	7.0
190	3039	3127	3.2 ± 0.2	0.42
200	3005	3102	2.0 ± 0.1	0.26
210	3007	3136	1.8 ± 0.1	0.25

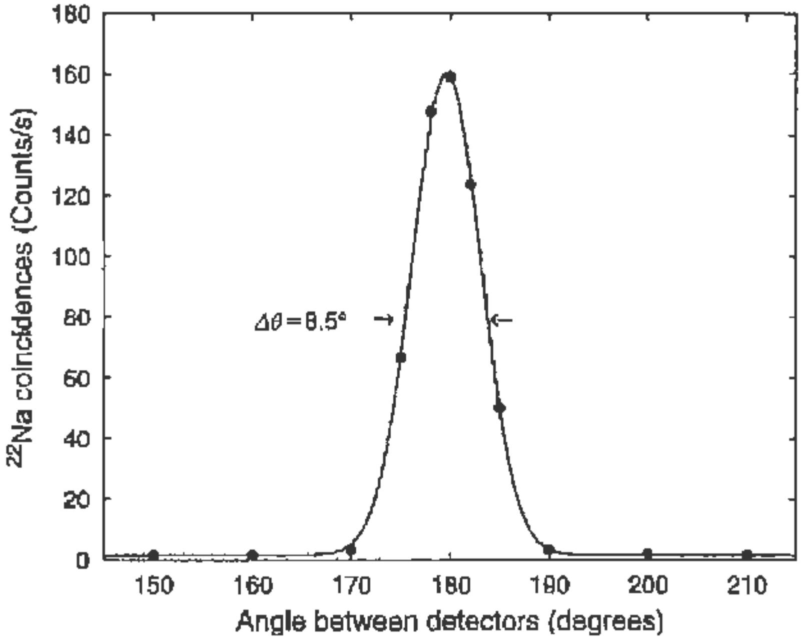


FIGURE 9.33 Angular correlation of the gamma rays from a ^{22}Na source. The coincidence rate is plotted as a function of the angle between the two counters. Note that the full width of the correlation curve is 8.5°, which is entirely due to the angular resolution of the two counters; the isotropic background outside the peak is very small. The curve is a Gaussian fit to the peak region, with a fixed constant background, but only serves to guide the eye.

Columns 2 and 3 of Table 9.5 give the singles rates for the stationary and the movable counter, respectively; the coincidence rate[30] is given in column 4. The counting time at each point was on the order of 1 min, which provides good statistics (about 1% in the peak region).

Indeed we do notice a very pronounced correlation at $\theta = 180°$, with an angular width of ±4.25°. This width is on the order of the angular resolution of our system, which might be taken as the angle subtended at the position of the source by one of the counters

$$\tan\left(\frac{\Delta\theta}{2}\right) = \frac{0.5}{8} \rightarrow \Delta\theta = 7.2°. \tag{9.41}$$

We therefore conclude that this correlation is compatible with

$$C(\theta) = \delta(\pi - \theta). \tag{9.42}$$

The anisotropy as defined by Eq. (9.34) is

$$\alpha = \frac{C(180°) - C(90°)}{C(90°)} = \frac{150 - 1.5}{1.5} \approx 100, \tag{9.43}$$

which is extremely large and compatible with $\alpha \rightarrow \infty$ as predicted by Eq. (9.42).

The counts observed at large angles are still real coincidences, but due mainly to the *isotropic* correlation of the 1.277-MeV gamma ray with one of the annihilation gamma rays; it should be on the order of the correlated counts multiplied by the solid angle for one detector $\Delta\Omega \sim 10^{-2}$, as is indeed the fact. Also, a small fraction of the background originates from annihilation gamma rays that have scattered through a large angle in the source or the converter.

In column 5 of Table 9.5, the coincidence rate has been divided by the angular acceptance $\Delta\theta$ of the movable counter as given by Eq. (9.41). Indeed, since the correlation is a function of θ, it is obvious that our system measures $C(\theta)$ at θ within the differential range $\pm\Delta\theta$.

From the results presented we conclude that ^{22}Na provides a very good technique for aligning and adjusting the equipment, especially since the strong correlation from the annihilation gamma rays is quite easy to detect.

[30] The rate for accidentals should have been subtracted from the results of column 4; however, it is so small (see Eq. (9.40)) that we neglect it.

Also, the obtained correlation provides strong evidence for the annihilation of the positron–electron pair into two gamma rays; if a differential discriminator is used after the detector, it is also possible to measure the energy of the coincident gamma rays. The angular resolution of the equipment may be easily improved by simply increasing the distance between the source and the counters. In fact, precise data on positron annihilation are quite sensitive to the momentum of the positronium just before it annihilates; this in turn provides information on the structure of the Fermi surface of the converter material.

9.5.4. The γ–γ Correlation of ^{60}Co

Once the equipment has been adjusted and aligned (for example, with ^{22}Na) as described before, any correlation may be measured. A ^{60}Co source of the same strength as the ^{22}Na source (100 μCi) was placed at the center of the apparatus, and data were taken every 15°. The discriminator levels could be readjusted, but it is usually preferable to leave everything as is.

Since the ^{60}Co γ–γ correlation has a small anisotropy (as compared to ^{22}Na, Eq. (9.43)) the expected coincidence rate is

$$C(\theta) = N(\Delta\Omega)^2\epsilon^2 \approx 4 \text{ counts/s}, \tag{9.44}$$

which is much smaller than that given by Eq. (9.39) for the same source strength. Consequently, also, the signal-to-noise ratio (Eq. (9.40)) will be only about 30, and the "accidental" rate, which was 0.070 counts/s, must be subtracted. Furthermore in view of the smaller correlation, better statistical accuracy is required.

Representative data taken in one run are presented in Table 9.6 and plotted in Fig. 9.34. In column 5 the coincidence rate after the subtraction of accidentals is given, while in column 6 the rate at each angle is normalized to the rate at 90°. At each point sufficient coincidence counts were taken to give 1% statistical accuracy (10,000 s $\approx$ 3 h); these errors are indicated by the error bars shown in Fig. 9.34, where we plot $\alpha(\theta) = C(\theta)/C(90^\circ)$ against angle. We see that the *fractional* errors on $(\alpha(\theta) - 1)$ are now much larger, and on the order of 10%.

It is known from theoretical considerations that the ^{60}Co correlation function is of the form

$$\alpha(\theta) = \frac{C(\theta)}{C(90^\circ)} = 1 + a_1 \cos^2\theta + a_2 \cos^4\theta. \tag{9.45}$$

TABLE 9.6 Representative Data on the γ–γ Correlation of ^{60}Co

	Counts/s				
θ (°)	Stationary counter	Movable counter	Coincidences	Corrected coincidences	$\frac{C(\theta)}{C(90°)}$
60	2203	2129	0.880	0.810	1.080
90	2132	2157	0.820	0.750	1.000
105	2152	2127	0.857	0.787	1.049
120	2144	2130	0.864	0.794	1.059
135	2109	2125	0.886	0.816	1.088
150	2132	2136	0.933	0.863	1.151
165	2121	2123	0.931	0.861	1.148
180	2116	2124	0.944	0.874	1.165
210	2086	2134	0.889	0.819	1.087

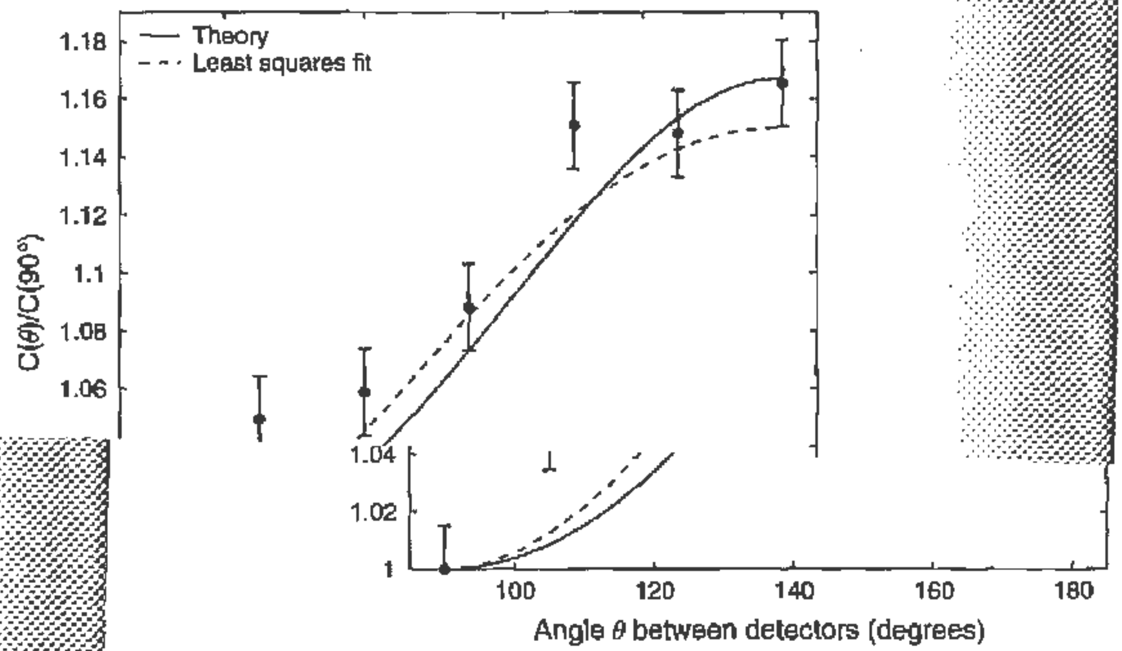


FIGURE 9.34 Data on the angular correlation of the two gamma rays fror correlation function $C(\theta)/C(90°)$ is plotted against the angle between the t Note, however, that the ordinates begin at the value 1.00. The experiment shown, and the dashed curve is a least-squares fit to the data. The solid lir theoretical curve, which is given by the function $1 + 0.125\cos^2\theta + 0.042$ co

ı ^{60}Co. The
/o counters.
l points are
: shows the
$^4\theta$.

A least-squares fit to Eq. (9.45) was made, using the entire set of experimental data,[31] and the following values were obtained for the coefficients a_1 and a_2.

$$a_1 = 0.190 \pm 0.08 \qquad a_2 = -0.04 \pm 0.08$$

The theoretical values resulting from the spin assignments $I_a = 4^+$, $I_b = 2^+$ and $I_c = 0^+$ (see Fig. 9.28) are

$$a_1 = 0.125 \qquad a_2 = 0.042.$$

The correlation function that results from the above coefficients is included in Fig. 9.34; the dashed line represents the least-squares fit, and the solid line the theoretical curve.

From Fig. 9.34 we see clearly that an anisotropy in the angular distribution of the γ–γ coincidences from ^{60}Co exists; we obtain

$$\alpha = \alpha(180^\circ) - 1 = 0.165 \pm 0.016. \tag{9.46}$$

The error flags in Fig. 9.34 were set at 1.5%, but the data points scatter even more. This is not due to the "statistics," but to random fluctuations and drifts of the equipment over the long counting intervals.

[31] This included 21 more measurements in addition to those presented in Table 9.6.

CHAPTER 10

Elements from the Theory of Statistics

10.1. DEFINITIONS

Statistics is the science that tries to draw inferences from a finite number of observations constituting only a sample, so as to postulate rules that apply to the entire population from which the sample was drawn.

In the field of physics, statistics is needed (a) to fit data—that is, to estimate the parameters of assumed frequency functions; (b) to treat random errors; and (c) to interpret phenomena that are inherently of a statistical nature.

10.1.1. Definition of Probability

The probability of occurrence of an event can be axiomatically defined as equal to one (= 1) if the event occurred, or equal to zero (= 0) if the event did not take place. An alternative definition of probability is based on the *frequency* of occurrence of an event. Suppose that several trials of the same experiment have been made; then the probability of occurrence of an event A, that is $P(A)$, is given by the number of times event A was obtained divided by the total number of trials (in the limit that the total number of trials approaches infinity). This definition of probability retains its full value even in the case of nonrepetitive experiments, since the one trial can be considered as the first of a series of trials.

10.1.2. Sample Space

Any *set of points* that represents all possible outcomes of an experiment is a sample space. For example, if a coin is tossed twice, the sample space consists of the 4 points indicated in Fig. 10.1. (Sample spaces can be finite or infinite and discrete or continuous.)

Once the sample space for a particular experiment is constructed, we may assign (in the sense of Definition 10.1.1) a probability p_i to each point i of the space. From the definition of probability, we have

$$p_i > 0 \qquad \sum_{\text{all sample space points}} p_i = 1;$$

thus

$$p_i \leq 1$$

and the probability of occurrence of an event A is

$$P(A) = \frac{\sum_A p_i(A)}{\sum p_i} = \sum_A p_i(A),$$

where $\sum_A$ indicates summation over all points that include event A.

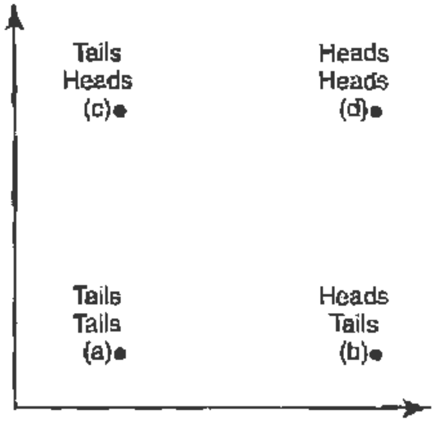


FIGURE 10.1 Simple example of a discrete and finite sample space. Here the sample-space points correspond to all possible outcomes of "tossing a coin" twice.

In most situations treated by statistics, equal probability is assigned to each sample-space point, a condition we will maintain throughout this discussion. Then

$$p_i = \frac{1}{n},$$

n being the total number of sample-space points, and

$$P(A) = \frac{n(A)}{n},$$

where $n(A)$ is the number of sample-space points containing event A.

For example, in the case of the sample space of Fig. 10.1, the probability of obtaining heads at least once is

$$P(\text{heads}) = \frac{n(\text{heads at least once})}{n} = \frac{3}{4}$$

while the probability of obtaining heads once and tails once (irrespective of order) can again be found by counting the appropriate points in the sample space of Fig. 10.1. We obtain

$$P(\text{heads, tails}) = \frac{n(\text{heads, tails})}{n} = \frac{2}{4}.$$

10.1.3. Probability for the Occurrence of a Complex Event

The probability that both events A and B will occur is called the joint probability

$$P[AB] = \frac{n(A \text{ and } B)}{n},$$

where n = total number of sample-space points. The probability that either A or B will occur is called the *either probability*

$$P[A + B] = \frac{n(A \text{ or } B)}{n},$$

and the probability that A will occur when it is certain that B occurred is called the *conditional probability*

$$P[A|B] = \frac{n(A \text{ and } B)}{n(B)}.$$

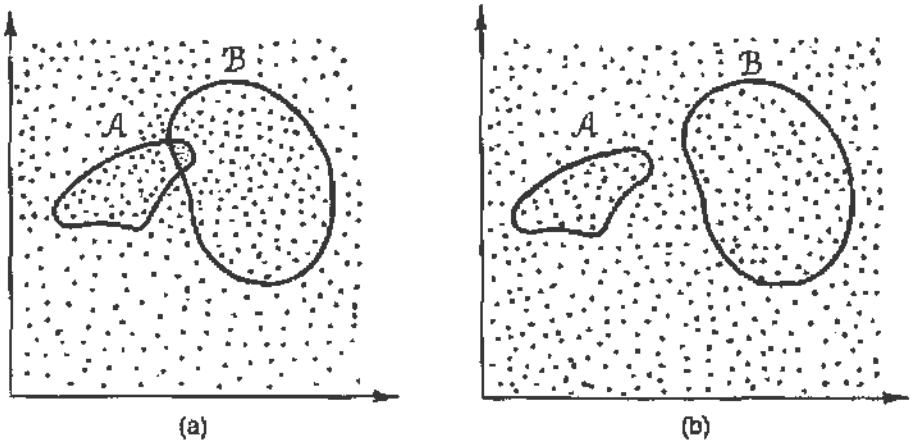


FIGURE 10.2 In the sample spaces shown it is assumed that all sample-space points in domain $\mathcal{A}$ contain event A, whereas all points in domain $\mathcal{B}$ contain event B. (a) There exists a region where both event A and event B can occur simultaneously. (b) No such region exists; events A and B are mutually exclusive.

All these probabilities are defined in the sense of Definition 10.1.2 as the number of sample-space points that contain the stated condition divided by the total number of sample-space points *allowed for* by the statement.

Figures 10.2a and 10.2b illustrate two sample spaces. All points within domain $\mathcal{A}$ include event A while all points within domain $\mathcal{B}$ include events B. The points contained in any intersection of the two domains $\mathcal{A}$ and $\mathcal{B}$ include both events A and B.

If such a common intersection does not exist in sample space, the two events are *mutually exclusive*, and

$$P[AB] = 0.$$

It follows from consideration of Fig. 10.2 that

$$P[A + B] = P[A] + P[B] - P[AB].$$

For the conditional probability

$$P[A|B] = \frac{n(\mathcal{A} \text{ intersection } \mathcal{B})}{n(\mathcal{B})},$$

since the condition that event B occurred restricts our sample within domain $\mathcal{B}$. However,

$$P[B] = \frac{n(\mathcal{B})}{n}$$

and

$$P[AB] = \frac{n(\mathcal{A}\text{ intersection }\mathcal{B})}{n} = P[A|B]\cdot P[B] = P[B|A]\cdot P[A]. \tag{10.1}$$

If $P[A|B] = P[A]$, it means that the occurrence of B does not affect the probability of occurrence of A. We say that the two events A and B are *independent*. It then follows from Eq. (10.1) that

$$P[AB] = P[A]\cdot P[B]. \tag{10.2}$$

Equation (10.2) in turn implies (when combined with Eq. (10.1)) that for independent events

$$P[B|A] = P[B].$$

To illustrate some of the ideas we have just expressed, consider the following. For the sample space of Fig. 10.1 we may define: event A = heads in first throw, and event B = heads in second throw. The domains are shown in Fig. 10.3, and it follows (assigning $p = 1/4$ to each point) that

$$P[A] = \frac{1}{2}; \qquad P[B] = \frac{1}{2}$$

$$P[AB] = \frac{1}{4}$$

$$P[A+B] = P[A] + P[B] - P[AB] = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$$

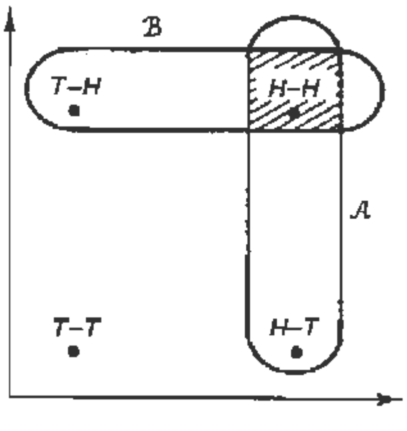


FIGURE 10.3 The sample space of Fig. 10.1 including the domain $\mathcal{A}$ (heads in the first throw) and the domain $\mathcal{B}$ (heads in the second throw).

$$P[A|B] = \frac{1}{2}; \qquad P[B|A] = \frac{1}{2}$$

$$P[AB] = P[A|B] \cdot P[B] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = P[A] \cdot P[B].$$

Thus events A and B are *not* mutually exclusive but are independent.

10.1.4. Random Variable

To study a sample space analytically (instead of geometrically), it is convenient to use a numerical variable that takes a definite value for each and *every* point of the sample space; however, the same value may be assigned to several points. Thus, a random variable used for the representation of a finite and discrete sample space will have a definite range and will take only discrete values. As an example, for the sample space of Fig. 10.1, we can assign to the random variable x the value 0 for points (b) and (c) (one each of heads and tails), the value -1 for point (a) (both tails), and the value $+1$ for point (d) (both heads).

10.1.5. Frequency Function

A frequency function (of a random variable) is a function $f(x)$ such that $f(x_0)$ is the probability that the random variable x may take the specific value x_0. By Definition 10.1.1, $f(x)$ gives the number of points in the sample space that have been assigned the value of x of the random variable, divided by the total number of sample-space points. The function $f(x)$ is defined only within the range of x and need not have a definite analytic form. For the example considered above (the sample space of Fig. 10.1), $f(x)$ is just a table, as shown in Table 10.1 (see also Fig. 10.4).

TABLE 10.1 Example of a Frequency Function $f(x)$ of the Random Variable x

Sample-space point	x	$f(x)$
(a)	-1	$\frac{1}{4}$
(b,c)	0, 0	$\frac{1}{2}$
(d)	$+1$	$\frac{1}{4}$

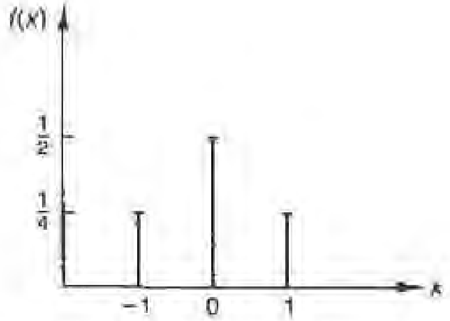


FIGURE 10.4 The distribution function of the discrete random variable x defined in Table 10.1.

The summation of $f(x)$ over the entire range of x must give 1:

$$\sum_{\text{all } x} f(x) = 1.$$

The probability that the random variable may take any value smaller or equal to x is given by

$$F(x) = \sum_{t < x} f(t)$$

and is called the *distribution function* of x (or integral distribution function).

