OPTICS

Ajoy Ghatak

OPTICS

सूर्यस्य विवधवर्णाः पवनेन
विघट्टिताः कराः साभ्रे ।
वियति धनुः संस्थानाः
ये दृश्यन्ते तदिन्द्रधनुः ।।

Bruhatsamhita-chapter 35
(6th century CE)

The multi coloured rays of the Sun, being dispersed
in a cloudy sky, are seen in the form of a bow,
which is called the Rainbow.

ABOUT THE AUTHOR

Ajoy Ghatak has recently retired as Professor of Physics from Indian Institute of Technology, Delhi. He obtained his M.Sc. from Delhi University and Ph.D. from Cornell University. His area of research is fiber optics. He has several books in this area including *Introduction to Fiber Optics* and *Optical Electronics* (both books coauthored with Prof. K. Thyagarajan and published by Cambridge University Press, United Kingdom). Professor Ghatak is a recipient of several awards including the 2008 SPIE Educator award and the 2003 Optical Society of America Esther Hoffman Beller award in recognition of his outstanding contributions to optical science and engineering education. He is also a recipient of the CSIR S. S. Bhatnagar award, the Khwarizmi International award and the International Commission for Optics Galileo Galilei award. He received DSc (Honoris Causa) from University of Burdwan in 2007.

OPTICS

Ajoy Ghatak

Emeritus Professor
Department of Physics
Indian Institute of Technology
Delhi

Mc Graw Hill Higher Education

Boston Burr Ridge, IL Dubuque, IA New York San Francisco St. Louis
Bangkok Bogotá Caracas Kuala Lumpur Lisbon London Madrid Mexico City
Milan Montreal New Delhi Santiago Seoul Singapore Sydney Taipei Toronto

The McGraw-Hill Companies

Mc Graw Hill Higher Education

OPTICS

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 0 QPD/QPD 0 9

ISBN 978–0–07–338048–3
MHID 0–07–338048–2

Global Publisher: *Raghothaman Srinivasan*
Director of Development: *Kristine Tibbetts*
Developmental Editor: *Lorraine K. Buczek*
Senior Marketing Manager: *Curt Reynolds*
Senior Project Manager: *Jane Mohr*
Senior Production Supervisor: *Laura Fuller*
Associate Design Coordinator: *Brenda A. Rolwes*
Cover Designer: *Studio Montage, St. Louis, Missouri*
Lead Photo Research Coordinator: *Carrie K. Burger*
Compositor: *Aptara®, Inc.*
Typeface: *10/12.5 Times Roman*
Printer: *Quebecor World Dubuque, IA*

Cover: Laser pulses of 80 fs duration having a wavelength 800 nm (and total energy of 1.6 nJ) are incident on a special optical fiber known as a holey fiber, in which a silica core is surrounded by a periodic lattice of air holes; holey fibers are characterized by very small mode field diameters, which lead to very high intensities. Because of the high intensities, SPM (self phase modulation) and other nonlinear effects can be observed; these nonlinear effects result in the generation of new frequencies. In this experiment, the entire visible spectrum gets generated, which can be observed by passing the light coming out of the optical fiber through a prism. The repetition rate of the laser pulses is 82 MHz. The special fibers were fabricated by Dr. Shyamal Bhadra and Dr. Kamal Dasgupta and their group at CGCRI, Kolkata, and the supercontinuum generation was observed by Prof. Ajoy Kar and Dr. Henry Bookey at Heriot Watt University, Edinburgh. Photograph courtesy Prof. Ajoy Kar.

Library of Congress Cataloging-in-Publication Data

Ghatak, A. K. (Ajoy K.), 1939-
Optics / Ajoy Ghatak. — 1st ed.
p. cm.
Includes index.
ISBN 978–0–07–338048–3—ISBN 0–07–338048–2 (hard copy : alk. paper) 1. Optics. I. Title.
QC355.3.G43 2010
535—dc22

2008054008

www.mhhe.com

CONTENTS

Part 6 Photons

Part 7 Lasers and Fiber Optics

PREFACE

The first laser was fabricated in 1960, and since then there has been a renaissance in the field of optics. From optical amplifiers to laser physics, fiber optics to optical communications, optical data processing to holography, optical sensors to DVD technology, ultrashort pulse generation to super continuum generation, optics now finds important applications in almost all branches of science and engineering. In addition to numerous practical applications of optics, it is said that it was the quest to understand the "nature of light" that brought about the two revolutions in science: the development of quantum mechanics started with an attempt to understand the "light quanta," and the starting point of the special theory of relativity was Maxwell's equations which synthesized the laws of electricity and magnetism with those of light. Because of all this, an undergraduate course in optics has become a "must" not only for students of physics but also for students of engineering. Although it is impossible to cover all areas in a single book, this book attempts to give a comprehensive account of a large number of important topics in this exciting field and should meet the requirements of a course on optics meant for undergraduate students of science and engineering.