It is sometimes convenient to describe a sample space in terms of two or more random variables, a frequency function existing for each of them. If these random variables are independently distributed in the sense of Eq. (10.2), the joint frequency function is

$$f(x_1, x_2, \ldots) = f(x_1)\, f(x_2) \cdots f(x_n).$$

If the random variable is continuously varying (for example, it describes the height of individuals), the probability of occurrence of the specific value x when a measurement is performed defines the frequency function $f(x)\,dx$ of the random variable x. The random variable may now take any value within the range of its definition. Note, however, that the probability of occurrence of the exact value x is zero, while it is the probability of occurrence of some value in the infinitesimal interval dx about x that exists. For a continuously varying random variable, we have

$$f(x) \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{+\infty} f(x)\,dx = 1.$$

Similarly

$$\int_a^b f(x)\,dx = P[a < x < b] \qquad a < b$$

and

$$F(x) = \int_{-\infty}^{x} f(t)\,dt.$$

10.1.6. Some Definitions from Combinatorial Analysis

(a) *Permutations*. A permutation of n objects in groups of r objects is defined as follows. Consider n objects; any group of r of these objects, when ordered, forms a permutation; the same group of r objects, when ordered in a different fashion, forms a *new* permutation. As an example consider the three objects:

$$\square, \triangle, \bigcirc$$

There are only six possible permutations of three objects in groups of two:

$$\square\triangle, \quad \triangle\square; \quad \square\bigcirc, \bigcirc\square; \quad \triangle\bigcirc, \quad \bigcirc\triangle$$

We state without proof that the number of possible permutations of n objects in groups of r, ${}_nP_r$, is

$${}_nP_r = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}.$$

Then

$${}_nP_n = n!$$

as it must be.

(b) *Combinations.* A combination of n objects in groups of r objects is defined as any grouping of r objects out of the original n. The ordering within the grouping is not relevant. Thus for the previous example there are only three possible combinations

$$\square\triangle; \quad \bigcirc\square; \quad \bigcirc\triangle.$$

The number of possible combinations of n objects in groups of r, $\begin{bmatrix} n \\ r \end{bmatrix}$, is

$$\begin{bmatrix} n \\ r \end{bmatrix} = \frac{{}_nP_r}{{}_rP_r} = \frac{n!}{r!(n-r)!}.$$

(c) *Note.* Note that

$$n! = n \cdot (n-1)! \qquad 1! = 0! = 1.$$

10.2. FREQUENCY FUNCTIONS OF ONE VARIABLE

10.2.1. Definitions

Let us assume that a population (for example, all the possible outcomes of an experiment) can be described by a frequency function; we may attempt to find this function in two ways:

(a) By the use of a mathematical model based on the definitions of the previous section, thus obtaining a "theoretical frequency function."
(b) By observing a sample of the population and determining its "empirical frequency function."

The advantage of obtaining a frequency function for a population is that the few parameters involved in the frequency function suffice to describe completely the *entire* population and thus provide as much information as the most extensive data.

We will now deal only with populations that can be described by a frequency function depending on a single variable. To obtain the empirical frequency function it is best to divide the members of the sample into classes (defined by the random variable) and then make a graphical plot or *histogram* of the sample. If we try to describe the histogram, the first obvious features are its location and its spread.

A very useful set of measures are the *moments* of a histogram, defined in the usual way (moments of forces, electric moments, etc.). Thus, if x_i is the value of the random variable for the class i and if f_i is the number of events in this class, the k^{th} moment of the *empirical* frequency function about the origin is

$$m'_k = \frac{1}{n} \sum_{\text{all } i} x_i^k f_i,$$

where n is the size of the sample. Similarly, the kth moment about any other point x_0 is

$$m_k(x_0) = \frac{1}{n} \sum_{\text{all } i} (x_i - x_0)^k f_i.$$

10.2.2. Mean and Standard Deviation

The first moment about the origin, m'_1, is called the mean and will be denoted by m:

$$m = m'_1 = \frac{1}{n} \sum_{\text{all } i} x_i \, f_i \tag{10.3}$$

(commonly called the "average" of x). The second moment about the mean, m_2, is called the *variance*; its square root is called the *standard deviation* and is denoted by s: s has the same dimensions as the random variable x:

$$s = \sqrt{m_2} = \sqrt{\frac{1}{n} \sum_{\text{all } i} (x_i - m)^2 f_i}\,. \tag{10.4}$$

An often used relation pertaining to s is

$$s^2 = \frac{1}{n} \sum_{\text{all } i} (x_i - m)^2 f_i = \frac{1}{n} \sum (x_i^2 - 2mx_i + m^2) f_i$$

$$s^2 = \frac{1}{n} \sum_{\text{all } i} x_i^2 f_i - \frac{2m}{n} \sum (x_i f_i) + m^2$$

$$s^2 = \frac{1}{n} \sum_{\text{all } i} (x_i^2 f_i) - m^2$$

usually written as

$$\overline{\Delta x^2} = \overline{x^2} - \left(\bar{x}\right)^2 . \tag{10.5}$$

In most cases the mean and the standard deviation are the best measures (contain most information) of an empirical frequency function; there are, nevertheless, cases where they are very poor measures, and instead it is much better to give other location measures, such as the median or the geometric mean, and so on; and other variation measures such as the range or the mean variation, $(1/n) \sum |x_i - m| \, f_i$, and so on.

10.2.3. Theoretical Frequency Functions

As mentioned before, a theoretical frequency function $f(x)$ might be of the *discrete* type—that is, the random variable x takes only integer values, or of the "continuous" type. Most of the discrete random variables usually represent the number of successes, or of counts obtained, etc. In going from discrete frequency functions to continuous ones, obviously all summations are replaced by integrals.

Moments are defined as in Eq. (10.3), but instead of the empirical frequencies f_i, the theoretical frequency function $f(x)$ is used; the theoretical moments are designated by Greek letters, Latin letters being reserved for the empirical moments.

Thus, the k^{th} moment about the origin is

$$\mu'_k = \sum_{x=-\infty}^{x=+\infty} x^k f(x).$$

The first moment about the origin gives the mean, and is denoted by $\mu = \mu'_1$. The k^{th} moment of a theoretical frequency function about its mean is

$$\mu_k = \sum_{x=-\infty}^{x=+\infty} (x-\mu)^k f(x).$$

The square root of the second moment about the mean gives the standard deviation and is denoted by $\sigma = \sqrt{\mu_2}$:

$$\mu_2 = \sum_{x=-\infty}^{x=+\infty} (x-\mu)^2 f(x).$$

10.2.4. The Bernoulli or Binomial Frequency Function

This basic frequency function is applicable when there are only *two* possible outcomes of an experiment, as, for example, the occurrence of an event A or its nonoccurrence (we designate this by B). If the experiment is repeated n times, the random variable x describes the number of times event A occurred. The frequency function—that is, the probability of obtaining a certain x—is given by

$$f(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}, \tag{10.6}$$

where p is the probability that event A will occur in this experiment (defined in the sense of Section 10.1.1); and $q = 1 - p$ is the probability that B will happen, namely, that event A will not occur.

To prove Eq. (10.6), consider the probability of obtaining event A, x times in a definite sequence

$$\underbrace{AA\cdots A}_{x} \qquad \underbrace{BB\cdots B}_{n-x};$$

this joint probability of order n is according to Definition 10.1.3

$$\underbrace{pp\cdots p}_{x}\underbrace{qq\cdots q}_{n-x} = p^x q^{n-x}$$

since the outcome of consecutive experiments is independent. However, any other sequence, containing the same number x of occurrences, is also a satisfactory answer, since we are not interested in the order of occurrence of event A. Thus we must sum over all sample-space points that give x occurrences; the number of all such sample-space points is given by the permutations of n objects in groups of n when x of them are alike (have probability p), which is

$$\frac{n!}{x!(n-x)!},$$

completing the proof of Eq. (10.6).

The frequency function fulfills the normalization requirement as it should, since

$$\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} p^x q^{n-x} = (p+q)^n = [p+(1-p)]^n = 1. \tag{10.7}$$

10.2.5. Moments of the Binomial Frequency Function

From the definitions of Section 10.2.3, and since the range of x is from 0 to n, we have

$$\begin{aligned}\mu = \mu_1' &= \sum_{x=0}^{n} x f(x) = \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}\\ &= \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}\\ &= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}.\end{aligned}$$

If we let $y = x - 1$, it follows that

$$\mu = np \sum_{y=0}^{n-1} \frac{(n-1)!}{y![(n-1)-y]!} p^y q^{[(n-1)-y]},$$

where now the sum is equal to $(p+q)^{n-1} \equiv 1$. Thus

$$\mu = np. \tag{10.8}$$

Next we wish to obtain the second moment about the mean, $\mu_2 = \sigma^2$. We first calculate μ_2', given by

$$\mu_2' = \sum_{x=0}^{n} x^2 \frac{n!}{x!(n-x)!} p^x q^{n-x}.$$

We use

$$x^2 = x(x-1) + x$$

so that

$$\begin{aligned}\mu_2' &= \sum_{x=0}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x q^{n-x} + \mu \\ &= \sum_{x=2}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x q^{n-x} + \mu \\ &= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2} q^{n-x} + \mu\end{aligned}$$

and letting $y = x - 2$, as before, the sum is equal to $(p+q)^{n-2} \equiv 1$ and we obtain

$$\mu_2' = n(n-1)p^2 + \mu = n^2p^2 - np^2 + np.$$

Next we use Eq. (10.5) to obtain

$$\mu_2 = \sigma^2 = \mu_2' - \mu^2 = -np^2 + np = np(1-p) = npq.$$

Thus

$$\sigma = \sqrt{npq}. \tag{10.9}$$

The binomial frequency function is applicable to many physical situations, but it is cumbersome to calculate with. When n becomes large, however, the binomial frequency function approaches either the Poisson or the Gaussian frequency function, which will be discussed in Sections 10.2.6 and 10.2.7. In order for the binomial frequency function[1] to approach the

Poisson distribution n must be large, for example, $n > 100$, but $\mu = np$ must be finite and small, for example, $p < 0.05$.

Gaussian distribution n must be large, for example, $n > 30$, and also p must be large, for example, $p > 0.05$.

10.2.6. The Poisson Frequency Function

This is still a frequency function for the discrete random variable x, which describes, as in Section 10.2.4, the number of times event A will be obtained if the experiment is repeated n times when $n \to \infty$ for (large n). Contrary to Eq. (10.6), however, neither n nor p appears explicitly in the analytic expression of the frequency function, but instead only their product

$$y = np, \tag{10.10}$$

which remains finite despite $n \to \infty$, since $p \to 0$. The Poisson frequency function is given by

$$f(x) = \frac{y^x e^{-y}}{x!}, \tag{10.11}$$

and it is shown in the next section that y is the mean of the distribution governed by Eq. (10.11).

To prove Eq. (10.11), let us first note that since n is large, it (but not x) may be treated as a continuous variable; second, we will assume that for a small (differential) number of trials dn, the probability of obtaining event A once is proportional to this number of trials: that is,

$$P\{1, dn\} = \lambda\, dn, \tag{10.12}$$

where λ is a constant. Note that Eq. (10.6) fulfills this requirement for $x = 1$ in the limit that $p \to 0$ or $q \to 1$. In terms of sample space our assumption means that the density of sample-space points containing event A is uniform in the limit of a differential element of sample-space area.

[1] See, however, the detailed discussion in Section 10.2.9.

The Poisson frequency function then follows for all populations for which assumption (10.12) is valid.

Let $P\{x, n\}$ be the probability of obtaining event A, x times in n trials, so that $P\{0, n\}$ is the probability of obtaining no events A in n trials. Then the probability of obtaining no events in $n + dn$ trials is

$$P\{0, n + dn\} = P\{0, n\} \cdot [1 - P\{1, dn\}]$$

since the events are independent.[2] Using Eq. (10.12) we obtain

$$\frac{P\{0, n + dn\} - P\{0, n\}}{dn} = -P\{0, n\} \cdot \lambda$$

or

$$-\frac{dP\{0, n\}}{dn} = P\{0, n\} \cdot \lambda,$$

which has the solution

$$\ln P\{0, n\} = -n\lambda$$

$$P\{0, n\} = e^{-n\lambda} \qquad (10.13)$$

and use has been made of the initial condition that for $n = 0$

$$P\{0, 0\} = 1.$$

In a similar manner we obtain

$$P\{1, n + dn\} = P\{1, n\}P\{0, dn\} + P\{0, n\}P\{1, dn\},$$

where the two possible *either* probabilities are summed. Making use again of Eq. (10.12), we may write the above result as

$$P\{1, n + dn\} = P\{1, n\} \cdot [1 - \lambda dn] + P\{0, n\} \cdot \lambda dn$$

by further transforming and using Eq. (10.13) as well,

$$\frac{dP\{1, n\}}{dn} + \lambda P\{1, n\} - \lambda e^{-n\lambda} = 0.$$

The solution of this linear first-order equation is straightforward, leading to

$$P\{1, n\} = e^{-n\lambda}\left[\int e^{n\lambda}\lambda e^{-n\lambda} dn + C\right] = (n\lambda)e^{-n\lambda}, \qquad (10.14)$$

making use of the initial condition $P\{1, 0\} = 0$.

[2] Since the increase in the number of trials dn is differential, the possibility of obtaining more than one event in dn is excluded.

In general the following recursion formula holds

$$\frac{dP\{x,n\}}{dn} + \lambda P\{x,n\} - \lambda P\{(x-1),n\} = 0,$$

which is satisfied by

$$f(x) = P\{x,n\} = \frac{(\lambda n)^x e^{-n\lambda}}{x!}, \tag{10.15}$$

as can be verified by substitution.

Thus Eq. (10.11) has been proven, and we can identify the proportionality constant λ as the probability that event A will occur in one trial.[3] As pointed out before, however, it is only the product $y = \lambda n = pn$ that may be properly defined: it is the theoretical mean of the discrete random variable x when the same (large) number of n trials is repeated many times.

Equation (10.11) correctly fulfills the normalization requirement

$$\sum_{x=0}^{n=\infty} f(x) = e^{-y} \sum_{x=0}^{\infty} \frac{y^x}{x!} = e^{-y}e^y = 1.$$

It is shown in Section 10.2.9 that Eq. (10.11) is the limiting form of Eq. (10.6) when $p \to 0$ and $n \to \infty$.

10.2.7. Moments of the Poisson Frequency Function

Following the approach used in Section 10.2.5, the moments of the Poisson frequency function will be obtained by direct evaluation of the defining equations; note that as $n \to \infty$ the upper limit of x is also ∞:

$$\mu = \mu_1' = \sum_{x=0}^{x=n\to\infty} x \frac{y^x e^{-y}}{x!} = \sum_{x=1}^{\infty} \frac{y^x e^{-y}}{(x-1)!}$$
$$= e^{-y} y \sum_{x=1}^{\infty} \frac{y^{(x-1)}}{(x-1)!} = e^{-y} y e^y = y.$$

Thus

$$\mu = y \tag{10.16}$$

[3] $P\{1, 1\} = \lambda e^{-\lambda} \to \lambda$ when $\lambda \ll 1$.

as expected from our previous discussion. We see that through Eq. (10.16) we obtain the physical significance for the parameter y. Further,

$$\mu_2' = \sum_{x=0}^{\infty} x^2 \frac{y^x e^{-y}}{x!} = \sum_{x=0}^{\infty} \left(x(x-1)\frac{y^x e^{-y}}{x!} \right) + y$$

$$= e^{-y} \sum_{x=2}^{\infty} \left(\frac{y^x}{(x-2)!} \right) + y = e^{-y} y^2 \sum_{x=2}^{\infty} \frac{y^{(x-2)}}{(x-2)!} + y \Rightarrow y^2 + y,$$

and using Eq. (10.5) we obtain

$$\mu_2 = \sigma^2 = \mu_2' - \mu^2 = y^2 + y - y^2 = y.$$

Thus

$$\sigma = \sqrt{y}. \tag{10.17}$$

The close analogy of Eq. (10.16) to Eq. (10.8) and of Eq. (10.17) to Eq. (10.9) should be clear; also the derivation of these equations is completely analogous.

10.2.8. The Gaussian or Normal Frequency Function and Its Moments

This is indeed a most important frequency function because (a) it is a limiting case that many frequency functions approach; (b) the distribution of most physical observables is satisfactorily described by it; and (c) measurements containing *random* errors are distributed normally about the true value of the measured quantity.

The Gaussian distribution gives the frequency of the continuous random variable x in terms of two parameters a and b, which are the first and second moments of the frequency function. In its normalized form, the Gaussian distribution is given by

$$f(x)\,dx = \frac{1}{b\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{x-a}{b} \right)^2 \right] dx \tag{10.18}$$

and is shown in Fig. 10.5. The range of the variable x is from $-\infty$ to $+\infty$. In order to show the normalization of Eq. (10.18), as well as to find the moments, it is useful to know the values of the integral of $x^n e^{-ax^2}$,

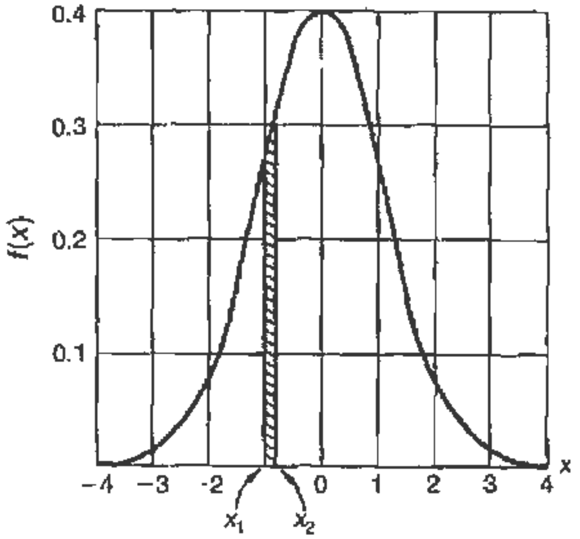


FIGURE 10.5 The Gaussian frequency function normalized to zero mean and unit variance $f(x)\,dx = (1/\sqrt{2\pi})e^{-x^2/2}\,dx$. Note that the probability of finding a value of x between x_1 and x_2 is proportional to the corresponding area under the Gaussian.

TABLE 10.2 Value of the Integral $f(n) = \int_0^\infty x^n \exp(-ax^2)\,dx$

n	$f(n)$	n	$f(n)$
0	$\frac{1}{2}\sqrt{\pi/a}$	1	$1/2a$
2	$\frac{1}{4}\sqrt{\pi/a^3}$	3	$1/2a^2$
4	$\frac{3}{8}\sqrt{\pi/a^5}$	5	$1/a^3$

$$f(n) = \int_{-\infty}^{\infty} x^n \exp(-ax^2)\,dx = \begin{cases} 2f(n) & \text{when } n \text{ is even} \\ 0 & \text{when } n \text{ is odd} \end{cases}$$

which are summarized in Table 10.2. To obtain the moments we proceed as before

$$\mu = \mu_1' = \frac{1}{b\sqrt{2\pi}} \int_{-\infty}^{+\infty} x \exp\left[-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right] dx.$$

We let $x = tb + a$, $dx = b\,dt$; thus

$$\mu = \frac{1}{\sqrt{2\pi}}\left[\int_{-\infty}^{+\infty} bte^{-(t^2/2)}\,dt + \int_{-\infty}^{+\infty} ae^{-(t^2/2)}\,dt\right].$$

According to Table 10.2, integrals with odd powers of t vanish, thus

$$\mu = a. \tag{10.19}$$

Similarly

$$\mu_2' = \frac{1}{b\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 \exp\left[-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right] dx$$

with the same substitution

$$\mu_2' = \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{+\infty} b^2 t^2 e^{-(t^2/2)}\, dt + \int_{-\infty}^{+\infty} 2abte^{-(t^2/2)}\, dt + \int_{-\infty}^{+\infty} a^2 e^{-(t^2/2)}\, dt\right],$$

so that by using Table 10.2 we obtain

$$\mu_2' = \frac{1}{\sqrt{2\pi}} \left[b^2 \frac{1}{2}\sqrt{8\pi} + a^2\sqrt{2\pi}\right] = a^2 + b^2$$

and, using Eq. (10.5),

$$\mu_2 = \sigma^2 = \mu_2' - \mu^2 = b^2.$$

Thus

$$\sigma = b. \tag{10.20}$$

We see that through Eqs. (10.19) and (10.20), we obtain the physical significance of the parameters a and b of Eq. (10.18). Thus, Eq. (10.18) takes the form

$$f(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mu - x}{\sigma}\right)^2\right] dx. \tag{10.21}$$

It is sometimes useful to transform the random variable linearly so as to obtain a frequency function with zero mean and unit standard deviation; the transformation is

$$y = \frac{x-\mu}{\sigma}; \qquad dy = \frac{dx}{\sigma},$$

and Eq. (10.18) becomes (as shown in Fig. 10.5)

$$f(y)dy = \frac{1}{\sqrt{2\pi}} e^{-(y^2/2)}\, dy. \tag{10.22}$$

10.2.9. The Gaussian Frequency Function as a Limiting Case

In the previous section we gave Eq. (10.18) without proof. We will now show that it can be obtained from the binomial frequency function, Eq. (10.6), in the limit of $n \to$ large and $|np - x| \ll np$.

Consider Eq. (10.6):

$$f(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}.$$

If $n \to \infty$ but $np \to \mu$ remains finite, we may write

$$f(x) = \frac{n(n-1)\cdots(n-x+1)}{n^x} \cdot \frac{(np)^x}{x!} \cdot (1-p)^{n-x}$$

$$f(x) = \frac{1[1-(1/n)]\cdots[1-(x-1)/n]}{(1-p)^x} \cdot \frac{(np)^x}{x!} \cdot (1-p)^n. \qquad (10.23)$$

However,

$$(1-p)^n = [(1-p)^{-(1/p)}]^{-np} \to e^{-\mu}$$

since from the definition of e,

$$\lim_{z\to 0}(1+z)^{1/z} = e$$

and in the present case we have $p \to 0$. Further

$$\lim_{n\to\infty} \frac{1[1-(1/n)]\cdots[1-(x-1)/n]}{(1-p)^x} = 1$$

because $p \to 0$ and x is finite; by substituting the last two expressions into Eq. (10.23) we obtain the Poisson frequency function, Eq. (10.11):

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}.$$

We now use the further condition that x be a continuous variable and $|np - x| \ll np$, namely, its deviations from the mean μ be small; then the following approximate expression is valid:

$$\ln\frac{\mu}{x} = \ln\left(1 + \frac{\mu - x}{x}\right) = \left(\frac{\mu - x}{x}\right) - \frac{1}{2}\left(\frac{\mu - x}{x}\right)^2 + \cdots.$$

Hence

$$\frac{\mu}{x} \approx \exp\left(\frac{\mu - x}{x}\right) \exp\left[-\frac{1}{2}\left(\frac{\mu - x}{x}\right)^2\right]$$

and

$$\mu^x \approx x^x \exp(\mu - x) \exp\left[-\frac{1}{2}\frac{(\mu - x)^2}{x}\right].$$

From Stirling's formula we have

$$x! \simeq \sqrt{2\pi x}x^x e^{-x}$$

and by substituting $(\mu)^x$ and $x!$ into Eq. (10.11) we obtain

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} = \frac{e^{-\mu} x^x e^{(\mu - x)} \exp\left\{-\frac{1}{2}[(\mu - x)^2/x]\right\}}{\sqrt{2\pi x}x^x e^{-x}}$$
$$= \frac{1}{\sqrt{2\pi x}} \exp\left[-\frac{1}{2}\left(\frac{\mu - x}{\sqrt{x}}\right)^2\right]. \qquad (10.24)$$

Thus the binomial frequency function in its limit approaches a Gaussian frequency function with

$$\begin{array}{ll} \text{mean} & \mu = np \\ \text{standard deviation} & \sigma = \sqrt{x} \approx \sqrt{npq}, \end{array} \qquad (10.25)$$

where $x \approx npq$ follows from $|\mu - x| \ll \mu$ and $p \to 0$. From Eq. (10.25) we see that the moments of the limiting Gaussian frequency function are the limits of the moments of the original binomial frequency function.

10.2.10. Properties of the Gaussian Frequency Function

Let us now interpret the frequency function given by Eq. (10.18). We could refer to our original example of obtaining event A, x times when a choice between A or B is made n times; x then can vary from 0 to n in integer values. It is easier, however, to consider the measurement with a ruler of the length of a rod; we let the continuous random variable x represent the result of *one* measurement. If the true length of the rod is x_0, Eq. (10.18) specifies that a result between x and $x + dx$ will be obtained with a frequency

$$f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_0 - x}{\sigma}\right)^2\right] dx. \tag{10.26}$$

One may also say that the probability that the measurement will "yield a result x" between x and $x + dx$ is given by Eq. (10.26). In simpler words, if N measurements are performed, a result between x_1 and x_2 is likely to be obtained in $n(x_1, x_2)$ of these measurements, where

$$n(x_1, x_2) = N \cdot F(x_1, x_2) = \frac{N}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} \exp\left[-\frac{1}{2}\left(\frac{x_0 - x}{\sigma}\right)^2\right] dx \tag{10.27}$$

as shown in Fig. 10.5.

Note that in Eqs. (10.26) and (10.27) the standard deviation σ is determined by the conditions of the measurement. The applicability of the Gaussian distribution to the results obtained from such measurements lies in the fact that: (a) n, the number of (least) divisions of the ruler, is large and (b) the errors in measurement $|x_0 - x|$ are small as compared to x.

In Table 10.3 are given the values of $f(x)$ and its integral, $F(c)$, for the normalized Gaussian function (Eq. (10.22)).

From Table 10.3, for example, we see that half of the measurements do yield a result x between

$$x_0 - 0.69\sigma < x < x_0 + 0.69\sigma$$

or that only 2.23% of the results may yield x, such that

$$x > x_0 + 2\sigma.$$

TABLE 10.3 Some Numerical Values of the Normalized Gaussian Function

$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$	$F(-c, c) = \int_{-c}^{+c} f(x)\,dx$
$f(0) = 0.3989$	$F(-1, 1) = 0.6826$
$f(1) = f(-1) = 0.2420$	$F(-2, 2) = 0.9554$
$f(2) = f(-2) = 0.0540$	$F(-3, 3) = 0.9974$
	$F(-0.69, 0.69) = 0.5000$

As another example we see that a result x in the small interval Δx about x_0, will be obtained $(0.3989)/(0.0540) = 7.4$ times more frequently than a result in the same small interval Δx about $x_0 + 2\sigma$.

10.3. ESTIMATION OF PARAMETERS AND FITTING OF DATA

In Section 10.1 the basic definitions were given; in Section 10.2, analytic expressions for some frequency functions were obtained. We will now see how statistics can be applied to the interpretation of a measurement or an experiment.

We can consider one or more measurements to form a sample of a population that obeys a certain frequency function; we are then faced with one of two estimation problems:

(a) Given the frequency function and its parameters, what is the probability of obtaining from a measurement the result x?

(b) Given the result x of a measurement, what are the parameters of the frequency function (or the frequency function itself)?

In physics we are usually faced with estimation of type (b), since a set of experimental data are obtained, and it is then desired to reduce them to a few parameters that should describe the whole population and therefore, also any new measurement that may be performed.

There are several methods for obtaining "estimators" to an unknown parameter. Some of these methods are almost subconsciously applied, but most of them can be derived from the principle of "maximum likelihood" introduced by R. A. Fisher in 1920.

10.3.1. Maximum Likelihood

To apply this principle we must have knowledge of the normalized frequency functions of the variables x_i that form the data,

$$f(x_i, \theta),$$

where θ is the parameter to be estimated and upon which the frequency function depends. We may then form the product of the frequency functions for all observed variables,

$$\mathcal{L}(x_1, x_2, \ldots x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta), \tag{10.28}$$

which is called the likelihood function for the parameter θ (note that $\mathfrak{L}$ is *not* a frequency function for the parameter θ). The theorem of maximum likelihood then states that the value of θ, θ^*, that maximizes $\mathfrak{L}$ (for the set of observed data) is the best estimator of θ:

$$\left.\frac{\partial \mathfrak{L}(x_1, x_2, \ldots x_n, \theta)}{\partial\theta}\right|_{\theta=\theta^*} = 0.$$

In practice, it is almost always convenient to work with the logarithm of $\mathfrak{L}$, since when $W = \log \mathfrak{L}$ is maximum, so will also be $\mathfrak{L}$.

As an example, we consider a set of n data x_i that obey a normal frequency function about a, with a standard deviation σ; let us seek the best value for the parameter a:

$$f(x_i, a) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{a - x_i}{\sigma}\right)^2\right]. \qquad (10.29)$$

Then

$$\mathfrak{L} = \prod_{i=1}^{n} f(x_i, a)$$

and

$$W = \log \mathfrak{L} = -n \log\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{a - x_i}{\sigma}\right)^2. \qquad (10.30)$$

$$\frac{\partial W}{\partial a} = -\sum_{i=1}^{n} \frac{a - x_i}{\sigma^2}.$$

Setting $(\partial W)/(\partial a) = 0$ leads to the estimator a^*;

$$\sum_{i}^{n} \frac{a^* - x_i}{\sigma^2} = 0 \qquad \frac{na^*}{\sigma^2} - \sum_{i}^{n} \frac{x_i}{\sigma^2} = 0$$

or

$$a^* = \frac{1}{n}\sum_{1}^{n} x_i. \qquad (10.31)$$

Thus if a set of measurements is distributed *normally*, the best estimator for the true value of the parameter is the *mean* of the measurements (first moment).

Similarly we may obtain the estimator, σ^*, for σ, by differentiating Eq. (10.30) with respect to σ

$$\frac{\partial W}{\partial \sigma} = -\frac{n}{\sigma} + \sum_i^n \left[\left(\frac{a - x_i}{\sigma} \right) \left(\frac{a - x_i}{\sigma^2} \right) \right]$$

and setting $\partial W / \partial \sigma = 0$ gives

$$(\sigma^*)^2 = \frac{1}{n} \sum_i^n (a - x_i)^2, \tag{10.32}$$

where, in Eq. (10.32), a should be replaced by its estimator a^* given by Eq. (10.31). Again we obtain the familiar result that the best estimator for the standard deviation of the theoretical frequency function is given by the second moment (about the mean) of the observed measurements.

The principle of maximum likelihood can be further extended to give the variance S^2 of the estimator θ^*; that is, if the determination of estimators θ^* is repeated, the values so obtained will have a standard deviation S, where

$$\frac{1}{S^2} = -\frac{\partial^2 W}{\partial \theta^2}. \tag{10.33}$$

We may apply Eq. (10.33) to our sample of measurements that obeys a normal frequency function, where W was given by Eq. (10.30). We obtain

$$\frac{1}{S^2} = -\frac{\partial^2 W}{\partial a^2} = \sum_i^n \frac{1}{\sigma^2} = \frac{n}{\sigma^2}.$$

Thus the standard deviation of the estimator will be

$$S = \frac{\sigma}{\sqrt{n}}, \tag{10.34}$$

where n is the number of measurements used for obtaining each estimator. Equation (10.34) is a well-known result that we will obtain again when we discuss the combination of errors in Section 10.4.

10.3.2. The Least-Squares Method

Until now we have discussed the case where all n measurements are made on the same physical quantity whose true value is a, for example, the data of

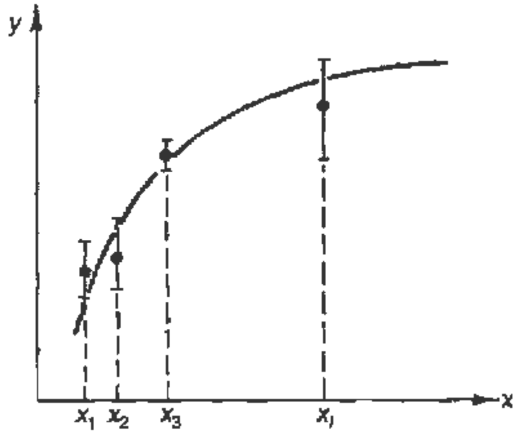


FIGURE 10.6 Least-squares fit of a two-dimensional curve to a set of data points obtained for different values of x. Note that each data point has associated with it a different error as indicated by the flags; this is taken into account when forming the least-squares sum.

Eq. (10.29). However, consider now a set of measurements yielding values $y_1, y_2, \ldots, y_n$ depending on another variable x; the corresponding true values of y, which we designate by $\bar{y}$, are assumed to be a function of x and of one or more parameters α_ν common to the whole sample. Thus we write

$$\bar{y}_i = y(x_i; a_\alpha, \ldots, a_\nu). \qquad (10.35)$$

Further, each measurement y_i has associated with it a standard deviation σ_i, which is not the same for each point. This situation is shown in Fig. 10.6.

It is possible that the form of Eq. (10.35) is known or may be correctly inferred from the physics of the process under investigation, in which case the estimation is reduced to finding the best estimators for the parameters a_ν. If, however, the form of Eq. (10.35) is not known, various functional relationships must be assumed, for example, a polynomial of order k. We then speak of fitting a curve to the data. Even though special techniques are developed in Section 10.3.4 to ascertain which curve fits best, the following discussion is generally applicable.

The method of least squares follows directly from the assumption that each individual measurement y_i is a member of a *Gaussian* population with a mean given by the true value of y_i, $\bar{y}(x_i; a_\lambda)$; for the standard deviation of this Gaussian we use the experimental error σ_i of each measurement. Then in analogy to Eq. (10.29) we write for the frequency function of y_i

$$f(y_i; x_i; a_\lambda) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{y_i - \bar{y}(x_i; a_\lambda)}{\sigma_i}\right]^2\right\}, \qquad (10.36)$$

and in analogy with Eq. (10.28) we form the likelihood function

$$\mathcal{L}(y_1 \cdots y_n; x_1 \cdots x_n; a_\lambda) = \prod_{i=1}^{n} f(y_i; x_i; a_\lambda).$$

We seek the estimators a_λ^* that maximize this function, or its logarithm W

$$\begin{aligned} W &= \log \mathcal{L} \\ &= -\sum_{i=1}^{n} \log\left(\sigma_i\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{n}\left[\frac{y_i - \bar{y}(x_i; a_\lambda)}{\sigma_i}\right]^2. \end{aligned} \qquad (10.37)$$

Since the values of σ_i are fixed by the measurement, the estimators a_λ^* are those values of a_λ that *minimize* the sum

$$\mathcal{M} = \sum_{i=1}^{n} \frac{[y_i - \bar{y}(x_i; a_\lambda)]^2}{\sigma_i^2}, \qquad (10.38)$$

that is, those that give the "least-squares sum." They are obtained by solving the simultaneous equations

$$\frac{\partial \mathcal{M}}{\partial a_\lambda} = 0 \qquad \lambda = 1 \text{ to } \nu.$$

10.3.3. Application of the Least-Squares Method to a Linear Functional Dependence

The simplest case of functional dependence $y(x)$ is the linear one:

$$y = ax + b.$$

If we assume that every measurement y_i has the *same standard deviation* (statistical weight), we may obtain the estimators a^* and b^* that minimize Eq. (10.38) in closed form.

Since $\sigma_1 = \sigma_2 = \cdots = \sigma_n = \sigma$, instead of Eq. (10.38) we need only minimize

$$\mathcal{R} = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2. \qquad (10.39)$$

Hence

$$\frac{\partial \mathfrak{R}}{\partial a} = -2\sum_{i=1}^{n}[y_i - (a + bx_i)] = 0$$
$$\frac{\partial \mathfrak{R}}{\partial b} = -2\sum_{i=1}^{n}\{[y_i - (a + bx_i)]x_i\} = 0, \tag{10.40}$$

which after some manipulation[4] leads to

$$a^* = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum (x_i y_i)}{n\sum x_i^2 - \sum x_i \sum x_i}$$
$$b^* = \frac{n\sum (x_i y_i) - \sum y_i \sum x_i}{n\sum x_i^2 - \sum x_i \sum x_i}. \tag{10.41}$$

The standard deviations for the above estimators may be obtained by an extension of Eq. (10.33), which now yields a symmetric square matrix

$$\mathbf{H}_{\lambda\upsilon} = -\frac{\partial^2 W}{\partial a_\lambda \partial a_\mu} = \frac{1}{2\sigma^2}\frac{\partial^2 \mathcal{M}}{\partial a_\lambda \partial a_\mu}. \tag{10.42}$$

The elements of the *inverse* matrix give the variance of the estimators a^*. A complete discussion of this error matrix is given in Section 10.4; suffice it to say here that the usually given expressions (Eqs. (10.43)) for the standard deviation of the estimators (Eqs. (10.41)) are the square roots of the diagonal elements of $\mathbf{H}^{-1}$ (see Eq. (10.63)). We then obtain

$$\sigma_{a^*} = \sqrt{\left(\mathbf{H}^{-1}\right)_{aa}} = \sigma\sqrt{\frac{\sum x_i^2}{n\sum x_i^2 - \sum x_i \sum x_i}}$$
$$\sigma_{b^*} = \sqrt{\left(\mathbf{H}^{-1}\right)_{bb}} = \sigma\sqrt{\frac{n}{n\sum x_i^2 - \sum x_i \sum x_i}}. \tag{10.43}$$

In case $\sigma_1 \neq \sigma_2 \neq \cdots \neq \sigma_n$, it is $\mathcal{M}$ and not $\mathfrak{R}$ that must be minimized.

Clearly, such calculations are best done using computer programs. In fact, many packages and self-contained programs that are designed to handle these kinds of problems are available (both commercially and

[4] Note that the second of the above equations is by no means equal to the first one multiplied by x_i.

through "shareware"). In this book, we default to MATLAB (see Appendix B), which is in fact well suited for dealing with problems formulated in terms of matrices. For the problem of linear (or, more generally, polynomial) function fitting with equally weighted data points, MATLAB provides the polyfit utility for exactly this purpose.

For more general problems, the reader is referred to other textbooks on the subject of data analysis. For example, the problem of linear fitting with unequally weighted data points is discussed in Chapter 5 of *Numerical Methods for Physics*, 2nd ed., by Alejandro Garcia (Prentice-Hall, Englewood Cliffs, NJ, 2000). A program linreg for this task, is described and the code is available online from the publisher as a MATLAB m-file, as well as in the languages C++ and FORTRAN.

10.3.4. Goodness of Fit; the χ^2 Distribution

We have seen how the least-squares method, as a consequence of the principle of maximum likelihood, may be used to fit a curve to a set of data. Once the curve has been found, however, the necessity to ascertain quantitatively how good the fit is arises. This is important especially if the functional dependence is not known, a poor fit might indicate the necessity for fitting with a curve of higher order, or a poor fit might indicate inconsistencies in the data.

Similarly, we may wish to test whether a certain hypothesis is supported by the data, in which case the goodness of the fit may establish the level of confidence with which the hypothesis should be accepted.

Let us first suppose that we know the true functional relationship of y to x, that is, $\bar{y}(x) = f(x)$; we may then form the least-squares sum

$$\mathcal{M} = \sum_{i=1}^{n} \frac{[y_i - \bar{y}(x_i)]^2}{\sigma_i^2}. \tag{10.38}$$

The range of $\mathcal{M}$ is $0 < \mathcal{M} < +\infty$ but we would be surprised if $\mathcal{M} = 0$ and would be equally surprised if $\mathcal{M}$ was extremely large. Thus we have already a quantitative indication as to how well the data fit the known (or assumed) curve $y = f(x)$.

If a new set of data pertaining to the same experimental situation is obtained, and Eq. (10.38) is again formed, a new value $\mathcal{M}$ will result. Clearly, if enough such measurements are repeated, each time yielding a value for $\mathcal{M}$, we will obtain the frequency function for $\mathcal{M}$. Once the frequency function is known, it is then easy to tell what the probability of obtaining a specific $\mathcal{M}$ is. We may, for example, calculate that in 95% of the cases $\mathcal{M} < \mathcal{M}_0$; if then a specific set of data yields $\mathcal{M}_s \geq \mathcal{M}_0$, we know that such data should be obtained only in 5% of the experiments and can therefore be rejected.

Obtaining the frequency function for the least-squares sum in this way is obviously impractical. Nevertheless, it is true that the *distribution* of $\mathcal{M}$ is *independent* of the curve $y = f(x)$ and of σ_i, and can therefore be calculated theoretically; it depends only on the number n of points that are compared, and is called the χ^2 distribution (pronounced "chi-squared")

$$f(\mathcal{M})\,d\mathcal{M} = \frac{\mathcal{M}^{(\nu/2)-1}\exp(-\mathcal{M}/2)}{2^{\nu/2}\Gamma(\nu/2)}\,d\mathcal{M} \equiv f(\chi^2)\,d\chi^2, \tag{10.44}$$

where ν is the number of "degrees of freedom" of $\mathcal{M}$. In the present case we set

$$\nu = n$$

because this is the number of truly independent points being compared. In Eq. (10.44) $\Gamma(x)$ is the "gamma function," which for positive integer arguments[5] is simply

$$\Gamma(n) = (n-1)!.$$

Consider next that $y = f(x)$ is not known, but that a two-parameter curve is fitted to n data points, yielding estimators a^* and b^*. Then one forms again the least-squares sum $\mathcal{M}$ using $\bar{y} = f(x; a^*, b^*)$ but now the frequency function for the $\mathcal{M}$ values is given by Eq. (10.44) with the n degrees of freedom reduced by the number of estimators obtained from the data, that is,

$$\nu = n - 2.$$

The χ^2 distribution may also be used for comparing the frequency of occurrence of a class of events with the theoretical frequency (function). Let us consider, for example, 100 measurements of a radioactive sample, and divide the sample into seven classes, with mean value $\overline{N} = 85$ counts/min

[5] The general definition of the gamma function is

$$\Gamma(z) = \int_0^\infty t^{z-1}\exp(-t)\,dt;$$

for more details see any text on advanced calculus.

TABLE 10.4 Observed and Expected Frequencies of the Results of 100 Measurements of a Radioactive Sample

Class	0–75	75–79	79–83	83–87	87–91	91–95	95–∞	Counts/min
o_i	15	11	15	15	18	12	14	Observed freq
e_i	13	12	15	16	16	13	15	Expected freq
$(e_i - o_i)^2/e_i^2$	0.307	0.083	0	0.062	0.25	0.077	0.067	χ^2

and approximately equal expected frequencies; the resulting frequency of the experimental observations o_i in each class is given in Table 10.4. Next we obtain from the data the estimators for the parameters of a Gaussian (1) $\mu^* = \overline{N}$, (2) $\sigma^* = \sqrt{N}$, and (3) the overall normalization, namely, $\sum o_i = \sum e_i$; thus the degrees of freedom of χ^2 are four, corresponding to seven classes less three estimators. From the Gaussian distribution we calculate the expected frequencies e_i for each class; they are also given in Table 10.4.

In complete analogy with the least-squares sum, Eq. (10.38), we form the χ^2 sum

$$\chi^2 = \sum_{i=1}^{n} \frac{(e_i - o_i)^2}{e_i^2}.$$

Note that χ^2 is now a discrete variable, since frequencies of classes are compared; however, Eq. (10.44), which holds for a continuously variable χ^2, is valid provided the number of classes $n \gtrsim 5$ and the expected frequencies $e_i \geq 5$.

For this experiment we obtain

$$\chi^2 = 0.846,$$

and we explained before that $\nu = 4$. From a table of the χ^2 distribution we find that in 93% of the cases the χ^2 distribution would be larger than the result obtained here. Thus one may suspect that the data are "too good" a fit to the estimated Gaussian.

The χ^2 distribution of Eq. (10.44) for different degrees of freedom is shown in Fig. 10.7. Tables of this distribution may be found in reference manuals, or easily calculated in any number of computer programs. It should not be surprising that when the number of degrees of freedom

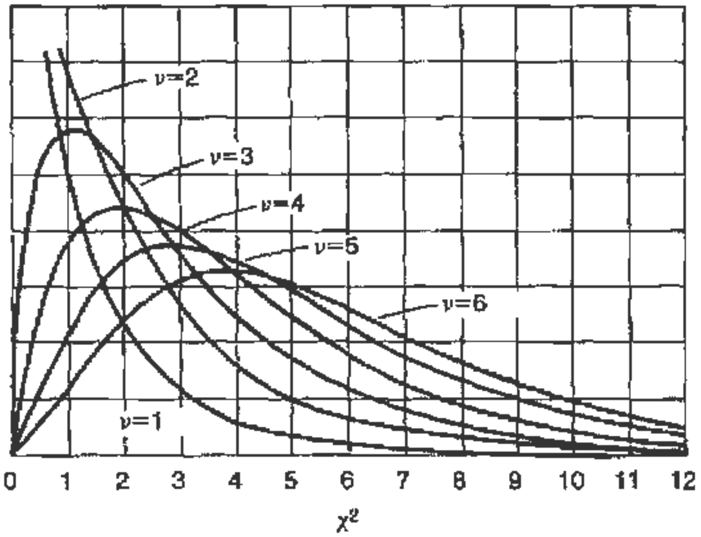


FIGURE 10.7 The frequency function for the distribution of χ^2, for different degrees of freedom. All curves are normalized to the same unit area. Note that for large ν the χ^2 distribution approaches a Gaussian.

increases $\nu > 30$, the χ^2 distribution approaches a Gaussian[6] with mean $\mu = \nu - 1/2$.