Organization of the Book

The book attempts to give a balanced account of traditional optics as well as some of the recent developments in this field. The plan of the book is as follows:

- Chapter 1 gives a brief history of the development of optics. I have always felt that one must have a perspective of the evolution of the subject that she or he wants to learn. Optics is such a vast field that it is extremely difficult to give a historical perspective of all the areas. My own interests lie in fiber optics, and hence there is a bias toward the evolution of fiber optics and related areas. In the process, I must have omitted the names of many individuals who made important contributions to the growth of optics. Fortunately, there is now a wealth of information available through the Internet; I have also included a number of references to various books and websites.
- Chapter 2 gives a brief historical evolution of different models describing the nature of light. It starts with the corpuscular model of light and then discusses the evolution of the wave model and the electromagnetic character of light waves. Next we discuss the early twentieth-century experiments, which could only be explained by assuming a particle nature of light, and we end with a discussion on "wave-particle duality."
- Chapters 3 to 6 cover geometrical optics. Chapter 3 starts with Fermat's principle and discusses ray tracing through graded index media, explaining in detail the phenomena of mirage and looming, ray propagation through graded index optical waveguides, and reflection from the ionosphere. Chapter 4 covers ray tracing in lens systems, and Chap. 5 discusses the matrix method in paraxial optics, which is used in the industry. Chapter 6 gives a brief account of aberrations.
- Chapters 7 to 12 discuss the origin of refractive index and the basic physics of wave propagation including Huygens' principle. Many interesting experiments (such as the redness of the setting Sun, water waves, etc.) are discussed. The concept of group velocity and the dispersion of an optical pulse as it propagates through a dispersive medium are discussed in detail. Self phase modulation, which is one of the phenomena leading to the super continuum generation (see photograph on the cover), is also explained.
- Chapters 13 to 16 cover the very important and fascinating area of interference and many beautiful experiments associated with it—the underlying principle is the superposition principle, which is discussed in Chap. 13. Chapter 14 discusses interference by division of the wave front including the famous Young double-hole interference experiment.

In Chap. 15, interference by division of amplitude is discussed which allows us to understand the colors of thin films and applications such as antireflection films. The basic working principle of the fiber Bragg gratings (usually abbreviated as FBG) is discussed along with some of their important applications in the industry. In the same chapter, the Michelson interferometer is discussed which is *perhaps one of the most ingenious and sensational optical instruments ever,* and for which Michelson received the Nobel Prize in Physics in 1907. Chapter 16 discusses the Fabry–Perot interferometer that is based on multiple-beam interference and is characterized by a high resolving power and hence finds applications in high-resolution spectroscopy.

- Chapter 17 discusses the basic concept of temporal and spatial coherence. The ingenious experiment of Michelson, which used the concept of spatial coherence to determine the angular diameter of stars, is discussed in detail. Topics such as optical beats and Fourier transform spectroscopy are also discussed.
- Chapters 18, 19, and 20 cover the very important area of diffraction and discuss the principle behind topics such as the diffraction divergence of laser beams, resolving power of telescopes, laser focusing, X-ray diffraction, optical media technology, Fourier optics, and spatial frequency filtering.
- Chapter 21 discusses the underlying principle of holography and some of its applications. Dennis Gabor received the 1971 Nobel Prize in Physics for discovering the principle of holography.
- Chapters 22 to 24 cover are on the electromagnetic character of light waves. Chapter 22 discusses the polarization phenomenon and propagation of electromagnetic waves in anisotropic media including first-principle derivations of wave and ray velocities. Phenomena such as optical activity and Faraday rotation (and its applications to measuring large currents) are explained from first principles. In Chap. 23, starting with Maxwell's equations, the wave equation is derived which led Maxwell to predict the existence of electromagnetic waves and to propound that light is an electromagnetic wave. Reflection and refraction of electromagnetic waves by a dielectric interface are discussed in Chap. 24. Results derived in this chapter directly explain phenomena such as Brewster's law, total internal reflection, evanescent waves, and Fabry–Perot transmission resonances.
- Chapter 25 covers the particle nature of radiation, for which Einstein received the 1921 Nobel Prize. The chapter also discusses the Compton effect (for which Compton received the 1927 Nobel Prize in Physics), which established that the photon has a momentum equal to $h\nu/c$.
- Chapter 26 is on lasers—a subject of tremendous technological importance. The basic physics of optical amplifiers and of lasers along with their special characteristics is also discussed.
- Chapters 27 to 29 discuss waveguide theory and fiber optics, an area that has revolutionized communications and has found important applications in sensor technology. Chapter 27 discusses the light guidance property of the optical fiber (using ray optics) with applications in fiber-optic communication systems; the chapter also gives a very brief account of fiber-optic sensors. Chapter 28 discusses basic waveguide theory and concept of modes with Maxwell's equations as the starting point. Chapter 29 discusses the propagation characteristics of single-mode optical fibers, which are now extensively used in optical communication systems.
- In 1905 Einstein put forward the special theory of relativity which is considered one of the revolutions of the last century. The starting point of the special theory of relativity was Maxwell's equations, which synthesized the laws of electricity and magnetism with those of light. Chapters 30 and 31 describe briefly the important consequences of the special theory of relativity, i.e., time dilation, length contraction, the mass-energy relation, and Lorentz transformations.
- Very often a good photograph clarifies an important concept and also makes the student interested in the subject. It is with this intention that we have given a few colored photographs (in the insert at the end of the book) that describe important concepts in optics.

In summary, the book discusses some of the important topics that have had a tremendous impact in the growth of science and technology.

Other Important Features of the Book

- A large number of figures correspond to actual numerical calculations which were generated using software such as GNUPLOT and Mathematica. There are also some diagrams which give a three-dimensional perspective of the phenomenon.
- Most chapters start with important milestones in the area. This gives a historical perspective of the topic.

- All important formulae have been derived from first principles so that the book can also be used for self-study.
- Numerous worked out examples are scattered throughout the book to help clarify difficult concepts.
- Each chapter ends with a summary of important results derived in the chapter.

Experiments in Fiber Optics

My own research interests are in the general area of fiber optics. I have found that there are many beautiful experiments in fiber optics, which are not very difficult to set up, that allow us not only to understand difficult concepts but also to find very important applications. For example,

- Optical fibers with parabolic index variation are used in optical communication systems. Ray paths in such fibers and their dispersion characteristics are of great importance. This is discussed from first principles in Chaps. 3 and 27.
- Chapter 10 discusses in great detail the dispersion of an optical pulse as it propagates through a dispersive medium. This is an extremely important concept. The chapter also discusses self phase modulation (usually abbreviated as SPM) that is probably the simplest nonlinear optical phenomenon which can be easily understood from first principles. Indeed, when a monochromatic laser pulse propagates through a special optical fiber, SPM (along with other phenomena) can lead to the awesome super continuum generation; we discuss this in Chap. 10.
- The working of a fiber Bragg grating (usually abbreviated as FBG) is a beautiful application of the interference phenomenon, and FBGs find very important applications in sensors and other optical devices. In Chap. 15, the basic physics of an FBG is discussed along with its very important application in temperature sensing at places where no other device would work.
- The experiment on Faraday rotation in optical fibers (discussed in Chap. 22) allows one to understand the concept of rotation of plane of polarization in the presence of a longitudinal magnetic field. This experiment finds important application in the industry for measuring very large currents (about 10,000 A or more). The theory of Faraday rotation is also given from first principles. In Chap. 22, the change in the state of polarization (usually abbreviated as SOP) of a light beam as it propagates through an elliptic core single-mode optical fiber has been discussed; the experiment not only allows one to understand the changing SOP of a beam propagating through a birefringent fiber, but also helps one to understand the radiation pattern of an oscillating dipole.
- Erbium-doped fiber amplifier (usually abbreviated as EDFA) and fiber lasers are discussed in Chap. 26. The working of an EDFA allows one to easily understand the concept of optical amplification.
- Chapters 27 through 29 are on waveguide theory and fiber optics, an area that has revolutionized communications and finds important applications in sensor technology. Optical fibers are now widely used in endoscopy, display illumination, and sensors, and of course the most important application is in the field of fiber-optic communication systems. We discuss all this in Chap. 27. Chapter 28 discusses basic waveguide theory (and concept of modes) with Maxwell's equations as the starting point. The chapter allows one to understand the transition from geometrical optics to wave optics, which happens to be similar to the transition from classical mechanics to quantum mechanics. Chapter 29 discusses the waveguiding properties of single-mode optical fibers, which are now extensively used in optical communication systems. The prism film coupling experiment (discussed in Chap. 28) allows one to understand the concept of quantization, an extremely important concept in physics and electrical engineering.