10.4. ERRORS AND THEIR PROPAGATION

10.4.1. Introduction

When we perform a measurement of a physical quantity x, it can be expected that the result obtained, x_1, will differ from x; this difference is the *error* of the measurement and consists of a *systematic* and a *random* contribution. Suppose, now, that the measurement is repeated under the same conditions n times; then the results x_n will be distributed (in most cases) normally about a mean $\bar{x}$ with a standard deviation σ. The difference between $\bar{x}$ and the true value x is then the systematic error, and the standard deviation σ of the Gaussian is a measure of the *dispersion* of the results due to the random error.

The object of the measurement, however, is the determination of the unknown true value x; since this is not possible, we seek to find whether x lies between certain limits, or whether the true value x is distributed

[6] It is really the distribution of $\sqrt{2\chi^2}$ that approaches the Gaussian with mean $\mu = \sqrt{(2\nu - 1)}$ and unit standard deviation (R. A. Fisher's approximation).

about some mean x^* with a standard deviation σ^*. Note that in a rigorous sense, this statement is incorrect, since the unknown true value x is not distributed, but is fixed; what we mean is that the probability, $x = x^*$, $x > x^*$, etc., is given by the normal frequency function with mean $\bar{x}$ and $\sigma = \mu_2$, the second moment of the measured data about their mean $\bar{x}$.

Thus, by repeating the measurement several times, it is possible in principle to circumvent the random errors because (a) a knowledge'of $\bar{x}$ and σ contains all possible information about the unknown true value x, and (b) as n increases, the second moment should decrease as $1/\sqrt{n}$ and may be made arbitrarily small. On the other hand, the systematic errors cannot be extracted from a set of identical measurements. They can either be estimated by the observer or be judged from a performance of the same measurement with a different technique. Therefore, it is unadvisable to reduce the random errors much below the expected limits of the systematic errors. In what follows we will discuss only the treatment of random errors and work under the assumption that the results of the measurements follow a normal distribution.

Until now we have considered the simple case where the unknown value x is directly measured and an error σ_x can be associated with the measurement; that is, the frequency function of x depends only on one variable:

$$f(x) = \frac{a}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{1}{2}\left(\frac{\bar{x}-x}{\sigma}\right)^2\right].$$

Most frequently, however, the unknown value x is not directly measured, and we distinguish two cases:

(a) x is an explicit function of the quantities $y_1, y_2, \ldots, y_n$ that are measured and have with them associated errors $\sigma_1, \sigma_2, \ldots, \sigma_n$. Namely,

$$x = \phi(y_1, y_2, \ldots, y_n), \tag{10.45}$$

and it is desired to find the estimator x^* and its standard deviation σ_x.

(b) x is an implicit function of other unknown variables $u_1, u_2, \ldots, u_m$, and of the quantities $y_1, y_2, \ldots, y_n$ that are measured and have with them associated errors $\sigma_1, \sigma_2, \ldots, \sigma_n$. Namely,

$$\phi(x; u_1, u_2, \ldots, u_m; y_1, y_2, \ldots, y_n) = 0, \tag{10.46}$$

and it is desired to find the estimators $x^*, u_1^*, u_2^*, \ldots, u_m^*$ and the symmetric error matrix $\sigma_{ij}(i, j = 1, \ldots, m+1)$. Such an example was treated in Section 10.3.3, and we know that at least $m+1$ sets of measurements are required to obtain the $m+1$ estimators.

The techniques for obtaining the best estimators were discussed in Section 10.3. In this section we will discuss how the random error of x may be determined from knowledge of the errors of the independent variables y_n; this procedure is frequently referred to as the combination or the propagation of the errors of the measured values y_n.

10.4.2. Propagation of Errors

Let us first assume x to be an explicit function of the measured y_n as discussed previously (Section (10.4.1)):

$$x = \phi(y_1, y_2, \ldots, y_n). \tag{10.45}$$

By applying the maximum likelihood method, it can be shown that the estimator x^* is obtained by using the mean values, μ_n, of the measured y_n (provided[7] the y_n are distributed normally). Here the mean values μ_n are obtained from r different measurements

$$\mu_n = \frac{1}{r}\sum_{i=1}^{r}(y_n)^i.$$

Thus

$$x^* = \phi(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n) = \phi(\mu_1, \mu_2, \ldots, \mu_n). \tag{10.47}$$

Next we make a Taylor expansion of Eq. (10.45) about x^*, through first order

$$x = \phi(\mu_1, \mu_2, \ldots, \mu_n) + \left[\frac{\partial\phi}{\partial y_1}\right]_\mu (\mu_1 - y_1)$$
$$+ \left[\frac{\partial\phi}{\partial y_2}\right]_\mu (\mu_2 - y_2) + \cdots + \left[\frac{\partial\phi}{\partial y_n}\right]_\mu (\mu_n - y_n).$$

[7] Clearly if x is variable, all measurements y_n^i are made so as to correspond to the same point x.

where $[\partial\phi/\partial y_n]_\mu$ means evaluation of the derivative at the point about which we expand—that is, $(\mu_1, \mu_2, \ldots, \mu_n)$. We can now form the second moment of the distribution of the x^i values as they result from the observed $y_n{}^i$ values. The superscript i here refers to the r different sets of measurements:

$$\begin{aligned}
\sigma_x^2 &= \frac{1}{r}\sum_{i=1}^{r}\left(\bar{x} - x^i\right)^2 \\
&= \frac{1}{r}\sum_{i=1}^{r}\left[\left(\frac{\partial\phi}{\partial y_1}\right)_\mu(\mu_1 - y_1^i) + \cdots + \left(\frac{\partial\phi}{\partial y_n}\right)_\mu(\mu_n - y_n^i)\right]^2 \\
&= \left(\frac{\partial\phi}{\partial y_1}\right)_\mu^2 \frac{1}{r}\sum_{i=1}^{r}(\mu_1 - y_1^i)^2 + \left(\frac{\partial\phi}{\partial y_2}\right)_\mu^2 \frac{1}{r}\sum_{i=1}^{r}(\mu_2 - y_2^i)^2 + \cdots \\
&\quad + 2\left(\frac{\partial\phi}{\partial y_1}\right)_\mu \left(\frac{\partial\phi}{\partial y_2}\right)_\mu \frac{1}{r}\sum_{i=1}^{r}(\mu_1 - y_1^i)(\mu_2 - y_2^i) + \cdots \\
\sigma_x^2 &= \left(\frac{\partial\phi}{\partial y_1}\right)_\mu^2 \sigma_1^2 + \left(\frac{\partial\phi}{\partial y_2}\right)_\mu^2 \sigma_2^2 + \cdots + 2\left(\frac{\partial\phi}{\partial y_1}\right)_\mu \left(\frac{\partial\phi}{\partial y_2}\right)_\mu \sigma_{12}^2 + \cdots .
\end{aligned} \tag{10.48}$$

Equation (10.48) is the most general expression for the propagation of errors. If we assume that the errors are uncorrelated, namely, $\sigma_{ij} = 0$ when $i \neq j$, we can obtain the results for the simplest functional relationships:

(a) Addition

$$x = y_1 + y_2 + \cdots + y_n$$

$$\sigma_x = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}. \tag{10.49}$$

(b) Subtraction

$$x = y_1 - y_2$$

$$\sigma_x = \sqrt{\sigma_1^2 + \sigma_2^2}. \tag{10.50}$$

(c) Multiplication

$$x = y_1 \times y_2 \times \cdots \times y_n$$

$$\left(\frac{\partial\phi}{\partial y_1}\right)_\mu = \mu_2 \times \cdots \mu_n$$

$$\sigma_x = \sqrt{\sigma_1^2 \times (\mu_2 \cdots \mu_n)^2 + \cdots + \sigma_n^2 \times (\mu_1\mu_2 \cdots)^2} \qquad (10.51)$$

$$= x^* \sqrt{\left(\frac{\sigma_1}{\mu_1}\right)^2 + \left(\frac{\sigma_2}{\mu_2}\right)^2 + \cdots + \left(\frac{\sigma_n}{\mu_n}\right)^2}.$$

(d) Division

$$x = \frac{y_1}{y_2}$$

$$\left(\frac{\partial\phi}{\partial y_1}\right)_\mu = \frac{1}{\mu_2}, \qquad \left(\frac{\partial\phi}{\partial y_2}\right)_\mu = \frac{-\mu_1}{(\mu_2)^2} \qquad (10.52)$$

$$\sigma_x = \sqrt{\frac{\sigma_1^2}{(\mu_2)^2} + \frac{\sigma_2^2(\mu_1)^2}{(\mu_2)^4}} = x^* \sqrt{\left(\frac{\sigma_1}{\mu_1}\right)^2 + \left(\frac{\sigma_2}{\mu_2}\right)^2}. \qquad (10.53)$$

From the above examples we see that in general the errors are combined in quadrature—that is, it is their squares that are added. Consequently, if the error in one of the variables σ_i is large, it will dominate all other terms and the error of x, σ_x, will be almost equal to σ_i, despite good measurements made on the other independent variables.

Our simple rule for the case of addition, Eq. (10.49), may be used to obtain in a different way the result derived in Eq. (10.34). Let a variable x be measured and let the mean of a set of measurements be $\bar{x}_i$, with a standard deviation σ_i; if this set of measurements is repeated under identical conditions, a new mean result $\bar{x}_j \neq \bar{x}_i$ will be obtained, but let the standard deviations be equal, that is, $\sigma_j = \sigma_i$. If n such sets of measurements are performed, the new estimator for x will be

$$x^* = \frac{1}{n}(\bar{x}_1 + \bar{x}_2 + \cdots \bar{x}_n),$$

and thus

$$\left(\frac{\partial\phi}{\partial \bar{x}_i}\right) = \frac{1}{n}.$$

Hence, from Eq. (10.48) or (10.49),

$$\sigma_x^* = \sqrt{\left(\frac{\sigma_1}{n}\right)^2 + \left(\frac{\sigma_2}{n}\right)^2 + \cdots + \left(\frac{\sigma_n}{n}\right)^2} = \sqrt{n\frac{\sigma^2}{n^2}} = \frac{\sigma}{\sqrt{n}}. \qquad (10.54)$$

Namely, the standard deviation of the mean of n measurements of a Gaussian distribution is $\sigma/\sqrt{n}$, where σ is the standard deviation of the individual measurements.

10.4.3. Example of Calculation of Error Propagation

As an example, let us consider an experiment to determine Stefan's constant b, from the relation

$$E = bT^4,$$

where the following values of E and T were obtained with the indicated standard deviations:

T (K)	E (W/m^2)
$800(1 \pm 0.02)$	$(3.0 \pm 0.3) \times 10^4$
$1000(1 \pm 0.02)$	$(8.0 \pm 0.8) \times 10^4$
$1200(1 \pm 0.02)$	$(15.6 \pm 0.6) \times 10^4$

We wish to calculate the estimator b^* and its standard deviation σ_b.

There are two ways to proceed in this case. We either may calculate b_j^* from each of the three sets of measurements and then combine these values to obtain $b^* = \bar{b}_j^*$, but weighing each b_j^* according to its standard deviation, or we may use least squares in the observed variables E and T^4. Note that a mean of T or E of the three listed measurements makes no sense whatsoever since each measurement is made for a *different* T.

We will follow the first procedure, and we first obtain the error on T^4 from the known error on T. For this we should use the general expression, Eq. (10.48), but since $\phi = T^4$ is a function of only one variable,[8] simple differentiation gives the desired result directly

$$\frac{d\phi}{dT} = 4T^3 \qquad \frac{\Delta\phi}{\phi} = 4\frac{\Delta T}{T}. \qquad (10.55)$$

[8] If we choose to write $\phi = T \times T \times T \times T$, we may *not* apply Eq. (10.51), since these variables are correlated; use of Eq. (10.48) and $\sigma_{TT} = \sigma_T$ gives back the result of Eq. (10.55).

TABLE 10.5 An Example of a Calculation of Propagation of Errors

Set of data	T^4	$E/T^4 = b_j^*$	$\sigma(T^4)/T^4$	$\sigma(b_j)/b_j^*$
1	0.41×10^{12}	7.3×10^{-8}	0.08	0.13
2	1.0×10^{12}	8.0×10^{-8}	0.08	0.13
3	2.0×10^{12}	7.8×10^{-8}	0.04	0.06

We note from Eq. (10.54) that it is easier to work with relative errors, and we thus form Table 10.5, where

$$\frac{\sigma(b)}{b} = \sqrt{\left[\frac{\sigma(T^4)}{T^4}\right]^2 + \left[\frac{\sigma(E)}{E}\right]^2}$$

since the errors in T and E are uncorrelated.

For the best estimator of b, we will use the mean of the three measurements but weighed in inverse proportion to the square of their standard deviation (see Section 10.3.3). Thus

$$\bar{b} = \frac{1}{6}(7.3 + 8.0 + 4 \times 7.8) \times 10^{-8} = 7.75 \times 10^{-8};$$

for $\sigma(\bar{b})$ we used Eq. (10.49),

$$\sigma(\bar{b}) = \frac{1}{6}\sqrt{\sigma^2(b_1) + \sigma^2(b_2) + 4\sigma^2(b_3)}$$

or the convenient approximation

$$\frac{\sigma(\bar{b})}{\bar{b}} = \frac{1}{6}\sqrt{\left[\frac{\sigma(b_1)}{b_1}\right]^2 + \left[\frac{\sigma(b_2)}{b_2}\right]^2 + 4\left[\frac{\sigma(b_3)}{b_3}\right]^2} = 0.043,$$

so that the final result is

$$b^* = 7.75(1 \pm 0.043) \times 10^{-8}\ \mathrm{W/{}^\circ K^4\text{-}m^2}.$$

10.4.4. Evaluation of the Error Matrix

In the two previous sections we have discussed the case where only one unknown variable x was sought. We will now consider the random errors when several unknown variables are simultaneously estimated or measured.

When only one variable is measured, we know how to obtain from the data the second moment about the mean

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2.$$

If now p variables are simultaneously measured in an experiment, we must form the $p(p+1)/2$ second moments about the mean; for example, if we measure x, y, and z, we must calculate the six expressions

$$\begin{aligned}
&\sigma_{xx} = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)(\bar{x} - x_i); \qquad \sigma_{yy} = \cdots; \qquad \sigma_{zz} = \cdots;\\
&\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)(\bar{y} - y_i) = \sigma_{yx}; \qquad (10.56)\\
&\sigma_{xz} = \cdots = \sigma_{zx}; \qquad \sigma_{yz} = \cdots = \sigma_{zy}.
\end{aligned}$$

(In this notation, the dimensionality of a quantity σ_{pq} is that of the product pq. Hence, σ_x^2 has the same dimensions as σ_{xx}. We avoid the notation σ_{xx}^2, etc., because it misleads one to think that σ_{xy}, for example, is positive definite.) If the distribution of the variables x, y, and z is normal, then these six moments form the symmetric error matrix; if the variables are uncorrelated, the matrix is diagonal.

Clearly, the error matrix must be known if it is desired to apply Eq. (10.48). Consider, for example, that from the measured variables x, y, and z we wish to obtain a new unknown u and its standard deviations $\sigma(u)$, where

$$u = \phi(x, y, z). \qquad (10.57)$$

Then the values of σ_{ij}^2 that were obtained from the data with the help of Eq. (10.56) are substituted in Eq. (10.48) along with the partial derivatives of u, which are obtained from Eq. (10.57).

Conversely, if the frequency function of the three variables x, y, and z, and thus of u, is known,

$$f(u) = f[\phi(x, y, x)]$$

it is possible to calculate theoretically the elements of the error matrix through the usual expression

$$\mu_2'(x, y) = \iiint f(x, y, z)xy\,dx\,dy\,dz \qquad (10.58)$$

or

$$\mu_2(x, y) = \iiint f(x, y, z)(\mu_x - x)(\mu_y - y)\,dx\,dy\,dz,$$

where

$$\sigma_{xy} = \mu_2(x, y), \qquad \text{etc.}$$

In most practical applications, however, it is difficult to use Eq. (10.56) or (10.58). Equation (10.56) may not be usable because the unknown variables may not be measured directly (although they are measured implicitly); also, extensive data are required to yield meaningful results, and the calculation is cumbersome. Equation (10.58) may not be usable because the multidimensional integrals are frequently too difficult to calculate. Instead, the method of maximum likelihood provides an easy way for obtaining the error matrix.

As already discussed in Section 10.3, if the set of data $x_i, y_i, \ldots, z_i$ has been measured, and the estimators for the m unknown variables $\theta_a, \theta_b, \ldots, \theta_m$ are sought, we may form the likelihood function

$$\begin{aligned}\mathcal{L}(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n, \ldots z_1, z_2, \ldots, z_n; \theta_a, \theta_b, \ldots, \theta_m)\\ = f(x_1, y_1, \ldots, z_1; \theta_a, \theta_b, \ldots, \theta_m)\, f(x_2, y_2, \ldots, z_2; \theta_a, \theta_b, \ldots, \theta_m) \cdots\\ \times f(x_n, y_n, \ldots, z_n; \theta_a, \theta_b, \ldots, \theta_m),\end{aligned}$$

where f is the frequency function of the measured variables and is usually assumed to be a product of Gaussians. Then the estimators $\theta_a^*, \theta_b^*, \ldots, \theta_m^*$ are given by the values that *simultaneously* maximize $\mathcal{L}$, namely,

$$\left.\frac{\partial \mathcal{L}}{\partial \theta_a}\right]_{\theta_a^*, \theta_b^*, \ldots, \theta_m^*} = \cdots = \left.\frac{\partial \mathcal{L}}{\partial \theta_m}\right]_{\theta_a^*, \theta_b^*, \ldots, \theta_m^*} = 0, \qquad (10.59)$$

requiring the solution of m coupled equations. Equation (10.41) is a simple example of such a solution of Eq. (10.59). We note that the number of independent data points taken, n, must be larger than or equal to m.

The elements of the error matrix can be obtained from the inverse of the matrix

$$\mathbf{H}_{kl} = \left.\frac{\partial^2 W}{\partial\theta_k \partial\theta_l}\right]_{\theta_a^*, \theta_b^*, \ldots, \theta_m^*}, \tag{10.60}$$

where the second-order partial derivatives must be calculated at the values of the estimators, and $W = \log \mathcal{L}$. We have

$$\sigma_{kl} = (\mathbf{H})_{kl}^{-1},$$

where the rule for matrix inversion is

$$(\mathbf{H}^{-1})_{ij} = (-1)^{i+j} \frac{\text{Det } (ji \text{ minor of } \mathbf{H})}{\text{Det } \mathbf{H}} \tag{10.61}$$

and the minor is the matrix resulting from $\mathbf{H}$ when the jth row and ith column are removed; obviously, the inverse matrix does not exist unless Det $\mathbf{H} \neq 0$.

We will now apply this method of obtaining the error matrix to the simple example treated in Section 10.3.3. The measured variables are x and y, and estimators are sought for the variables a and b; we assume that x is known exactly and that y is distributed normally for each measurement, and related to x through

$$y = a + bx.$$

Using Eq. (10.37), we have

$$\mathcal{L} = \prod_{i=1}^{n} \left[\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma_i^2} [y_i - \bar{y}(x_i; a, b)]^2 \right\} \right]$$

and

$$W = \log \mathcal{L} = -\frac{n}{2} \log(2\pi) - \sum_{i=1}^{n} \log \sigma_i - \frac{1}{2} \sum_{i=1}^{n} \left[\frac{y_i - (a + bx_i)}{\sigma_i} \right]^2.$$

To simplify the calculations we assume $\sigma_1 = \sigma_2 = \cdots = \sigma_n$, so that

$$-\frac{\partial^2 W}{\partial a^2} = \frac{n}{\sigma^2}; \qquad -\frac{\partial^2 W}{\partial a \partial b} = \frac{\sum x_i}{\sigma^2}; \qquad -\frac{\partial^2 W}{\partial b^2} = \frac{\sum x_i^2}{\sigma^2}.$$

Hence

$$\mathbf{H} = \frac{1}{\sigma^2} \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum (x_i^2) \end{bmatrix} \tag{10.62}$$

and

$$\text{Det}\,\mathbf{H} = \frac{1}{\sigma^2}\left[n\sum(x_i^2) - \left(\sum x_i\right)^2\right].$$

Thus

$$\sigma_{\nu\mu} = \mathbf{H}^{-1} = \frac{\sigma^2}{n\sum(x_i^2) - \left(\sum x_i\right)^2}\begin{bmatrix}\sum(x_i^2) & -\sum x_i \\ -\sum x_i & n\end{bmatrix}, \qquad (10.63)$$

which gives the results stated in Eq. (10.43); the indices ν, μ stand for a or b.

10.4.5. The Monte Carlo Method

It is clear that the calculation of the propagation of errors may become extremely involved, especially when the frequency functions of the variables cannot be expressed analytically and when intermediate processes of statistical nature take place. It is then preferable to use computer programs based on the so-called "Monte Carlo" method.

By this technique, we follow a particular event through the sequence of processes it may undergo. For each process, all possible outcomes are weighed according to the frequency function and divided into x classes of equal probability. Then, from a table of these classes, one class is selected at random: for example, by looking up a table of x random numbers. The outcome of this process is incorporated in the progress of the event until a new decision point is reached, when again random selection is made. Thus, at the end of the sequence of all processes, certain final conditions will be reached from the initial conditions with which we started and through the intermediary of the random choices made at each decision point.