There are many such examples scattered throughout the book, and each example is unique and not usually found in other textbooks.

Online Resources for Instructors

Various resources are available to instructors for this text, including solutions to end-of-chapter problems, lecture PowerPoints and the text images in PowerPoint form. All these can be found at the text's website: www.mhhe.com/ghatak

Acknowledgments

At IIT Delhi, I was very fortunate to have the opportunity to interact with outstanding colleagues and with outstanding students, so it was always a pleasure and challenge to teach any course there. We had the opportunity and freedom to modify and develop any course and present it in a form, that would make the subject more interesting. That is how the

present book evolved. In this evolution, many persons have helped me and have made important suggestions. First I would like to mention the name of my very close friend and colleague Prof. Ishwar Goyal, who used earlier Indian editions of this book many times while teaching Optics at IIT Delhi and offered numerous suggestions and many constructive criticisms; I am sure he would have been very happy to see this edition of the book, but unfortunately, he is no longer with us—I greatly miss my interactions with him. I am very grateful to Prof. M. S. Sodha for his constant encouragement and support. My sincere thanks to Prof. K. Thyagarajan for continuous collaboration and for letting me use some of his unpublished notes. My grateful thanks to Prof. Arun Kumar, Prof. Lalit Malhotra, Prof. Bishnu Pal, Prof. Anurag Sharma, Prof. K. Thyagarajan (from IIT Delhi); Dr. Kamal Dasgupta and Dr. Mrinmay Pal (from CGCRI, Kolkata); Dr. Rajeev Jindal, Subrata Dutta, and Giriraj Nyati (from Moser Baer in Noida); Prof. Vengu Lakshminarayanan (from University of Waterloo, Canada) and Prof. Enakshi Sharma (now at University of Delhi South Campus) for their help in writing some portions of the book. I sincerely thank Dr. Gouranga Bose, Dr. Parthasarathi Palai (now at Tejas Networks in Bangalore), Prof. Chandra Sakher, Prof. R. S. Sirohi, Prof. K. Thyagarajan, and Dr. Ravi Varshney (from IIT Delhi); Prof. Govind Swarup (from GMRT, Pune); Dr. Somnath Bandyopadhyay, Dr. Shyamal Bhadra, Dr. Kamal Dasgupta, Dr. Tarun Gangopadhyay, Atasi Pal, and Dr. Mrinmay Pal (from CGCRI, Kolkata); Dr. Suresh Nair (from NeST, Cochin); Avinash Pasricha (from the U.S. Information Service at New Delhi); Prof. Ajoy Kar and Dr. Henry Bookey (from Heriot Watt University, Edinburgh); Dr. R. W. Terhune, Prof. R. A. Phillips, and Dr. A. G. Chynoweth (from the United States) and Dr. R. E. Bailey (from Australia) for providing me important photographs that I have used in this book. I also thank V. V. Bhat for providing me very important literature on the scientific contributions made in ancient India. I would also like to thank my other colleagues, Prof. B. D. Gupta, Dr. Sunil Khijwania, Prof. Ajit Kumar, Dr. Vipul Rastogi, Prof. M. R. Shenoy, and Prof. Kehar Singh for collaboration in research and stimulating discussions. I also thank all the authors and their publishers for allowing me to use many diagrams from their published work. I thank Prof. G. I. Opat of University of Melbourne for his invitation to attend the 1989 conference on teaching of optics which gave me many ideas on how to make difficult concepts in optics easy to understand. I am grateful to Dr. Sunil Khijwania, Monish Das, and Debasish Roy for their help in the preparation of the manuscript and in the drawing of some difficult diagrams. A part of the present writing was carried out with support from Department of Science and Technology, Government of India, which I gratefully acknowledge.

I would also like to thank the following individuals who completed reviews that were instrumental in the development of this first edition:

Alan Cheville
Oklahoma State University

Alfonso D'Alessio
New Jersey Institute of Technology

Dennis Derickson
California Polytechnic State University–San Luis Obispo

Michael Du Vernois
University of Minnesota

Thomas Plant
Oregon State University

Finally, I owe a lot to my family—particularly to my wife, Gopa—for allowing me to spend long hours in preparing this difficult manuscript and for her support all along.

I will be very grateful for suggestions for further improvement of the book. My e-mail addresses are ajoykghatak@yahoo.com and ajoykghatak@gmail.com.