We follow in this fashion several events, always starting with the same initial conditions, but because of the random choices, the final conditions will be spread over some range. If enough events have been followed through, we are able to find the frequency function of the combined process and of its parameters, namely, the mean and the standard deviation for the final conditions that result from a given set of initial conditions.

For more discussion, including examples with accompanying computer codes, the reader is referred to the material listed at the end of this chapter.

10.5. THE STATISTICS OF NUCLEAR COUNTING

In many experiments related to nuclear physics, we count the particles or photons emitted in the decay of a nucleus. Usually only a very small fraction of the total sample undergoes such decay. The decay of *one* nucleus is a completely random phenomenon, yet from the number of counts in a given time interval, we may determine the decay probability of this species of nuclei or unstable particles. We have already made use of these`concepts in Chapters 8 and 9.

10.5.1. The Frequency Function for the Number of Decays

We start with the assumption that the decay of *one* nucleus is purely random and the probability (unnormalized) for decay in a time interval Δt is proportional to Δt and some constant λ with dimensions of inverse time[9]:

$$p_d = \lambda \Delta t. \tag{10.64}$$

If we have a sample of N nuclei, since the presence of one nucleus does not affect the decay of another, the probability that *one* nucleus out of the *sample* of N nuclei will decay, in time Δt, is

$$P(1, \Delta t) = \lambda N \Delta t. \tag{10.65}$$

Equation (10.65) is completely analogous to Eq. (10.12) of Section 10.2.6, which leads to the Poisson distribution; the only difference is that the *product* Nt of Eq. (10.65) is the equivalent of the number of trials n of Eq. (10.12). Consequently the probability (frequency function) for obtaining n decays in a time interval t is

$$P(n, t) = \frac{e^{-\lambda N t}(N\lambda t)^n}{n!}. \tag{10.66}$$

The first moment of Eq. (10.66) (in the discrete unknown variable n), as we know from Eq. (10.16), is

$$\overline{n} = \lambda N t. \tag{10.67}$$

[9] E. Schweidler, 1905; this assumption has been proven absolutely correct from the agreement of experiment with the deductions following from Eq. (10.64) as developed in the following paragraphs.

Since $\bar{n}/t$ is the average number of decays per unit time (the average decay rate), we find the physical significance of the constant parameter λ. That is, $N\lambda$ gives the average decay rate of the sample; N is the total number of nuclei in the sample.

Similarly, the second moment about the mean of Eq. (10.66), as we know from Eq. (10.17), is

$$\sigma^2 = \lambda N t = \bar{n}.$$

Hence the very frequently used expression,

$$\sigma = \sqrt{n}. \tag{10.68}$$

Note, however, that $\bar{n}/t = N\lambda$ is the theoretical average rate, which is usually unknown (unless λ and N are precisely known for the sample under consideration). The average rate that we measure, $R = n/t$ (counts per unit time), will, in general, differ from the true rate $N\lambda = \bar{n}/t$, but if n is large, R will be distributed normally about $N\lambda$. (See Eq. (10.66a) below.)

From the considerations of Section 10.2.9, it is clear that when the total number of observed counts n is large, Eq. (10.66) is well approximated by a Gaussian with mean $\mu = N\lambda t$ and standard deviation $\sigma = \sqrt{N\lambda t}$:

$$P(n,t) = \frac{1}{\sqrt{2\pi N\lambda t}} \exp\left[-\frac{(N\lambda t - n)^2}{2N\lambda t}\right] \tag{10.66a}$$

$$= \frac{1}{\sqrt{2\pi n}} \exp\left[-\frac{(\bar{n} - n)^2}{2n}\right]. \tag{10.66b}$$

Thus, unless we are dealing with very few counts, Gaussian statistics may be safely applied.

Finally, we summarize here some simple consequences of Eq. (10.64) for a *single* nucleus:

(a) If the probability for decay in dt is

$$p_d(dt) = \lambda\, dt,$$

(b) then the probability for not decaying (survival) in the time interval from $t = 0$ to $t = t$ is

$$p_s(t) = e^{-\lambda t}$$

(for proof see Eq. (10.13)).

(c) The probability for decay in dt at time t is

$$p_d(t, dt) = e^{-\lambda t}\lambda\, dt.$$

(d) The probability for decay in the time interval from $t = 0$ to $t = t$ is

$$p_d(t) = 1 - p_s(t) = 1 - e^{-\lambda t}.$$

Note that only (c) is properly normalized, so that

$$\int_0^\infty p_d(t)\, dt = \int_0^\infty e^{-\lambda t}\lambda\, dt = 1.$$

Expressions (b) and (d) are, correctly, always <1 and reduce to 0 and 1, respectively, as t approaches infinity. As to expression (a), we must keep in mind that it holds only for Δt such that $\lambda\Delta t \ll 1$.

10.5.2. Behavior of Large Samples

Having obtained the frequency functions, we may now examine the behavior of the total sample.[10] From Eq. (10.67) we see that given a sample of N nuclei, on the average, in a time interval Δt there will be

$$n = \lambda N \Delta t$$

decays; that is, the total sample will be decreased by an amount

$$-\Delta N = N\lambda\Delta t. \tag{10.69}$$

Equation (10.69) then leads to the differential equation for the number of nuclei in the sample

$$\frac{dN}{N} = -\lambda\, dt$$

with solution

$$N(t) = N_0 e^{-\lambda t}, \tag{10.70}$$

where N_0 is the number of nuclei at time $t = 0$. Frequently $\tau = 1/\lambda$ is used for the exponent in Eq. (10.70); τ is called the *lifetime* of that particular species of nuclei and is the time in which the population of the sample is

[10] The principles and formulas in this section have already been used in Section 8.6.

reduced to 37% ($1/e$) of its original value. The *half-life*

$$\tau_{1/2} = \tau\left[\ln_e \frac{1}{2}\right] = 0.693\tau$$

gives the time in which the population of the sample is reduced to half its original value. Using Eq. (10.70) we find, for the *decay rate* as a function of time, that

$$\frac{dN}{dt} = R(t) = -\lambda N(t) = -\lambda N_0 e^{-\lambda t}, \tag{10.71}$$

which has the same time dependence as Eq. (10.70). Experimentally we usually measure $R(t)$ and obtain a curve as shown in Fig. (10.8); from such a plot λ may be obtained. If the sample contains two or more different species of nuclei with different decay constants $\lambda_1, \lambda_2, \ldots$, the time dependence of the decay rate is no longer the simple exponential of Eq. (10.71); instead

$$\frac{dN}{dt} = R(t) = -\lambda_1 N_0^1 e^{-\lambda_1 t} - \lambda_2 N_0^2 e^{-\lambda_2 t} - \cdots .$$

If, however, $\lambda_1 \gg \lambda_2$, then for small t (that is, $t \sim 1/\lambda_1$) $R(t)$ is dominated by the first term; for large t (for example, $t \sim 1/\lambda_2$), $R(t)$ is dominated by

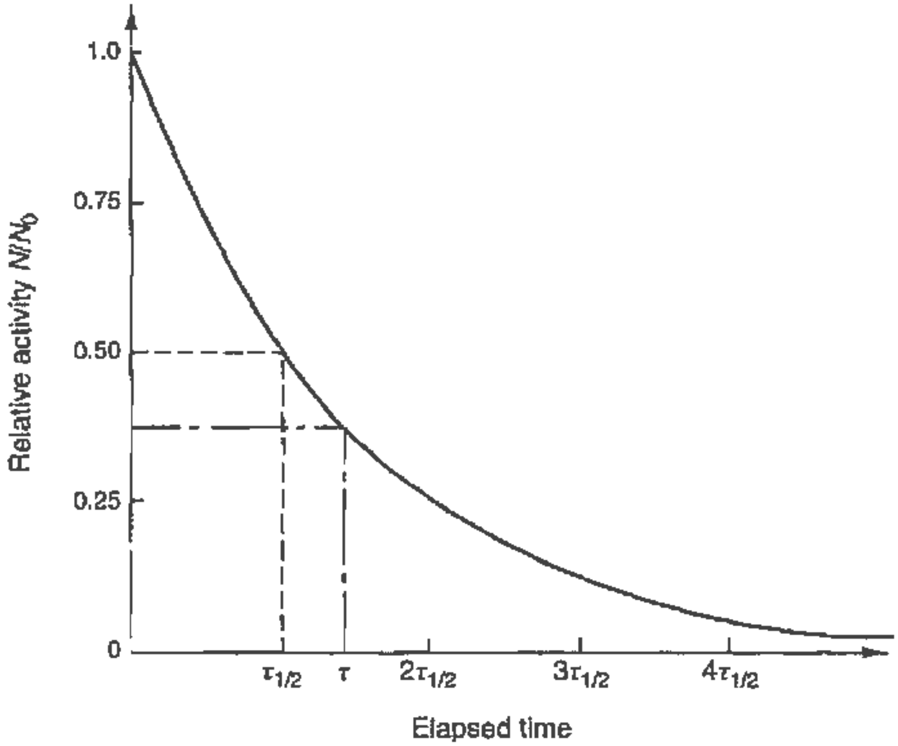


FIGURE 10.8 Exponential decay of a sample of radioactive nuclei. The abscissa is calibrated in units of the half-life of the sample; the lifetime is also indicated.

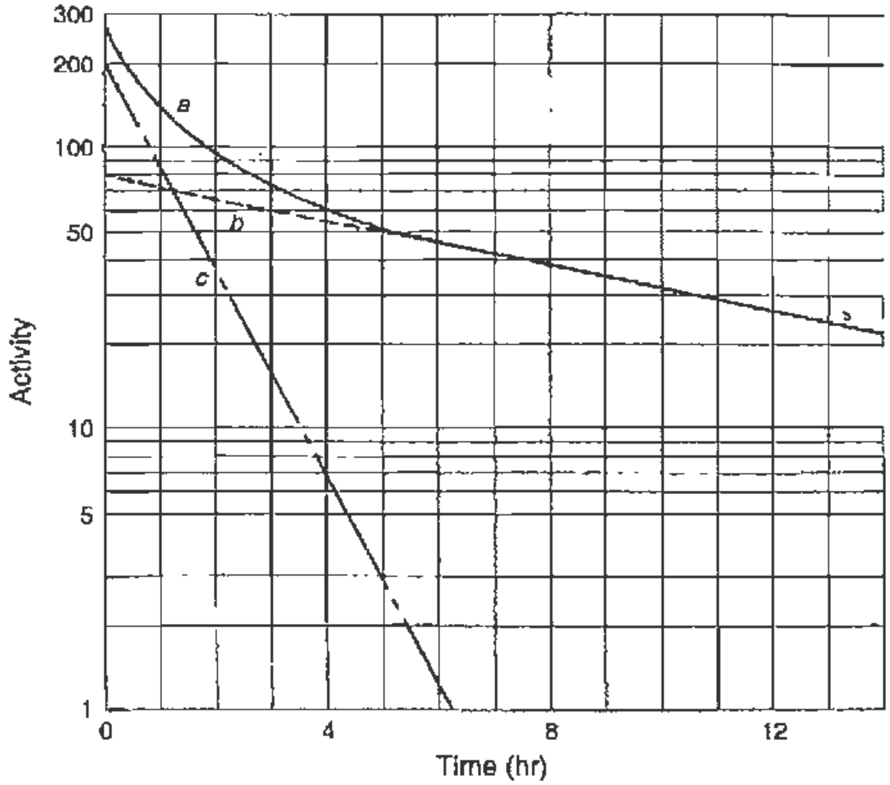


FIGURE 10.9 The decay curve for a sample containing two species of radioactive nuclei, each decaying with a different lifetime. Note that the composite decay curve *a* is the sum of curves *b* and *c*.

the second term. This is shown in Fig. 10.9, which gives the decay curves on a semilogarithmic plot. See also Section 8.6.3, in particular Fig. 8.37.

Another situation of interest arises when nuclei of species A decay into species B with a constant λ_A; nuclei B, however, decay in turn into species C with a constant λ_B. Let, at time $t = 0$, the number of nuclei of species A be N_0 and that of species B be 0.

Then the number of nuclei of species A as a function of time is still given by Eq. (10.70), $N_A = N_0 e^{-\lambda_A t}$. However, for the number of nuclei of species B, the following differential equation holds:

$$\frac{dN_B}{dt} = +\lambda_A N_A - \lambda_B N_B.$$

The solution of this first-order linear differential equation is straightforward, and with the initial condition $N_B(t = 0) = 0$, we have

$$N_B = N_0 \frac{\lambda_A}{\lambda_B - \lambda_A} \left[e^{-\lambda_A t} - e^{-\lambda_B t}\right]. \qquad (10.72)$$

e that Eq. (10.72) always gives $N_B > 0$, as it must be, irrespective vhether $\lambda_A > \lambda_B$ or $\lambda_B > \lambda_A$. Equation (10.72) correctly reduces to $= 0$ for $t = 0$ and $t = \infty$. The two limiting cases for the decay rate n B to C can also be obtained from Eq. (10.72) if we take into account : $R_{BC}(t) = N_B \lambda_B$. Thus

$$\begin{aligned} &\text{for } \lambda_B \gg \lambda_A \qquad R_{BC}(t) \approx N_0 \lambda_A e^{-\lambda_A t} \\ &\text{for } \lambda_A \gg \lambda_B \qquad R_{BC}(t) \approx N_0 \lambda_B e^{-\lambda_B t} \end{aligned}$$

5.3. Testing of the Distribution of Radioactive Decay; the Distribution of the Time Intervals between Counts

s frequently desirable to test whether a sample of counting data does eed come from the decay of radioactive nuclei, that is, that it follows the quency function of Eq. (10.66). A very sensitive test can be devised if plot the distribution of the time intervals between successive decays, or ;ry second, third, etc., decay. This method was applied to the distribution the arrival times of cosmic rays in Section 9.4.2.

First we obtain the distribution of the time intervals between two succese decays. Let $t = 0$ when a decay occurs; we then seek the probability it no decay occurs until $t = t$, but a decay occurs within dt at $t = t$. This)bability is given by Eq. (10.66) with $n = 0$, multiplied by Eq. (10.65); nely,[11]

$$P(t, dt) \equiv q_1(t)\, dt = e^{-N\lambda t} N\lambda\, dt. \tag{10.73}$$

uation (10.73) indicates that the shortest time intervals between two ınts are much more frequent than the longer ones; this is true for any *ıdom* events, since they obey Eq. (10.64) and is shown in Fig. 9.22.

Next we consider the distribution of the time intervals between every :ond, third, etc., mth count. In practice this arises when the counts from : output of a "scaling circuit" are recorded. Consider, therefore, a circuit 'ing one output count for every m input count. If the true rate is r, then : output rate R is related to r by

$$N\lambda = r = Rm.$$

[1] Compare this equation with the probability for the decay of a single nucleus, as given ;ection 10.5.1(c).

= 0 when an output pulse arrives, and let $Q_m(t)$ be the probability another output pulse arrives *in the time interval t*; $q_m(t)\,dt$ will then e probability that this other output pulse arrives *at t* (between t and lt).

nother output pulse will arrive if the input counts $n \geq m$, so that

$$Q_m(t) = \sum_{n \geq m}^{\infty} P(n,t) = \sum_{n \geq m}^{\infty} \frac{(rt)^n e^{-rt}}{n!}$$

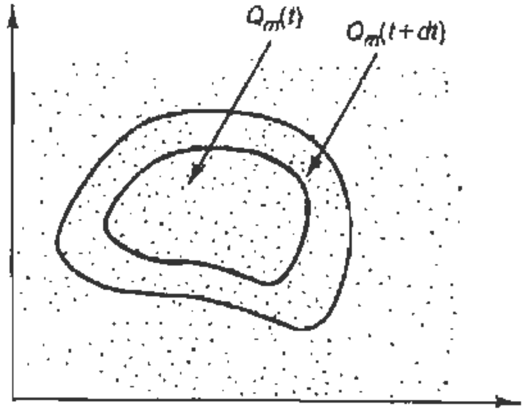
$$= 1 - \sum_{n=0}^{n=m-1} \frac{(rt)^n e^{-rt}}{n!}, \qquad (10.74)$$

e the last equality follows from the normalization of Eq. (10.66)

$$\sum_{n=0}^{\infty} P(n,t) = 1.$$

by considering the sample space of Fig. 10.10 we see that the set of ts $Q_m(t)$ is a subset of $Q_m(t+dt)$, so that any sample-space point nging to $Q_m(t+dt)$ but not to $Q_m(t)$ represents an output count een t and $t + dt$. Thus

$$q_m(t)\,dt = Q_m(t+dt) - Q_m(t)$$



RE 10.10 Sample space indicating the domain $Q_m(t)$, which contains all points ponding to the arrival of an output count in the time interval from 0 to t after the us count. This domain forms a subset of $Q_m(t+dt)$, which contains all points ponding to the arrival of the output count in the time interval from 0 to $t + dt$. The of the output count at t is $q_m(t) = Q_m(t+dt) - Q_m(t)$.

or

$$q_m(t) = \frac{dQ_m(t)}{dt}.$$

Taking the derivative of Eq. (10.74)

$$\begin{aligned} q_m(t) &= -\sum_{n=0}^{n=m-1}\left[\frac{rn(rt)^{n-1}e^{-rt}}{n!} - \frac{r(rt)^n e^{-rt}}{n!}\right] \\ &= r\sum_{n=0}^{n=m-1}\frac{(rt)^n e^{-rt}}{n!} - r\sum_{n=1}^{n=m-1}\frac{(rt)^{n-1}e^{-rt}}{(n-1)!}. \end{aligned}$$

By replacing in the second sum n by $l = n - 1$, we see that only the last term of the first sum survives, so that

$$q_m(t) = r\frac{(rt)^{m-1}e^{-rt}}{(m-1)!}. \tag{10.75}$$

Equation (10.75) correctly reduces to Eq. (10.73) for $m = 1$ (since $r = N\lambda$). For $m \geq 2$, Eq. (10.75) has a maximum at $dq_m(t)/dt = 0$, or

$$[r^2(m-1)(rt)^{m-2}e^{-rt}] - [r^2(rt)^{m-1}e^{-rt}] = 0.$$

Hence $t = (m-1)/r$ and for large m, $t \to m/r = 1/R$. Thus we see that the most probable time interval is not the shortest one, but instead approaches the mean time interval between *output* counts $1/R$; that is, the scaling circuit *regularizes* the counts. Equation (10.75) is shown in Fig. 10.11 for

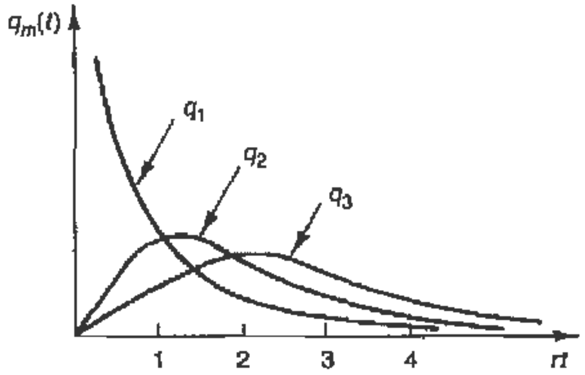


FIGURE 10.11 The probability $q_m(t)$ that the mth count will follow any original count at the time interval t. Note that the abscissa is calibrated in units of rt where r is the unscaled rate of events; for m large the curves approach a Gaussian with mean $\langle rt\rangle = m$ or $\langle t\rangle = m/r$.

different values of m. Comparison of these curves with experimental data has been presented in Section 9.4.2.

10.6. REFERENCES

There are many texts, both elementary and advanced, on the subject of statistics, data fitting, treatment of errors, and computational modeling. The references given below were consulted for the preparation of this chapter.

L. Lyons, *A Practical Guide to Data Analysis for Physical Science Students*, Cambridge Univ. Press, Cambridge, UK, 1994. A succint guide with plenty of examples.

J. R. Taylor, *An Introduction to Error Analysis*, second ed., University Science Books, Sausalito, CA, 1997. A thorough treatment with applications to the physical sciences.

B. P. Roe, *Probability and Statistics in Experimental Physics*, Springer-Verlag, Berlin, 1992. A slightly more advanced and mathematical text.

P. G. Hoel, *Introduction to Mathematical Statistics*, Wiley, New York, 1958. The presentation of Sections 10.1 and 10.2 follows Hoel closely.

A. L. Garcia, *Numerical Methods for Physics*, second ed., Prentice-Hall, Englewood Cliffs, NJ, 2000. A general text including chapters on data analysis and Monte Carlo techniques, with plenty of coding examples in MATLAB, FORTRAN, and C++.

H. Gould and J. Tobochnik, *An Introduction to Computer Simulation Methods: Applications to Physical Systems*, second ed., Addison-Wesley, Reading, MA, 1996. A text devoted to simulations, with extensive use of Monte Carlo methods, with programming examples in BASIC, FORTRAN, C, and PASCAL.

APPENDIX A

Students

We gratefully acknowledge the many students who have contributed the data used to illustrate these experiments.

Students from the University of Rochester:

- R. Armstrong, class of 1994
- D. Boyd, class of 1963
- C. Border, class of 1994
- M. Dobbins, class of 1994
- R. Dockerty, class of 1962
- P. D'Onofrio, class of 1962
- K. Douglass, class of 1964
- E. Glover, class of 1961
- R. Harris, class of 1963
- E. Holroyd, class of 1966
- M. Klein, class of 1962
- D. Kohler, class of 1962
- W. Lama, class of 1966
- T. Londergan, class of 1965
- E. May, class of 1962
- S. McColl, class of 1962
- T. Middleton, class of 1994
- R. Nebel, class of 1962
- P. Nichols, class of 1963
- D. Owen, class of 1963
- D. Peters, class of 1962
- S. Pieper, class of 1965
- W. Rakreungdet, class of 2000
- J. Reed, class of 1961

- A. Rosen, class of 1962
- T. Safford, class of 1994
- D. Sawyer, class of 1963
- P. Schreiber, class of 1962
- D. Stanchfield, class of 1995
- D. Statt, class of 1963
- R. Stevens, class of 1963
- M. Thomas, class of 1994
- J. Traer, class of 1994
- T. Wagner, class of 1961
- T. Walters, class of 1962
- J. S. Weaver, class of 1962
- J. Witkowski, class of 2001
- E. Yadlowski, class of 1962

Students from Rensselaer Polytechnic Institute:

- Daniel Bentz, class of 1996
- Jeff Fedison, class of 1994
- Adam Grossman, class of 2002
- Jackie Krajewski, class of 2005
- Jane Krenkel, class of 2003
- Katie Newhall, class of 2005
- John Orrell, class of 1997
- Ryan Quiller, class of 2003
- Herman Riese, class of 2003
- Kristen Rybij, class of 2003
- Jeffrey Schneider, class of 2003
- Joseph Schreier, class of 2003
- Peter Thies, class of 1996
- Tristan Ursell, class of 2003
- Jeff Wereszczynski, class of 2004
- Jeff Yu, class of 1997

A Short Guide to MATLAB

The experiments described in this book can be analyzed with any of a wide number of computer programs. All that is needed is the ability to sort and plot data, and basic statistical analysis. We have chosen to illustrate the analyses using MATLAB. Although it is a very sophisticated package, a relatively inexpensive student edition that is more than adequate for all of our illustrations is available.