DEDICATION

I dedicate this book to my students; my continuous interactions with them have led to a deeper understanding of optics. I end with the quotation (which I found in a book by G. L. Squires): “ I have learnt much from my teachers, but more from my pupils.” To all my pupils, I owe a very special debt.

Chapter One

HISTORY OF OPTICS

The test of all knowledge is experiment. Experiment is the *sole judge* of scientific "truth". . . . There are *theoretical* physicists who imagine, deduce, and guess at new laws, but do not experiment; and then there are *experimental* physicists who experiment, imagine, deduce and guess.

— Richard Feynman, *Feynman Lectures on Physics*

Optics is the study of light that has always fascinated humans. In his famous book *On The Nature of Light* Vasco Ronchi wrote:

> Today we tend to remember only Newton and Huygens and consider them as the two great men who laid the foundations of physical optics. This is not really true and perhaps this tendency is due to the distance in time which as it increases tends to strengthen the contrast and to reduce the background. In reality, the discussion on the nature of light was fully developed even before these two men were born . . .

It is with this perspective that I thought it would be appropriate to give a very brief history of the development of optics. For those who want to know more of the history, fortunately, there is a wealth of information that is now available through the Internet.

Archytas (428 – 347 BC) was a Greek philosopher, mathematician, astronomer, and statesman. It is said that he had propounded the idea that vision arises as the effect of an invisible "fire" emitted from the eyes so that on encountering objects it may reveal their shapes and colors.

Euclid, also known as **Euclid of Alexandria**, was a Greek mathematician who was born between the years of 320 and 324 BC. In his *Optica* (about 300 BC) he noted that light travels in straight lines and described the law of reflection. He believed that vision involves rays going from the eyes to the object seen, and he studied the relationship between the apparent sizes of objects and the angles that they subtend at the eye. It seems that Euclid's work on optics came to the West mainly through medieval Arabic texts.

Hero (or **Heron**) **of Alexandria** (c. 10 – 70 AD) lived in Alexandria, Roman Egypt, and was a teacher of mathematics, physics, and mechanics at the University of Alexandria. He wrote *Catoptrica*, which described the propagation of light, reflection, and the use of mirrors.

Claudius Ptolemaeus (ca. 90 – ca. 168 AD) known in English as **Ptolemy**, was a mathematician and astronomer who lived in Roman Egypt. Ptolemy's *Optics* is a work that survives only in a poor Arabic translation and in Latin translation of the Arabic. In it, he wrote about properties of light, including reflection, refraction, and color. He also measured the angle of refraction in water for different angles of incidence and made a table of it.

Āryabhatta (AD 476 – 550) is the first of the great mathematician-astronomers of the classical age of Indian mathematics and Indian astronomy. According to the ancient Greeks, the eye was assumed to be a source of light; this was also assumed by the early Indian philosophers. In the fifth century, Aryabhatta reiterated that it was light arriving from an external source at the retina that illuminated the world around us.

Ibn al-Haytham (965–1039), often called as **Alhazen**, was born in Basra, Iraq (Mesopotamia). Alhazen is considered the father of optics because of the tremendous influence of his *Book of Optics* (Arabic: *Kitab al-Manazir*, Latin: *De Aspectibus* or *Perspectiva*). Robert S. Elliot wrote the following about the book:

> Alhazen was one of the ablest students of optics of all times and published a seven-volume treatise on optics which had great celebrity throughout the medieval period and strongly influenced Western thought, notably that of Roger Bacon and Kepler. This treatise discussed concave and convex mirrors in both cylindrical and spherical geometries, anticipated Fermat's law of least time, and considered refraction and the magnifying power of

> lenses. It contained a remarkably lucid description of the optical system of the eye, which study led Alhazen to the belief that light consists of rays which originate in the object seen, and not in the eye, a view contrary to that of Euclid and Ptolemy.

Alhazen had also studied the reverse image formed by a tiny hole and indicated the rectilinear propagation of light. To quote Nobel Prize–winning physicist Abdus Salam:

> Ibn-al-Haitham was one of the greatest physicists of all time. He made experimental contributions of the highest order in optics. He enunciated that a ray of light, in passing through a medium, takes the path which is easier and 'quicker.' In this he was anticipating Fermat's Principle of Least Time by many centuries. . . . Part V of Roger Bacon's "Opus Majus" is practically an annotation to Ibn al Haitham's *Optics*.

There are many books written on the work of Alhazen; some discussion on Alhazen's work can be found in Ref. 1.