This appendix collects some information that should help you navigate your way through MATLAB. The MATLAB User's Guide is a very useful reference, but there is much more in there than you will need for these experiments. Also remember that you can get help online from http://www.mathworks.com. This site includes a long, searchable list of frequently asked questions, and it is likely that yours is among them. This site also offers you access to programs donated by other users, which you can download and use or modify yourself.

B.1. A MATLAB REVIEW

The following is a brief summary of key MATLAB commands and procedures.

Input Modes. Commands can be executed one by one in the command-line mode in MATLAB or you can write a program consisting of the appropriate command lines in a convenient word processor such as notes in Windows or emacs on a Unix system, and store it as a file with the ".m" extension such as *programname*.m.

Data Input. Lists of data points are usually input as one-dimensional matrices (vectors). You can do this in a command line within MATLAB:

```
x=[1 2 3 4 5 6];
y=[0.1 0.2 0.3 0.4];
```

(The semicolon at the end of the line is *not* necessary, but if you do not include it, then MATLAB will echo values.) You can also store data in ASCII columns in a file with the ".dat" extension, such as *mydata.dat*. If the x data are in the first column and the y data are in the second column of your ASCII file, then you would use the following commands to load it into your MATLAB session:

```
load mydata.dat
x=mydata(:,1);
y=mydata(:,2);
```

Simple Arithmetic. To get an online list of simple functions, type *help elfun*. Formatting for simple calculations with numbers is straightforward: Addition is a+b, subtraction is a-b, multiplication is a*b, division is a/b, and raising to a power is a^b. Scientific functions include:

- abs(x) for absolute value
- round(x) to round to the nearest integer
- real(x) to take the real part of a complex number
- sign(x) to find the sign (it returns +1, −1, or 0)
- log(x) for the natural logarithm
- log10(x) for the logarithm to base 10
- sqrt(x) to find the square root

as well as the familiar trigonometric and hyperbolic functions and their inverses, sin(x), cos(x), tan(x), asin(x), acos(x), atan(x), sinh(x), cosh(x), tanh(x), and so on.

Vector Construction. The easiest way to create a vector with regularly spaced elements is with the command

x = (*start*:*increment*:*last*)

where *start* is the first element of a vector, *last* is the last element, and *increment* is the step size between the elements. For example, x=(0:0.1:1) creates the vector

x = [0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0]

(The parentheses "( )" are optional, or they could be replaced with brackets "[ ]".) This is also equivalent to using the function linspace(*start,last,number*), where *number* is the number of entries in the vector. If you would like to define a vector where the increments are logarithmic, i.e., separated by a constant factor instead of a constant difference, use logspace(*start,last,number*).

Array Arithmetic. To get an online list of matrix functions, type *help elmat*. For operations between a *scalar and an array*, addition, subtraction, multiplication, and division of an array by a scalar look just like simple arithmetic, and the operation applies to every member of the array.

For operations between *two arrays of the same length*, addition, subtraction, multiplication, and division apply on an element-by-element basis, *but* the syntax for multiplication and division is different than that for simple arithmetic. Multiplication is written a.*b and division is a./b, where a and b are vectors of the same length. (Multiplication and division without the dot correspond to normal matrix multiplication and division.)

Data Analysis. There are some simple MATLAB functions for calculating often-used quantities for analyzing a vector x of data values:

- length(x) returns the number of elements in the vector
- sum(x) adds all the elements in the vector
- mean(x) averages all the elements in the vector
- std(x) finds the standard deviation of the elements.

Note that std(x) is equivalent to sqrt(sum((x-mean(x)).^2)/(length(x)-1)).

The command [n,x]=hist(y,nb) takes a vector y of data values, calculates a histogram with nb equally spaced bins, and returns vectors n and x, which give the frequencies and midpoints, respectively, of the binned data.

Least-Squares Fitting. The theory of the least-squares method is discussed, along with reference to MATLAB, in Section 10.3.3.

When the data points are equally weighted, all of the operations necessary to fit a polynomial to a set of (x,y) data points are included in the command p=polyfit(x,y,m), where m is the order of the polynomial. A fit to a straight line is therefore p=polyfit(x,y,1). The vector p holds the best-fit values in order of decreasing polynomial order. For example, if m=2, then you are fitting to a quadratic function $ax^2 + bx + c$ and polyfit returns p=[*a*,*b*,*c*].

The values of the fitted function can be computed for a set of x values x1 using the command y1=polyval(p,x1). (If you want to compute the fitted function at the data points, just use something like yfit=polyval(p,x).)

If the data points are not equally weighted, then you can use Garcia's function linreg to fit to a line. Note that you can retrieve this code from the MATLAB Web site.

Nonlinear Least-Squares Fitting. If you cannot express the function you want to fit as a polynomial, then you cannot use polyfit or linreg. If the function is still linear in the fitting parameters, though, you can use matrix techniques to solve the equations. However, it may be simpler just to resort to numerical techniques to minimize χ^2 directly. You are forced into this situation if the function is nonlinear in the fitting parameters anyway. For example, if you want to fit some decay data to $y = Ae^{-x/\lambda}$, then you can instead fit a straight line to $\log y = \log A - x/\lambda$, but if there is a background term, as in $y = Ae^{-x/\lambda} + B$, then you must use numerical techniques.

Defining the χ^2 function in MATLAB is quite straightforward, and there is a MATLAB function called fminsearch, which will do all the hard work of finding the values of the parameters that minimize the χ^2 function. (See, for example, Section 8.6.2.)

Simple Plots. There are several simple variations on the plot command that will give you everything you need for these experiments. If you really want to do more, see the next section of this appendix.

- plot(y) plots the column values of y versus index. It autoscales the axes. Points are connected by solid lines.
- plot(x,y) plots vector y (vertical) versus vector x (horizontal) on an autoscaled plot. Points are connected by solid lines.
- plot(x,y,'*linetype*') allows you to specify the type of line that connects the points of the type of symbol that is printed on a data point. For "linetype" use "-," ":," "- -," or "-." for solid, dotted,

dashed, or dot-dash lines, respectively, or use ".," "o," "x," "+," or "*" for the corresponding plot symbol.

- bar(y) draws a bar graph of the elements of y versus index.
- bar(x,y) draws a bar graph of y at the locations specified by vector x.
- stairs(y) and stairs(x,y) draw "stairstep" histogram plots.

You can plot more than one set of data, or data and a fit, by specifying more than one set of vectors in plot. For example, plot(x,y,'o',x,yfit,'-') plots "data" vector y versus x as little circles, and then overplots the "fit" vector yfit as a solid line through the points. Another way to overlay plots is to hold a plot and then just repeat the plot command with new vectors. When you are finished collecting overlays, use the command hold off.

Simple labels are put on the graph using the commands

- xlabel('*label on the x-axis*')
- ylabel('*label on the y-axis*')
- title('*title for your plot*')
- text(x,y,'*some text*') puts *some text* at point (x,y)
- legend('*string1*','*string2*',...) labels different sets of data added to the same plot

To print your plot on the default printer, use print. Printing to files or to other printers will depend on which system you are using to run MATLAB. Consult the online help or the User's Manual for details.

B.2. MAKING FANCY PLOTS IN MATLAB

It is simple to make MATLAB plots with the default characteristics. Sometimes, however, that is not quite what you want, especially if you are preparing a formal lab report.

You can also, of course, consult the Mathworks Web page help directly for some hints. For example, if you want to know how to add Greek characters to your plot, click "Tech Support Solution Search" on the Web page, and search for keywords "Greek AND plot." You will find "492 How can I place Greek characters in my plot?" in the search results list. Clicking on this solution tells you not only how to do it, but also tells you how to get an m-file, which will make a chart for you that shows the mappings for all the various Greek letters and symbols.

You can dress up plots quite a bit in MATLAB using what is called "handle graphics." Every plot, and plot element, has a "handle" that you can access in order to change properties of the corresponding element. Most plotting commands return the value of the handle if you ask for it. For example, `h=plot(x,y);` will return a value for the handle `h` that can be used with the command `set` for modifying properties of the plot. Refer to the on-line documentation for more information.

APPENDIX C

Laser Safety

Laser radiation can be dangerous, and in particular it can result in serious and *permanent* damage to the eye. Thus it is important to be aware of the hazards involved and to follow the rules for safe use and operation of lasers. Explicit rules and standards are given in publication ANSI Z136.+-1986 of the American National Standards Institute (1430 Broadway, New York, NY 10018).

The damage a laser can cause depends on the level of the emitted power for CW lasers and on a combination of power and energy for pulsed lasers. The energy per unit area is a better measure of the hazard from direct irradiation. The most serious danger, however, from laboratory lasers in the visible and near infrared (i.e., Nd:YAG) is that they can be focused by the eyeball onto the retina where they will create a *permanent blind spot.* This is particularly serious for infrared lasers where the beam is *invisible.* Thus protective IR absorbing glasses (typically of optical density 4) must always be worn in rooms where IR lasers or beams are present.

Lasers with power below 1 mW are classified as Class 1 lasers. At this power level the exposure in the time it takes for the eye to "blink," approximately 0.25 s, is considered safe. The HeNe laser used in this laboratory is a Class 1 laser. Still one should never stare directly into the beam, or let a specularly reflected ray enter the eye. No eyeglasses are needed but one must use common sense and remain alert. The lasers installed in commercial scanners to which the public is exposed are Class 1 devices. One advantage of the HeNe is that the beam is clearly visible so one is aware of stray beams. Stray beams result from reflection off the various optical elements and other smooth surfaces; they should be blocked or minimized.

Lasers with more than 1-mW power are generally classified as Class 4 devices, as are most pulsed lasers. Nd:YAG and argon-ion lasers can easily deliver several watts of power. Such lasers will cause permanent eye damage instantaneously before one is aware of it. In the case of Class 4 lasers only qualified trained personnel can enter the laser room, which must be kept locked with appropriate signs indicating laser operation. The nitrogen pulsed laser emits in the ultraviolet at $\lambda = 337$ nm. UV is invisible but can be absorbed by plexiglass, so that ordinary safety glasses are not effective; certain materials (i.e., a business card) will fluoresce and can be used to locate the beam. Similarly, IR beams are located with special fluorescent cards and/or with IR viewers.

The need for obeying safety rules and procedures around lasers is a real one, and not a "bureaucratic whim." *Never look into a laser beam, be aware of the stray beams, and wear glasses when required.* Do not let others be exposed to your laser.

APPENDIX D

Radioactivity and Radiation Safety

In a series of experiments on quantum physics, the student comes in contact with radioactive sources, either while studying the properties of the nucleus itself or when using the sources to obtain energetic beams of alpha or beta particles or gamma radiation. As is well known, radiation can be harmful to humans, and therefore precautions must be taken against undue exposure to it, and in the handling of radioactive materials.

In addition to the naturally occurring radioisotopes (which have long lifetimes), a great variety of isotopes have been produced artificially and many of them can be purchased. A convenient table of radioisotopes, many of which, like ^{60}Co, ^{22}Na, and ^{137}Cs, are quite standard for training, testing, and calibration purposes, is available online from the Particle Data Group (PDG) at Lawrence Berkeley National Laboratory:

http://pdg.lbl.gov/2000/sourcesrppbook.pdf

The table gives the type and energy of the radiation, as well as the half-life, with separate information for the different decay schemes, of each radioisotope. Much more detailed information is available from the National Nuclear Data Center (NNDC) at Brookhaven National Laboratory. This information includes level and decay schemes, radiations emitted, and thorough documentation on using the various online programs made available to the user:

8

http://www.nndc.bnl.gov/nndc/nudat/

In the handling of radioactive materials the following regulations should always be observed:

1. Wear a film badge when using radioisotopes.
2. Refrain from eating and smoking while using radioisotopes.
3. Check hands for activity after completing work with radioisotopes (use the appropriate detector, that is, for alphas, betas, etc.).
4. Use gloves when danger of contamination exists.
5. Use tongs for handling strong samples (but only if you can do so safely).
6. In case of a spill, wash it off immediately.
7. *Report all accidents and mishaps connected with radioisotopes.*
8. Do not take radioactive sources out of the laboratory.

Radiation is harmful to living organisms because by ionization it destroys individual cells, and also because it may induce genetic changes. It seems established that low levels of radiation do not produce permanent injury, but the effect is assumed to be cumulative. A genetic change, on the other hand, can be produced by low-level radiation as well as by high-level radiation, but it should not be forgotten that human beings have always been exposed to cosmic rays and natural radioisotopes.

In all establishments where some potential radiation hazard might prevail there must exist an agency (the health physics group) that is responsible for personnel and area monitoring, and for source custody. The health physics groups keeps a record of radioactive sources and other hazards, and of radiation accidents, and in general helps in the enforcement of safe procedures. It should be clear, however, that the sole responsibility for enforcement of proper practices rests with the individual who has been granted the privilege to work with a radioactive source. The aversion of many scientists to observe strict rules is a common phenomenon, but it must not be imitated by the student.

Two peculiar aspects of harm from radiation need special mention and warning: (a) radiation is neither visible nor painful; hence one may not be aware of having been exposed unless proper detectors are used; and (b) in general it is *too late* to do anything after one has been exposed.

Excluding nuclear reactors and particle accelerators, the most serious radiation hazards come from X-ray machines and from taking internally a small amount of radioactive material from a source used in a laboratory.

The PDG publishes online an excellent summary of the units and conversion factors for radiation and radiation doses, as well as recommended exposure limits and radiation protection procedures:

http://pdg.lbl.gov/2000/radiorppbook.pdf

Finally, we conclude with some remarks about radiation shielding. This is important not only for personnel protection, but also to reduce backgrounds in an experiment in which the primary radiation from a source is not meant to be detected.

The purpose of shielding is to attenuate the radiation beam. If the beam consists of charged particles, they do lose energy as they cross matter, and if the shield is sufficiently thick the beam will be completely *stopped.* Since the energy loss is proportional to the number of atomic electrons Z of the shielding material, low-Z materials have a larger stopping power per (nucleon) *gram*. On the other hand, tbe higher the density, the higher the stopping power per unit *length* of shielding.

The attenuation of a gamma-ray beam, however, is different; no gradual energy loss occurs, but there exists a finite probability (cross section) for an interaction. Interactions (electromagnetic) of a gamma-ray beam with matter are either the photoelectric effect, Compton scattering, or pair production, depending on the energy of the beam. As explained in detail in Chapter 8 through a series of such processes a fraction of the beam becomes completely *absorbed* in the material used for shielding. Since the interaction probability is proportional to the amount of material present, we have

$$-\frac{dI}{dx} = I\kappa,$$

hence

$$I = I_0 e^{-\kappa x},$$

where x is the length of the shield, $\kappa = 1/L = \sigma_t \rho N_0$ is the absorption coefficient, and L is the radiation length ($L = 0.51$ cm for lead).

If the beam consists of particles with strong interactions, such as neutrons or protons, the formalism is similar, but now $\kappa = 1/\lambda$, where λ is the mean free path, where λ_p can be roughly taken as 60 g/cm^2.

Despite these considerations, still the best shielding against a radioactive source is distance; since the inverse square law holds, keeping at a 10-m distance dilutes the flux over the value it had at contact with the source (assuming an extent of 5 cm) by a factor of 40,000; for gamma rays such attenuation is equivalent to shielding by 7 cm of lead.

APPENDIX E

Optical Detection Techniques

If we are going to do experiments with light, we have to learn how to measure it. There are several properties of light that can be measured, for example, its intensity, wavelength, or degree of polarization. In this section we discuss ways to measure the intensity, either as energy per unit time or number of photons per unit time.

In order to work with intensity quantitatively, we need to convert it to a voltage level that can be recorded or digitized or so on. However, the simplest option, namely photographic film, still lets you distinguish "dark" from "light" and has some advantages. We discuss it first.

E.1. PHOTOGRAPHIC FILM

Photographic film uses light and chemical reactions to record light intensity. It of course has some obvious drawbacks. For example, it is hard to convert this record into a voltage, although film-scanning machines are built for this purpose. Another disadvantage is that it is inconvenient to record large amounts of data this way, unless some fast and efficient scanning method is available. On the other hand, film has some great advantages as well.

First of all, film is economical. You can record light intensity over quite a large area for very little money. Astronomers, for example, photograph large sections of star fields on a single photographic plate, giving an accurate and reliable record, all for only a few dollars (in film) per picture.

Secondly, film gives you data that you can easily relate to. Distances between images are true, at least to the extent of your focusing device, and you can remeasure or check them easily. There can be an abundance of data on a single photograph, and you can always go back to the same picture if you want to recheck things.

Most importantly, however, film has outstanding position resolution, especially for its price. This resolution is limited by the grain size of the film, and 10 μm is simple to achieve while 1 μm is routine with a little care. What is more, this resolution can be achieved simultaneously over many centimeters of distance. This is almost impossible to achieve with direct electronic means, and can be quite important to astronomers measuring star maps and to optical spectroscopists measuring precise wavelengths.

An important trade-off is between resolution and speed. A film like Kodak Tech-Pan can be used routinely for 1-μm resolution or smaller, but it takes a lot of photons to convert a grain. Thus, such a film is limited to cases of rather large light intensity or where you can afford long exposure times. Somewhat faster films, like Kodak Pan-X, are much faster, and still give resolutions perfectly suitable for most applications.

E.2. PHOTOMULTIPLIER TUBES

The photomultiplier tube (sometime shortened to "phototube" or PMT) is probably the oldest device for converting optical photons directly into electrical signals. It does this with very high efficiency and is very reliable. Some can detect single photons and easily distinguish the signal from background noise. Others are made to measure beams of light. Photomultiplier tubes have been in development for more than 50 years, and have evolved into lots of varieties, some of which are quite sophisticated. The basic operation, though, is simple.

The photomultiplier tube is based on two effects, both of which involve the emission of electrons from the surface of materials. The first is the photoelectric effect, where a photon is absorbed by an electron on the material surface. The electron then emerges with some small kinetic energy; thus a photon is "converted" into an electron. The second effect is that when an electron of some moderate energy strikes a surface, a number of electrons are emitted. (This process is called "secondary emission.") Secondary emission is used to multiply the initial electron into a large number of secondary electrons. All of this takes place on surfaces enclosed within an evacuated glass tube, hence, the name photomultiplier tube.

A schematic photomultiplier tube is shown in Fig. E.1. The photoelectric effect acts at the front surface, or face, of the PMT, and there one photon is converted into one electron (with a certain efficiency less than 1). There is a potential difference of ~100–300 V between the face and the first "stage" of the tube, and this accelerates the electron. When this 100–300 eV electron strikes the first stage, it emits more electrons, which are accelerated to the next stage, and so on. These materials that act as stages are called "dynodes" since they act both as acceptors of electrons (i.e., anodes) and emitters of electrons (i.e., cathodes). After several (usually between 6 and 14) stages, a significant number of electrons emerge in place of the incident photon. Electrical connections are made with the outside world by pins that penetrate the glass envelope on the end.

The front window of the PMT is made of glass or some other transparent material. A thin layer of some optically active material is evaporated on the inner surface of the window. This layer, called the photocathode, is semitransparent and is usually brownish in color. If the tube breaks and air fills the inside, the photocathode oxidizes away and the brownish color disappears. In this case, the photomultiplier tube will never work again.

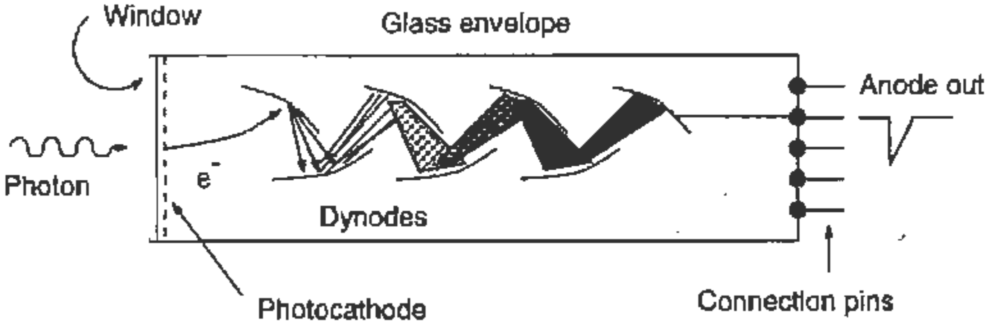


FIGURE E.1 How a photomultiplier tube works. The connection pins are used to supply high voltage to the individual dynodes, and to extract the anode output.

A photon incident on the window penetrates it if it can. In fact, glass window tubes become very inefficient in the near UV because photons with wavelengths below 350 nm are quickly absorbed in ordinary glass. Special UV transmitting glass is available on some photomultiplier tubes, and this can extend the range down to 250 nm or so. To get further into the UV, special windows made of quartz or CaF_2 are necessary, and the devices become very expensive.

If the photon penetrates the window, it reaches the photocathode and has a chance to eject an electron through the photoelectric effect. Recall that in the photoelectric effect, a photon of energy $h\nu$ gives rise to an electron of kinetic energy T given by

$$T = h\nu - \phi,$$

where ϕ is called the "work function" and represents the energy needed to remove the electron from the surface. Several different materials are used for photocathodes, but all are designed to have work functions small enough so that optical photons can eject electrons. It is in fact hard to find materials for which ϕ is less than ≈2 eV, so photomultipliers become quite insensitive at the red end of the visible spectrum.

The probability that an incident photon ejects an electron from the photocathode is called the "quantum efficiency" or QE. It is clearly a function of wavelength λ, tending to zero for both $\lambda \leq$ UV and $\lambda \geq$ red. It is also a function of window and photocathode material for the same reasons. Figure E.2, taken from the Burle photomultiplier tube handbook, shows the "spectral sensitivity" S (in mA/W) for various combinations of windows and photocathodes. Manufacturers tend to quote S rather than QE since it is closer to what the PMTs actually measure. By shining so much light energy per unit time (P) on the face of the PMT, and measuring the current (I) of electrons coming off the photocathode, they determine

$$S \equiv \frac{I}{P} = \frac{N_{\text{electron}} \times e/t}{N_{\text{photon}} \times h\nu/t} = \frac{N_{\text{electron}}}{N_{\text{photon}}} \times \frac{\lambda}{hc/e} = \text{QE} \times \frac{\lambda}{1.24},$$

where S is written in mA/W and λ is in nanometers. Curves of constant QE are drawn in on Fig. E.2. Typical quantum efficiencies are maximum in the blue region and range upward of 25% or so.

Now let's return to Fig. E.1 and see how the photomultiplier tube amplifies the signal. The incident photon has ejected an electron with something

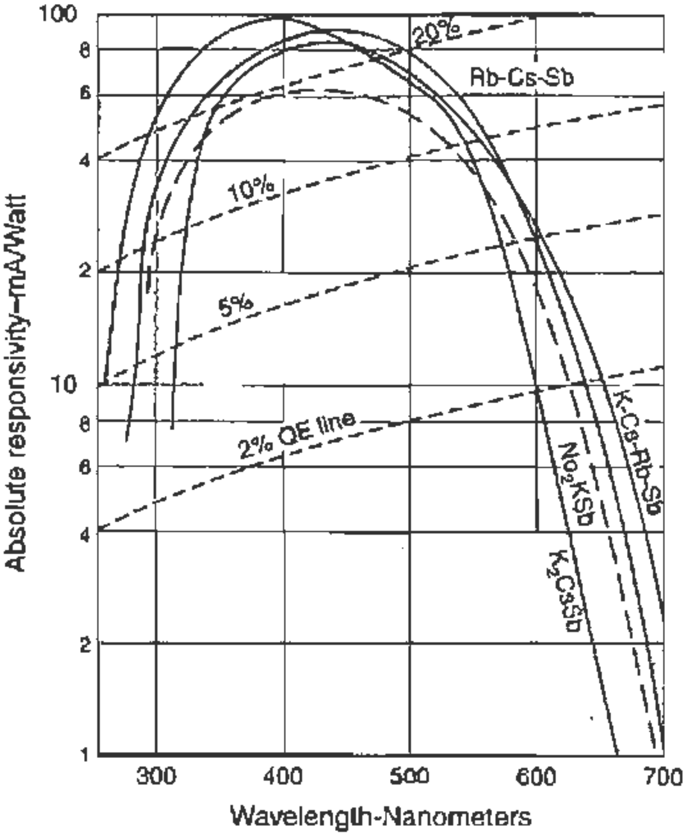


FIGURE E.2 Spectral sensitivity ("absolute responsibility") and quantum efficiency (QE) for some photomultiplier tube windows and photocathodes. From the Burle photomultiplier tube handbook, available online at http://www.burle.com/.

like an electronvolt of kinetic energy. This electron is accelerated to the first dynode and strikes it. The dynodes are constructed out of materials that give a significant mean number of electrons out for each that strikes the surface. This multiplication factor δ is a strong function of the incident electron energy, and is roughly linear with energy up to a few hundred electronvolts or so for most materials used in PMTs.

There is clearly some randomness associated with the operation of a photomultiplier. The quantum efficiency, for example, only represents the probability that a photon will actually eject an electron. The result is that the output voltage pulse corresponding to an input light signal will have random fluctuations about a mean value. We therefore frequently talk in terms of the "mean number of photoelectrons" N_{PE} that correspond to a particular signal.

Assuming that Poisson statistics dominate, this number will dominate the size of the fluctuations, since the number of electrons ejected in subsequent stages will be larger. That is, the fractional rms width of the signal fluctuations should be given by $\sqrt{N_{\text{PE}}}/N_{\text{PE}} = 1\sqrt{N_{\text{PE}}}$. This can be particularly important if the signal corresponds to a very low light level, i.e., a small value of N_{PE}. In this case, there is a probability $e^{-N_{\text{PE}}}$ that there will be no photoelectrons ejected and the signal will go unobserved.

The gain g of a photomultiplier tube is the number of electrons out the back (i.e., at the anode) for a single incident photon. So, for an n-stage tube,

$$g = \delta_1 \times \delta_2 \cdots \times \delta_n \approx \delta^n,$$

where we tacitly assume that δ is the same at each stage; i.e., all dynodes are identical and the potential difference across each stage is the same. If δ is proportional to V, then these assumptions[1] predict that g is proportional to V^n. Thus if you want to keep the gain constant to 1% in a 10-stage photomultiplier tube, you must keep the voltage constant to 0.1%. This is not particularly easy to do.

The accelerating voltage is usually applied to the individual stages by a single external high-voltage DC power supply, and a multilevel voltage divider. The voltage divider has output taps connected to each stage through the pins into the tube. This is connected to the circuit that extracts the signal from the anode. The extraction circuit and voltage divider string are housed together in the photomultiplier tube "base," and their design will vary depending on the application. The base is usually some sort of closed box with a socket that attaches to the tube pins. Two examples of base circuits, taken from the Philips photomultiplier tube handbook, are shown in Fig. E.3. If the signal is more or less continuous, and, for example, a meter reads the current off the anode to ground, you must use the negative high-voltage configuration so that the anode is at (or near) ground. If the output is pulse-like, such as when "flashes" of light, or perhaps individual photons, are detected intermittently, then it is usually best to use the positive high-voltage configuration since that leaves the photocathode at ground. In this case, an RC voltage divider at the anode output allows fast pulses to reach the counter, but the capacitor protects the downstream electronics from the high DC voltage.

[1] These assumptions are almost always wrong. We are using them just to illustrate the general performance of the PMT. For actual gain calculations, you must know the specific characteristics of the PMT.

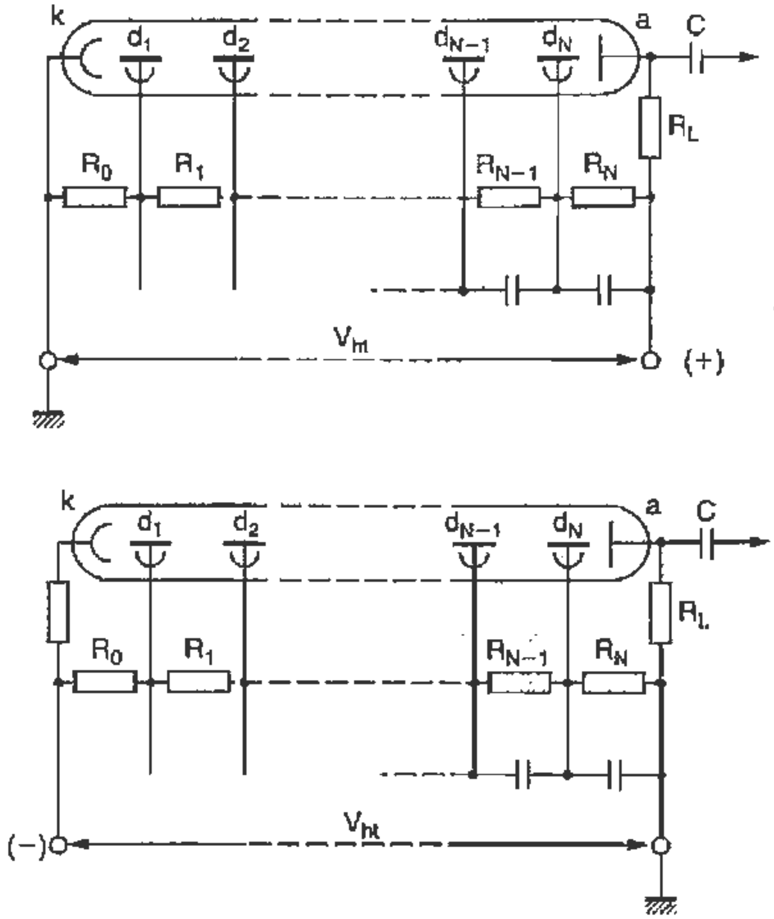


FIGURE E.3 Typical photomultiplier base circuits. The upper figure shows connections for a positive high-voltage configuration, while the lower shows negative high voltage.

No matter what circuit is used, either those in Fig. E.3 or otherwise, you must choose the resistor values carefully. Although the stage voltages only depend on the relative resistor values, you must make sure the average current passing through the divider string is much larger than the signals passing through the PMT. Otherwise, the electrons in the multiplier will draw current through the resistors and change the voltage drop across the stage. Even if this is a small change, it can affect the gain by a lot since the gain depends on voltage to a large power.

On the other hand, you cannot make the resistors arbitrarily small so the divider current gets very large, because this would require a large and expensive high-current, high-voltage DC power supply. What is more, the power dissipated in the divider string, i.e., I^2R, gets to be enormous, making things very hot. Trade-offs must be made, and always keep your eye on the gain.

E.3. PHOTODIODES

Photodiodes are an alternative to photomultipliers. Both turn light directly into electrical signals, but there are distinct differences. First, let's learn how photodiodes work.

Recall our discussion about diodes in Section 3.1.4. A piece of bulk silicon is essentially an insulator. Only thermally excited electrons can move to the upper, empty energy band to conduct electricity, and there are few of them at room temperature. By adding *n*- or *p*-type dopants, lots more charge carriers can be created, and it is a much better conductor. A piece of silicon doped *n* on one end and *p* on the other, a *pn* junction, only conducts in one direction. If a "reverse" voltage is applied, only a tiny current flows, due to the small number of thermally excited electrons.

A photodiode uses light (photons) to excite more electrons than those excited thermally. This is possible if the photon energy is larger than the band gap. Thus, the "reverse" voltage current would increase if you shine light on the diode. This is the principle of the photodiode.

The actual mechanism is a bit more complicated, because of how excited electrons actually conduct. So, for example, for a given applied voltage, the output current is not very linear with intensity. That is, if you double the light intensity, the output current does not change by quite a factor of 2 (over the "noise" from the thermal electrons). Furthermore, a photodiode can work if there is *no* applied voltage, reverse or otherwise. This all means that you must calibrate your photodiode response if you want a quantitative measure of the light intensity.

A popular form of photodiode puts a large region of pure, or "intrinsic," silicon in between the *p* and *n* ends. This increases the active area and decreases the thermal noise current. These photodiodes are called *p–i–n* or "pin" diodes.

Now let's look at a clear advantage that photodiodes have over phototubes. The energy gap in silicon is 1.1 eV, so photons with wavelengths up to ≈1.1 μm can be detected. This is well past red and into the IR. Photomultiplier tubes become inefficient at around 600 nm (see Fig. E.2) or so because of the work function of the photocathode. The band gap of germanium (another popular semiconductor) is 0.72 eV, so germanium photodiodes reach $\lambda \approx 2$ μm. So, if you need to detect red light, you probably want to use a photodiode, and not a photomultiplier tube.

Another big advantage of photodiodes over photomultiplier tubes is cost. A photomultiplier tube with voltage divider circuitry, high-voltage supply, and mechanical assemblies can easily cost upward of $2000. A photodiode costs around $1, and is very easy and cheap to instrument.

Photodiodes can also be made with very small active areas (say 50 μm across). This along with their low cost makes "photodiode arrays" practical. These are lines of photodiodes, separately instrumented, that measure photon position along the array. Such things are frequently used in spectrographic instruments. A typical example might be 1024 25 μm × 2.5 mm photodiodes arranged linearly in a single housing with readout capability, as discussed in Section 5.5. The cost for such a device is typically less than a thousand dollars.

Of course, photomultipliers have some advantages over photodiodes. The biggest is the relative signal-to-noise ratio. A microwatt of incident light power gives around a 1-μA signal in a photodiode, but around 1 A in a photomultiplier tube. This big enhancement in signal is due to the large gain ($\sim 10^6$ or more). Thermally excited electrons are plentiful in a photodiode, but rarely does such an electron spontaneously jump off the photocathode in a photomultiplier. Therefore, the noise is a lot larger in a photodiode. Thus, the signal-to-noise ratio is much worse in a photodiode.

So, if you need to detect very low light intensities ("photon counting" for example), you probably want to use a photomultiplier tube, and not a photodiode.

Photomultipliers also give a more linear response, particularly if care is given to the base design. Some of these relative advantages and disadvantages are shown in Table E.1. Another advantage of photodiodes is that they work in high magnetic fields. Photomultiplier tubes rely on electrons with ≈100–300 eV energy to follow the electric field lines to the dynodes. A few-gauss magnetic field disturbs the trajectories enough to render the PMT useless. In most cases, magnetic shielding solves the problem, but sometimes this is impractical and photodiodes are used instead.

TABLE E.1 Photomultiplier Tubes Versus Photodiodes

If you are	Then your choice should likely be	
interested in...	Photomultiplier	Photodiode
Low cost		√
Red sensitivity		√
Low intensity	√	
Linearity	√	

Finally, we mention that photosensitive transistors, or phototransistors, are also available. They use the natural amplification features of the transistor to get a ~100 times larger signal than the photodiode. Of course, the transistor also amplifies the noise, so there is no improvement in the sensistivity at low intensities.

APPENDIX F

Constants

Table F.1 of fundamental constants is taken from the "Review of particle properties," published in *Phys. Rev. D* **50** (1994). The uncertainties in the values are very small and can be neglected for the experiments in this book.

TABLE F.1 Fundamental Constants

Quantity	Symbol	Value
Speed of light in vacuum	c	299792458 m/s
Planck's constant	h	6.6260755×10^{-34} J s
	$\hbar/2\pi$	6.5821220×10^{-22} MeV s
Electron charge	e	1.60217733×10^{-19} C
	$\hbar c$	1.97327053×10^{-13} MeV m
Vacuum permittivity	ϵ_0	$8.854187817 \times 10^{-12}$ F/m
Vacuum permeability	μ_0	$4\pi \times 10^{-7}$ N/A^2
Electron mass	m_e	0.51099906 MeV/c^2
Proton mass	m_p	938.27231 MeV/c^2
Deuteron mass	m_d	1875.61339 MeV/c^2
Atomic mass unit	u	931.49432 MeV/c^2
Rydberg energy	hcR_∞	13.6056981 eV
Bohr magneton	$\mu_B = e\hbar/2m_e$	5.78838263×10^{-11} MeV/T
Nuclear magneton	$\mu_N = e\hbar/2m_p$	3.15245166×10^{-14} MeV/T
Avogadro constant	N_0	6.0221367×10^{23} atoms/mole
Boltzmann constant	k	1.380658×10^{-23} J/K

6

Exercises

The following exercises may be used.

1. The following table lists data points for the decay rate (in counts/s) of a radioactive source:

Time (s)	*Rate* (s^{-1})	*Time* (s)	*Rate* (s^{-1})	*Time* (s)	*Rate* (s^{-1})
0.6	18.4	2.0	3.02	3.6	1.72
0.8	10.6	2.4	2.61	4.0	1.61
1.2	8.04	2.8	2.08	4.2	1.57
1.6	6.10	3.0	1.50	4.3	1.85

a. Plot the data using an appropriate set of axes, and determine over what range of times the rate obeys the decay law $R = R_0 e^{-t/\tau}$.
b. Estimate the value of R_0 from the plot.
c. Estimate the value of τ from the plot.
d. Estimate the value of the rate you expect at $t = 6$ s.

2. An experiment determines the gravitational acceleration g by measuring the period T of a pendulum. The pendulum has an adjustable length L. These quantities are related as

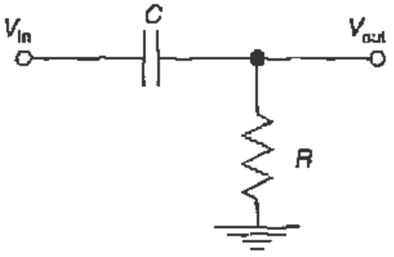
$$T = 2\pi\sqrt{\frac{L}{g}}.$$

A researcher measures the following data points in some arbitrary units.

Data point	*L*	*T*
1	0.6	1.4
2	1.5	1.9
3	2.0	2.6
4	2.6	2.9
5	3.5	3.4

One of these data points is obviously wrong. Which one?

3. Consider the following simple circuit:

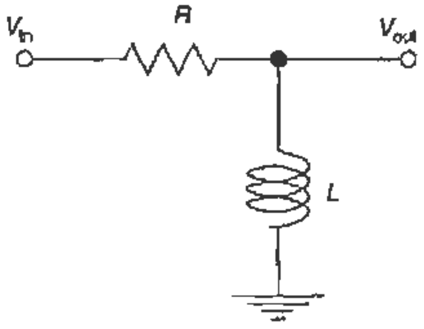


Let the input voltage V_{in} be a sinusoidally varying function with amplitude V_0 and angular frequency ω.

a. Calculate the gain g and phase shift ϕ for the output voltage relative to the input voltage.

b. Plot g and ϕ as a function of ω/ω_0 where $\omega_0 = 1/RC$. For each of these functions, use the combination of linear or logarithmic axes for g and for ϕ that you think are most appropriate.

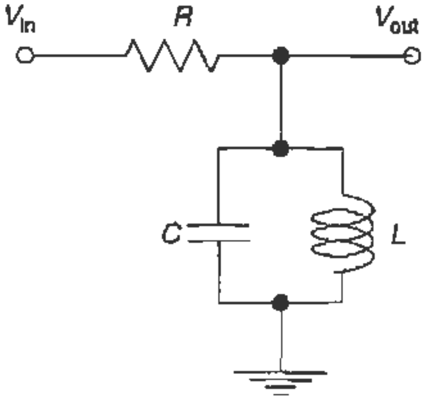
4. Consider the following simple circuit:



Let the input voltage V_{in} be a sinusoidally varying function with amplitude V_0 and angular frequency ω.

a. Calculate the gain g and phase shift ϕ for the output voltage relative to the input voltage.
b. Plot g and ϕ as a function of ω/ω_0 where $\omega_0 = R/L$. For each of these functions, use the combination of linear or logarithmic axes for g and for ϕ that you think are most appropriate.

5. Consider the following not-so-simple circuit:



a. What is the gain g for very low frequencies ω? What is the gain for very high frequencies? Remember that capacitors act like dead shorts and open circuits at high and low frequencies, respectively, and inductors behave in just the opposite way.
b. At what frequency do you suppose the gain of this circuit is maximized?
c. Using the rules for impedance and the generalized voltage divider, determine the gain $g(\omega)$ for this circuit and show that your answers to (a) and (b) are correct.

6. Suppose that you wish to detect a rapidly varying voltage signal. However, the signal is superimposed on a large DC voltage level that would damage your voltmeter if it were in contact with it. You would like to build a simple passive circuit that allows only the high-frequency signal to pass through.

a. Sketch a circuit using only a resistor R and a capacitor C that would do the job for you. Indicate the points at which you measure the input and output voltages.

b. Show that the magnitude of the output voltage equals the magnitude of the input voltage, multiplied by

$$\frac{1}{\sqrt{1 + 1/\omega^2 R^2 C^2}},$$

where ω is the (angular) frequency of the signal.

c. Suppose that $R = 1\ \mathrm{k}\Omega$ and the signal frequency is $1\ \mathrm{MHz} = 10^6/\mathrm{s}$. Suggest a value for the capacitor C.

7. An electromagnet is designed so that a 5-V potential difference drives 100 A through the coils. The magnet is an effective inductor with an inductance L of 10 MHz. Your laboratory is short on space, so you put the DC power supply across the room with the power cables along the wall. You notice that the meter on the power supply must be set to 6 V in order to get 5 V at the magnet. On the other hand, you are nowhere near the limit of the supply, so it is happy to give you the power you need.

Is there any reason for you to be concerned? Where did that volt go, and what are the implications? If there is something to be concerned about, suggest a solution.

8. You are given a low-voltage, high-current power supply to use for an experiment. The manual switch on the power supply is broken. (The power supply is kind of old, and it looks like someone accidently hit the switch with a hammer and broke it off.) You replace the switch with something you found around the lab, and it works the first time, but never again. When you take it apart, the contacts seem to be welded together, and you know it wasn't that way when you put it in. What happened? *(Hint: Recall that the voltage drop across an inductor is $L\,di/dt$, and assume the switch disconnects the circuit over 1 ms or so.)*

9. The following table is from the Tektronix Corp. 1994 catalog selection guide for some of their oscilloscopes:

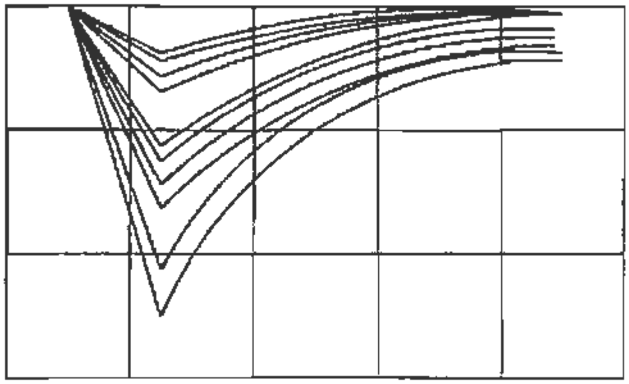
Model	*Bandwidth*	*Sample rate*	*Resolution*	*Time bases*
2232	100 MHz	100 MS/s	8 bits	Dual
2221A	100 MHz	100 MS/s	8 bits	Single
2212	60 MHz	20 MS/s	8 bits	Single
2201	20 MHz	10 MS/s	8 bits	Single

You are looking at the output of a waveform generator on one of these oscilloscopes. The generator is set to give a ±2-V sine wave output. If the sine-wave period is set at 1 μs, the scope indeed shows a 2-V amplitude.