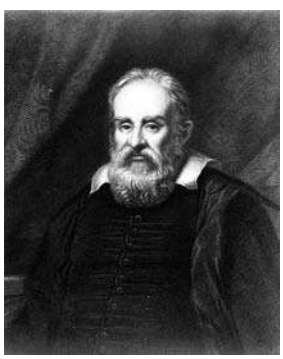
Erazmus Ciolek Witelo (born ca. 1230 and died around 1275) was a theologian, physicist, natural philosopher, and mathematician. Witelo called himself, in Latin, *Turingorum et Polonorum filius*, meaning "a son of Poland and Thuringia." Witelo wrote an exhaustive 10-volume work on optics entitled *Perspectiva*, which was largely based on the work of Ibn al-Haytham and served as the standard text on the subject until the seventeenth century (Refs. 2–4).

Leonardo da Vinci (April 15, 1452 – May 2, 1519), some people believed, was the first person to observe diffraction. Although Alhazen had studied the reverse image formed by a tiny hole, the first detailed description of the pinhole camera (*camera obscura*) was given in the manuscript *Codex atlanticus* (c. 1485) by Leonardo da Vinci, who used it to study perspective.

Johannes Kepler (December 27, 1571 – November 15, 1630) was a German mathematician, astronomer, and astrologer, and a key figure in the seventeenth-century astronomical revolution. In 1604, he published the book *Ad Vitellionem Paralipomena, Quibus Astronomiae pars Optica Traditur*. An English translation (by William H. Donahue) has recently been published as *Johannes Kepler Optics*. The announcement (see Ref. 5) says, "*Optics* was a product of Kepler's most creative period. It began as an attempt to give astronomical optics a solid foundation, but soon transcended this narrow goal to become a complete reconstruction of the theory of light, the physiology of vision, and the mathematics of refraction. The result is a work of extraordinary breadth whose significance transcends most categories into which it might be placed." Reviewing the book, David Lindberg writes:

> In this book Donahue has performed service of enormous value to Kepler scholars and historians of early optics. His lucid translation of the difficult Latin of Kepler's great optical treatise not only affords ready access to Kepler's optical achievement, but also reveals the clarity, rigor, and persuasive power of Kepler's arguments.

Hans Lippershey (1570 – September 1619) was a Dutch eyeglass maker. Many historians believe that in 1608, Lippershey saw two children playing with lenses in his shop and discovered that images were clearer when seen through two lenses. This inspired Lippershey to the creation of the first telescope. Some historians credit Galileo Galilei for the invention of the first telescope. Many historians believe that Lippershey also invented the compound microscope; however, there is controversy on that. See Ref. 6.



Galileo Galilei (February 15, 1564 – January 8, 1642) is often referred to as the father of modern physics. In 1609, Galileo was among the first to use a refracting telescope as an instrument to observe stars and planets. In 1610, he used a telescope as a compound microscope, and he made improved microscopes in 1623 and after. This appears to be the first clearly documented use of the compound microscope.

Willebrord Snel van Royen (1580–1626) was a Dutch astronomer and mathematician. In 1621, he discovered the law of refraction that is referred to as Snell's law.

Pierre de Fermat (August 17, 1601 – January 12, 1665) was a French mathematician and never went to college. In a letter to Cureau de la Chambre (dated January 1, 1662), Fermat showed that the law of refraction can be deduced by assuming that the path of a refracted ray of light was that which takes the least time! Fermat's principle met with objections. In May 1662, Clerselier, an expert in optics, wrote, "The principle you take as a basis for your proof, to wit, that nature always acts by the shortest and simplest path, is only a moral principle, not a physical one—it is not and *can not be* the cause of any effect in nature."

René Descartes (March 31, 1596 – February 11, 1650) was a highly influential French philosopher, mathematician, scientist, and writer. Descartes, in his book entitled *Dioptrique* (1638), gave the fundamental laws of propagation of light, the laws of reflection and refraction. He also put forward the corpuscular model, regarding *lumen* as a swarm of spherical corpuscles (see Refs. 7–8). In Ref. 8, it has been shown that "Descartes' insightful derivation of Snell's law is seen to be largely equivalent to the mechanical-particle or corpuscular derivation often attributed to Newton (who was seven years old at Descartes' death)."

Francesco Maria Grimaldi (April 2, 1618 – December 28, 1663). Around 1660, Grimaldi discovered the diffraction of light and gave it the name *diffraction*, which means "breaking up." He interpreted the phenomenon by stating that *light had to consist of a very fine fluid of some sort in a state of constant vibration*. He laid the groundwork for the later invention of diffraction grating. He formulated a geometrical basis for a wave theory of light in his *Physico-mathesis de lumine* (1666). It was this treatise which attracted Isaac Newton to the study of optics. Newton discussed the diffraction problems of Grimaldi in Part III of his *Opticks* (1704); Robert Hooke observed diffraction in 1672. For more details see Ref. 9.

Robert Hooke, FRS (July 18, 1635 – March 3, 1703). In his 1664 book *Micrographia*, Robert Hooke was the first to describe "Newton's rings." The rings are named after Newton because Newton explained it (incorrectly) in a communication to the Royal Society in December 1675 and presented it in detail in his book *Opticks* (1704). Hooke had also observed the colors from thin sheets of mica much later were explained through interference of light.

Rasmus Bartholin (Latinized *Erasmus Bartholinus*; August 13, 1625 – November 4, 1698) was a Danish scientist. In 1669, he discovered double refraction of a light ray by calcite and wrote a 60-page memoir about the results; the explanation came later. See Ref. 10.

Christiaan Huygens (April 14, 1629 – July 8, 1695) was a Dutch mathematician, astronomer, and physicist. In 1678, in a communication to the Academie des Sciences in Paris, he proposed the wave theory of light and in particular demonstrated how waves might interfere to form a wave front, propagating in a straight line. In 1672, Huygens gave the theory of double refraction which was discovered by Bartholinus in 1669. In 1690, he produced his famous book on optics *Traite de la Lumière*; the English translation of the book is now available as a Dover reprint (Ref. 11), and the entire book can be read at the website given in Ref. 12.

Ole Christensen Rømer (September 25, 1644 – September 19, 1710) was a Danish astronomer who in 1676 made the first quantitative measurements of the speed of light.

Sir Isaac Newton (January 4, 1643 – March 31, 1727) is considered one of the greatest figures in the history of science. In addition to his numerous contributions to science and mathematics, he made a systematic study of light and published it in the form of a book in 1704. The fourth edition of the book is available as a Dover reprint (Ref. 13) and also in the website given in Ref. 14.

In this book, Newton describes his experiments, first reported in 1672, on dispersion, or the separation of light into a spectrum of its component colors. Grimaldi had earlier observed light entering the shadow of a needle—Newton explained this by saying that the needle exerts a force that "pulled" the light from the straight-line path. Hooke had earlier observed the colors from thin sheets of mica—Newton explained this by "fits of easy transmission and reflection" of the light rays.

Thomas Young (June 13, 1773 – May 10, 1829) was an English scientist. In 1801, Young demonstrated the wave nature of light through a simple two-hole interference experiment; this experiment is considered one of 10 most beautiful experiments in physics (Refs. 15 and 16). Thomas Young used his wave theory to explain the colors of thin films (such as soap bubbles); and relating color to wavelength, he calculated the approximate wavelengths of

the seven colors recognized by Newton. In 1817, he proposed that light waves were transverse and thus explained polarization; for more details see Refs. 17–18.

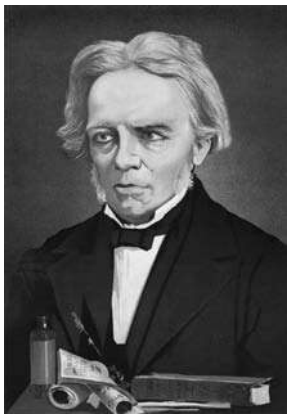
François Jean Dominique Arago (February 26, 1786 – October 2, 1853) was a French mathematician, physicist, astronomer, and politician; he became the twenty-fifth Prime Minister of France. In 1811, Arago observed the rotation of the plane of polarization in quartz. In 1818, Poisson deduced from Fresnel's theory the necessity of a bright spot at the center of the shadow of a circular opaque obstacle. With this result, Poisson had hoped to disprove the wave theory; however, Arago experimentally verified the prediction. Although this spot is usually referred to as the *Poisson* spot, many people call it *Arago's spot*.

Joseph von Fraunhofer (March 6, 1787 – June 7, 1826) was a German optician. In 1814, Fraunhofer invented the spectroscope and discovered 574 dark lines appearing in the solar spectrum; these lines are referred to as *Fraunhofer lines*. In 1859, Kirchhoff and Bunsen explained these lines as atomic absorption lines. In 1823, Fraunhofer published his theory of diffraction. He also invented the diffraction grating and demonstrated the accurate measurement of the wavelength.