However, if the period is 20 ns, the amplitude is 1 V. Assuming the oscilloscope is not broken, which one are you using?

10. You want to measure the energies of various photons emitted in a nuclear decay. The energies vary from 80 keV to 2.5 MeV, but you want to measure two particular lines that are separated by 1 keV. If you do this by digitizing the output of your energy detector, at least how many bits does your ADC need to have?

11. Pulses emitted randomly by a detector are studied on an oscilloscope: The vertical sensitivity is 100 mV/div and the sweep rate is 20 ns/div. The bandwidth of the scope is 400 MHz. The start of the sweep precedes the trigger point by 10 ns, and the input impedence is 50 Ω.



a. Estimate the pulse risetime. What could you say about the risetime if the bandwidth were 40 MHz?
b. Estimate the trigger level.
c. These pulses are fed into a charge-integrating ADC, also with 50 Ω input impedence. The integration gate into the ADC is 100 ns long and precedes the pulses by 10 ns. Sketch the spectrum shape digitized by the ADC. Label the horizontal axis, assuming $\frac{1}{4}$ pC of integrated charge corresponds to one channel.
d. The ADC can digitize, be read out by the computer, and reset in 100 μs. Estimate the number of counts in the spectrum after 100 s if the average pulse rate is 1 kHz. What is the number of counts if the rate is 1 MHz?

12. A detector system measures the photon emission rate of a weak light source. The photons are emitted randomly. The system measures a rate of 10 kHz, but the associated electronics requires 10 μs to register a photon, and the system will not respond during that time. What is the true rate at which the detector observes photons?

13. You measure the following voltages across some resistor with a three-digit DMM. As far as you know, nothing is changing so all the measurements are supposed to be of the same quantity V_R.

2.31	2.35	2.26	2.22	2.30
2.27	2.29	2.33	2.25	2.29

a. Determine the best value of V_R from the mean of the measurements.
b. What systematic uncertainty would you assign to the measurements?
c. Assuming the fluctuations are random, determine the random uncertainty from the standard deviation.
d. Somebody comes along and tells you that the true value of V_R is 2.23. What can you conclude?

14. (From G. L. Squires, *Practical Physics*, third ed., Cambridge (1985).) In the following examples, q is a given function of the independent measured quantities x and y. Calculate the value of q and its uncertainty δq, assuming the uncertainties are all independent and random, from the given values and uncertainties for x and y.

a. $q = x^2$ for $x = 25 \pm 1$.
b. $q = x - 2y$ for $x = 100 \pm 3$ and $y = 45 \pm 2$.
c. $q = x \ln y$ for $x = 10.00 \pm 0.06$ and $y = 100 \pm 2$.
d. $q = 1 - \frac{1}{x}$ for $x = 50 \pm 2$.

15. Police use radar guns to catch speeders. The guns measure the frequency f of radio waves reflected off of cars moving with speed v. This differs from the emitted frequency f_0 because of the Doppler effect

$$f = f_0 \left(1 - \frac{v}{c}\right)$$

for a car moving away at speed v. What fractional uncertainty must the radar guns achieve to measure a car's speed to 1 mph?

16. The period T of a pendulum is related to its length L by the relation

$$T = 2\pi\sqrt{\frac{L}{g}},$$

where g is the acceleration due to gravity. Suppose you are measuring g from the period and length of a particular pendulum. You have measured the length of the pendulum to be 1.1325±0.0014 m. You independently measure the period to within an uncertainty of 0.06%, that is, $\delta T/T = 6 \times 10^{-4}$. What is the fractional uncertainty (i.e., % uncertainty) in g, assuming that the uncertainties in L and T are independent and random?

17. You have a rod of some metal and you are changing its temperature T. A sensitive gauge measures the deviation of the rod from its nominal length $l = 1.500000$ m. Assuming the rod expands linearly with temperature, you want to determine the coefficient of linear expansion α, i.e., the change in length per Kelvin, and the actual length l_0 before any temperature change is applied. The measurements of the length deviation Δl as a function of the temperature change ΔT are as follows:

ΔT (K)	Δl (μm)	ΔT (K)	Δl (μm)	ΔT (K)	Δl (μm)
0.8	70	2.2	110	3.6	130
1.0	110	2.6	150	3.8	170
1.2	130	2.8	120	4.2	160
1.6	100	3.0	130	4.4	190
1.8	130	3.4	160	5.0	160

Plot the points and draw *three* straight lines through them:

- The line that best seems to go through the points.
- The line with the largest reasonable slope.
- The line with the smallest possible slope.

Use your own estimates by eye to determine these lines. (Do not use a fitting program.) Use the slopes and the intercepts of these lines to determine $\alpha \pm \delta\alpha$ and $l_0 \pm \delta l_0$.

18. For the previous problem, use the method of least squares to fit the data for Δl as a function of ΔT to a straight line. Use the fitted slope and the uncertainty to determine the coefficient of linear expansion α. Also calculate the uncertainty $\delta\alpha$. Are hand estimates just as good as a fitting program? What are the relative advantages or disadvantages?

19. Suppose you wish to measure the gravitational acceleration g by using something like the "Galileo" experiment. That is, you drop an object from some height h and you know that the distance it falls in a time t is given by $\frac{1}{2}gt^2$. For a given experimental run, the fractional uncertainty in h is $\delta h/h = 4\%$ and the fractional uncertainty in t is $\delta t/t = 1.5\%$. Find the fractional uncertainty in g from these data, assuming the uncertainties are random and uncorrelated.

20. You want to measure the value of an inductor L. First, you measure the voltage V across a resistor R when 1.21 ± 0.04 mA flows through it and find $V = 2.53\pm0.08$ V. Then you measure the decay time τ in an RC circuit with this resistor and a capacitor C and get $\tau = RC = 0.463 \pm 0.006$ ms. Finally, you hook the capacitor up to the inductor and measure the oscillator frequency $\omega = 1/\sqrt{LC} = 136 \pm 9$ kHz. What is the value of L and its uncertainty?

21. A simple pendulum is used to measure the gravitational acceleration g. The period T of the pendulum is given by

$$T = 2\pi\sqrt{\frac{L}{g}}\left(1 + \frac{1}{4}\sin^2\frac{\theta_0}{2}\right)$$

for a pendulum initially released from rest at an angle θ_0. (Note that $T \to 2\pi\sqrt{L/g}$ as $\theta_0 \to 0$.) The pendulum length is $L = 87.2 \pm 0.6$ cm. The period is determined by measuring the total time for 100 (round trip) swings.

a. A total time of 192 s is measured, but the clock cannot be read to better than ±100 ms. What is the period and its uncertainty?
b. Neglecting the effect of a finite value of θ_0, determine g and its uncertainty from these data. Assume uncorrelated, random uncertainties.
c. You are told that the pendulum is released from an angle less than 10°. What is the systematic uncertainty in g from this information?
d. Which entity (the timing clock, the length measurement, or the unknown release angle) limits the precision of the measurement?

22. The β-decay asymmetry, A, of the neutron has been measured by Bopp *et al. Phys. Rev. Lett.* **56**, 919 (1986) who found that

$$A = \frac{2\lambda(1-\lambda)}{1+3\lambda^2} = -0.1146 \pm 0.0019.$$

This value is perfectly consistent with, but more precise than, earlier results. The neutron lifetime, τ, has also been measured by several groups, and the results are not entirely consistent with each other. The lifetime is given by

$$\tau = \frac{5163.7\text{ s}}{1+3\lambda^2}$$

and has been measured to be

918 ± 14 s by Christenson *et al.*, *Phys. Rev. D* **5**, 1628 (1972),

881 ± 8 s by Bondarenko *et al.*, *JETP Lett.* **28**, 303 (1978),
937 ± 18 s by Byrne *et al.*, *Phys. Lett. B* **92**, 274 (1980), and
887.6 ± 3.0 s by Mampe *et al.*, *Phys. Rev. Lett.* **63**, 593 (1989).

Which, if any, of the measurements of τ are *consistent* with the result for A? Which, if any, of the measurements of τ are *inconsistent* with the result for A? Explain your answers. A plot may help.

23. The "weighted average" of a set of numbers is

$$\bar{x}_W = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \tag{7.1}$$

where the "weights" $w_i \equiv 1/\sigma_i^2$.

a. Prove that this definition for the weighted average is the value that minimizes χ^2.
b. Use propagation of errors to derive the uncertainty in the weighted average.

24. Let's suppose you have some peculiar dice which each have 10 faces. The faces are numbered from 0 to 9. You throw *eight* of these dice at a time and record which numbers land face down on the table. You repeat this procedure (i.e., throwing the dice) 50 times.

a. For how many throws do you expect there to be exactly three dice landing with either face 1 or face 5 landing face down?
b. What is the average number of dice you expect to land with either face 1 or face 5 down, for any particular throw? What is the standard deviation uncertainty in this number?
c. Use the Poisson approximation to calculate the same number as in (a).
d. Use the Gaussian approximation to calculate the same number as in (a).

25. A radioactive source emits equally in all directions, so that the intensity falls off like $1/r^2$ where r is the distance to the source. You are equipped with a detector that counts only radioactivity from the source, and nothing else. At $r = 1$ m, the detector measures 100 counts in 10 s.

a. What is the count rate, and its uncertainty, in counts per second?
b. What do you expect for the *fractional* uncertainty in the count rate if you count for 100 s instead of 10?

c. Based on the original 10-s measurement, predict the number of counts you should observe, and its uncertainty, if the detector is moved to a distance of 2 m and you count for 1 min.

26. Suppose you are using a Geiger counter to measure the decay rate of a radioactive source. With the source near the detector, you detect 100 counts in 25 s. To measure the background count rate, you take the source very far away and observe 25 counts in 25 s. Random counting uncertainties dominate.

a. What is the count rate (in counts/s) *and its uncertainty* when the source is near the Geiger counter?
b. What is the count rate (in counts/s) *and its uncertainty* when the source is far away?
c. What is the net count rate (in counts/s) *and its uncertainty* due to the source alone?
d. Suppose you want to reduce the uncertainties by a factor of 10. How long must you run the experiment?

27. An experimenter is trying to determine the value of "absolute zero" in degrees Celsius using a pressure bulb and a Celsius thermometer. She assumes that the pressure in the bulb is proportional to the *absolute* temperature. That is, the pressure is zero at absolute zero. She makes five measurements of the temperature at five different pressures:

Pressure (mm of Hg)	65	75	85	95	105
Temperature (°C)	−21	19	41	93	129

Use a straight line fit to determine the value of absolute zero, and its uncertainty, from these data.

28. Fit the following (x, y) values

$x =$	2.5	63	89	132	147
$y =$	406.6	507.2	551.3	625.5	651.7

to a straight line and plot the data points and the fitted line.

a. Does it look like a straight line describes the data well?
b. Study this further by plotting the *deviations* of the fit from the data points. What does this plot suggest?
c. Try fitting the points to a quadratic form, i.e., a polynomial of degree 2. Is this fit significantly better than the straight line?

29. The following results come from a study of the relationship between high school averages and the students' overall average at the end of the first year of college. In each case, the first number of the pair is the high school average, and the second is the college average.

78,65	80,60	85,64	77,59
80,56	82,67	81,66	89,78
87,71	80,66	85,66	87,76
84,73	87,63	74,58	91,78
81,72	91,74	86,66	90,68

a. Draw a scatterplot of the college average against the high school average.

b. Evaluate the correlation coefficient. Would you conclude there is a strong correlation between the grades students get in high school and the grades they get in their first year of college?

30. Using the data in Table 2.1, draw a scatterplot of electrical conductivity versus thermal conductivity for various metals. (Electrical conductivity is the inverse of electrical resistivity.) Calculate the linear correlation coefficient.

31. Graph the ratio of the Poisson distribution to the Gaussian distribution for mean values $\mu = 2$ and for $\mu = 20$. Use this to discuss where the Gaussian approximation to the Poisson distribution is applicable. Repeat the exercise, but compare the Gaussian approximation directly to the binomial distribution with $p = \frac{1}{2}$.

32. Consider blackbody radiation.

a. Show that the wavelength at which the intensity of a blackbody radiator is the greatest is given by "Wien's displacement law":

$$\lambda_{\max}\ (\mathrm{m}) = \frac{2.9 \times 10^{-3}}{T\ (\mathrm{K})}.$$

Hint: You will need to solve an equation like $xe^x/(e^x - 1) = A$ for some value A. If $A \gg 1$ then this is trivial to solve, but you can be more exact using MATLAB. In MATLAB you would use the "function" fzero to find the place where $f(x) = A(e^x - 1) - xe^x$ crosses zero.

b. Stars are essentially blackbody radiators. Our sun is a "yellow" star because its spectrum peaks in the yellow portion of the visible light. Estimate the surface temperature of the sun.

33. A particular transition in atomic neon emits a photon with wavelength $\lambda = 632.8$ nm.

a. Calculate the energy E of this photon.
b. Calculate the frequency ν of this photon.
c. An optical physicist tells you the "line width" of this transition is $\Delta\nu = 2$ GHz. What is the line width ΔE in terms of energy?
d. Use the Heisenberg uncertainty principle to estimate the lifetime Δt of the state that emitted the photon.
e. How far would a photon travel during this lifetime?
f. Suppose the neon is contained in a narrow tube 50 cm long, with mirrors at each end to reflect the light back and forth and "trap" it in the tube. What is the nominal "mode number" for 632.8-nm photons, that is, the number of half-wavelengths that fit in the tube?
g. What is the spacing in frequency between the nominal mode number m, and the wavelength corresponding to the mode $m + 1$?
h. Compare the mode spacing $\delta\nu$ (part G) with the line width $\Delta\nu$.
i. What is this problem describing?

34. Estimate the "transit time" for a typical photomultiplier tube. That is, how much time elapses between the photon ejecting an electron from the photocathode and the pulse emerging from the anode? Assume the photomultiplier has 10 stages and 2000 V between cathode and anode, divided equally among all stages, and that the dynodes are each separated by 1 cm.

35. Some high-quality photomultipliers can detect the signal from a single photoelectron, and cleanly separate it from the background noise. Such a PMT is located some distance away from a pulsed light source, so that on the average, the PMT detects $\langle N_{\mathrm{PE}}\rangle$ photoelectrons. If $\langle N_{\mathrm{PE}}\rangle \ll 1$ and N_0 pulses are delivered, show that the number of pulses detected by the photomultiplier is given by $\langle N_{\mathrm{PE}}\rangle N_0$.

36. A photomultiplier tube observes a flash of green light from an Ar^+ laser. (Assume the photons have wavelength $\lambda = 500$ nm.) The photomultiplier is a 10-stage tube, with a RbCsSb photocathode. The voltages are set so that the first stage has a secondary emission factor $\delta_1 = 5$, while the other 9 stages each have $\delta = 2.5$. The laser delivers some huge number of photons to a diffusing system, which isotropically radiates the light, and only a small fraction of them randomly reach the photomultiplier. On the average, 250 photons impinge on the window for each flash of the laser.

a. What is the average number of electrons delivered at the anode output of the photomultiplier tube, per laser flash?

b. Assume these electrons come out in a rectangular pulse 20 ns wide. What is the height of the *voltage* pulse as measured across a 50-Ω resistor?
c. You make a histogram of these pulse heights. What is the standard deviation of the distribution displayed in the histogram?
d. Suppose the photomultiplier tube is moved four times farther away from the source. For any given pulse of the laser, what is the probability that no photons are detected?

37. A Geiger counter is a device that counts radioactive decays, typically used to find out whether something is radioactive. A particular Geiger counter measures 8.173 background counts per second; i.e., this is the rate when there are no known radioactive sources near it. Your lab partner hands you a piece of material and asks you whether it is radioactive. You place it next to the Geiger counter for 30 s and it registers a total of 253 counts.

a. What do you tell your lab partner?
b. What do you do next?

38. The Tortoise and the Hare have a signal-to-noise problem. A very weak signal sits on top of an enormous background. They are told to determine the signal rate with a fractional uncertainty of 25%, and they decide to solve the problem independently. The Tortoise dives into it and takes data with the setup, and he determines the answer after running the apparatus for a week. The Hare figures she is not only faster than the Tortoise, but smarter too, so she spends two days reducing the background in the apparatus to *zero*, without affecting the signal. She then gets the answer after running the improved setup for one hour. (The Hare really is a lot smarter than the Tortoise, at least this time.)

Assuming Poisson statistics,

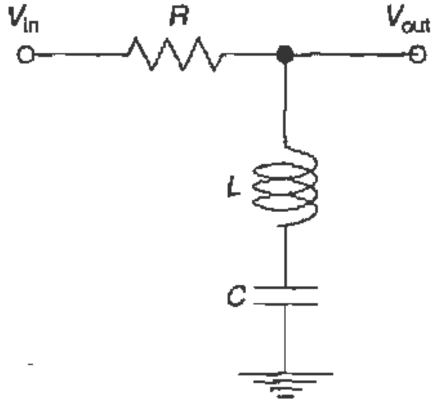
a. What is the signal rate?
b. What is the Tortoise's background rate?

39. Consider the passive filters shown in Fig. 3.11.

a. Determine the gain as a function of $\omega = 2\pi\nu$ for each filter.
b. Plot the gain as a function of ω/ω_C for the three low-pass filters. Define the critical frequency ω_C using the simplest combination of the two components in the circuit, that is, $\omega_C = 1/RC$, $\omega_C = 1/\sqrt{LC}$, or $\omega_C = R/L$. It is probably best to plot all three on the same set of log–log axes.

c. Do the same as (b) for the high-pass filters.
d. Can you identify relative advantages and disadvantages for the different combinations of low-pass and high-pass filters?

40. Consider the following variation on the circuit shown in Fig. 3.12:



a. How does this circuit behave at high frequency?
b. How does this circuit behave at low frequency?
c. Calculate the gain $g = |V_{out}/V_{in}|$ as a function of frequency. What is the behavior for intermediate frequencies?
d. Give an example of where this sort of filter would be useful.

41. A particle detector gives pulses that are 50 mV high when measured as a voltage drop across a 50-Ω resistor. The pulse rises and falls in a time span of 100 ns or less. Unfortunately, there are lots of noisy motors in the laboratory and the ground is not well isolated. The result is that a 10-mV, 60-Hz sine wave is also present across the resistor, and adds linearly with the pulses.

a. Draw a simple circuit, including the 50-Ω resistor and a single capacitor, that allows the pulses to pass, but blocks out the 60-Hz noise.
b. Determine a suitable capacitance value for the capacitor.

42. You are measuring a quantity Q that is proportional to some small voltage. In order to make the measurement, you amplify the voltage using a negative feedback amplifier, as discussed in Section 3.5.

a. Show that the gain g of the full amplifier circuit can be written as

$$g = g_0\left[1 - \frac{1}{\alpha\beta} + \mathcal{O}\left(\frac{1}{\alpha^2\beta^2}\right)\right],$$

where $g_0 = 1/\beta$ and $\alpha \gg 1$ is the internal amplifier gain, β is the feedback fraction, and $\alpha\beta \gg 1$.

b. You measure Q with such an amplifier, with $\beta = 0.01$. The temperature in the lab fluctuates by 5°F while you make the measurement, and the specification sheet for the opamp tells you that its gain varies between 2.2×10^4 and 2.7×10^4 over this temperature range. What is the fractional uncertainty in Q due to this temperature fluctuation?

43. A ^{22}Na radioactive source emits 0.511- and 1.27-MeV γ-rays. You have a detector placed some distance away. You observe a rate of 0.511-MeV photons to be 2.5×10^3/s, and of 1.27-MeV photons to be 10^3/s, with just air between the source and the detector. Calculate the rate you expect for each γ-ray if a 1/2-in.-thick piece of iron is placed between the source and the detector. Repeat the calculation for a 2-in.-thick lead brick.

44. A radioactive source is situated near a particle detector. The detector counts at a rate of 10^4/s, completely dominated by the source. A 2-cm-thick slab of aluminum (density 2.7 gm/cm^3) is then placed between the source and the detector. The radiation from the source must pass through the slab to be detected.

a. Assuming the source emits only 1-MeV photons, estimate the count rate after the slab is inserted.

b. Assuming the source emits only 1-MeV electrons, estimate the count rate after the slab is inserted.

45. Consider a small rectangular surface far away from a source. The surface is normal to the direction to the source, and subtends an angle α horizontally and β vertically. Show that the solid angle subtended is given by $\alpha\beta$.

46. A photomultiplier tube with a 2-in. active diameter photocathode is located 1 m away from a blue light source. The face of the PMT is normal to the direction of light. The light source isotropically emits 10^5 photons/s. Assuming a quantum efficiency of 20%, what is the count rate observed by the photomultiplier?

47. Two scintillation detectors separated by 3 m can measure the "time-of-flight" for a particle crossing both of them to a precision of ±0.20 ns. Each detector can also measure the differential energy loss $dE/dx =$ constant$/\beta^2$, $\beta = v/c$, to ±10%. For a particle with a velocity of 80% the speed of light (i.e., $\beta = 0.8$), how many individual detectors are needed along the particle path to determine the velocity v using dE/dx to the same precision as is possible with time-of-flight?

48. A Čerenkov detector is sensitive to particles that move faster than the speed of light in some medium, i.e., particles with $\beta > 1/n$, where n is the index of refraction of the medium. When a particle crosses such a detector, it produces an average number of detected photons given by

$$\mu = K\left(1 - \frac{1}{\beta^2 n^2}\right).$$

The actual number of detected photons for any particular event obeys a Poisson distribution, so the probability of detecting no photons when the mean is μ is given by $e^{-\mu}$. When 1-GeV electrons ($\beta = 1$) pass through the detector, no photons are observed for 31 out of 19,761 events. When 523-MeV/c pions ($\beta = 0.9662$) pass through, no photons are observed for 646 out of 4944 events. What is the best value of the index of refraction n as determined from these data? What is peculiar about this value? (You might want to look up the indices of refraction of various solids, liquids, and gases.)

Index