Augustin-Jean Fresnel (May 10, 1788 – July 14, 1827) was a French physicist. Fresnel contributed significantly to the establishment of the wave theory of light. In 1818, he wrote a memoir on diffraction for which in the following year he received the prize of the *Académie des Sciences* at Paris. In 1819, he was nominated Commissioner of Lighthouses, for which he was the first to construct a special type of lens, now called a *Fresnel lens*, as substitutes for mirrors. By the year 1821, he showed that polarization could be explained only if light was *entirely* transverse.

Joseph Nicephore Niepce (March 7, 1765 – July 5, 1833) was a French inventor and a pioneer in photography.

Michael Faraday (September 22, 1791 – August 25, 1867) contributed significantly to the fields of electromagnetism and electrochemistry. Faraday had established that a changing magnetic field produces an electric field. This relation subsequently was one of the four equations of Maxwell and is referred to as *Faraday's law*. In 1845, Faraday discovered the phenomenon that is now called the *Faraday rotation*. In this experiment, the plane of polarization of linearly polarized light (propagating through a material medium) gets rotated by the application of an external magnetic field aligned in the direction of propagation. The experiment established that magnetic force and light were related. Faraday wrote in his notebook, "I have at last succeeded in . . . magnetising a ray of light."



Etienne-Louis Malus (July 23, 1775 – February 24, 1812) was a French engineer, physicist, and mathematician. **David Brewster**, FRS (December 11, 1781 – February 10, 1868) was a Scottish scientist.

In 1809, Malus had published his discovery of the polarization of light by reflection; however, he was unable to obtain the relationship between the polarizing angle and refractive index. In 1811, David Brewster *repeated* the experiments of Malus for many materials and realized that when a ray is polarized by reflection, the reflected ray makes an angle of 90° with the refracted ray; he promptly called this *Brewster's law!* Malus is best known for the law named after him which states that the intensity of light transmitted through two polarizers is proportional to the square of the cosine of the angle between the polarization axes of the polarizers. In 1810, Malus published his theory of double refraction of light in crystals.

James Clerk Maxwell (June 13, 1831 – November 5, 1879) was an outstanding Scottish mathematician and theoretical physicist. Around 1865, Maxwell showed that the laws of electricity of magnetism can be described by four partial differential equations; these equations are known as *Maxwell's equations* and appeared in his book *A Treatise on Electricity and Magnetism*, published in 1873. Maxwell also predicted the existence of electromagnetic waves (which were later observed by Hertz) and showed that the speed of propagation of electromagnetic waves is approximately equal to the (then) measured value of the speed of light; this made him

predict that light must be an electromagnetic wave. In 1864, he wrote:

> This velocity is so nearly that of light that it seems we have strong reason to conclude that light itself (including radiant heat and other radiations) is an electromagnetic disturbance in the form of waves propagated through the electromagnetic field according to electromagnetic laws.

This synthesis represents one of the great scientific achievements of the nineteenth century. In 1931 (during the birth centenary celebration of Maxwell), Max Planck had said, "(Maxwell's theory) . . . remains for all time one of the greatest triumphs of human intellectual endeavor." Albert Einstein had said, "(The work of Maxwell was) . . . the most profound and the most fruitful that physics has experienced since the time of Newton." For more details about Maxwell, see Ref. 19. Some of the original papers of Maxwell can be seen in the website in Ref. 20.

John William Strutt usually referred to as Lord Rayleigh (November 12, 1842 – June 30, 1919) and **John Tyndall** (August 2, 1820 – December 4, 1893) was an Irish natural philosopher. In 1869, John Tyndall had discovered that when light passes through a transparent liquid with small particles in suspension (such as a small amount of milk put in water), the shorter blue wavelengths are scattered more strongly than the red; thus from the side, the color looks blue and the light coming out straight appears reddish. Many people call this *Tyndall scattering*, but it is more often referred to as *Rayleigh scattering* because **Rayleigh** studied this phenomenon in great detail and showed (in 1871) that scattering is inversely proportional to the fourth power of the wavelength (see Ref. 21). Thus the blue color is scattered 10 times more than the red color (because the red color has a wavelength which is about 1.75 times the wavelength of blue). This is the reason why the sky appears blue. Although violet has an even smaller wavelength, the sky does not appear violet because there is very little violet in the sunlight! Some of the scientific papers of Lord Rayleigh can be seen at the website given in Ref. 22. Lord Rayleigh received the 1904 Nobel Prize in Physics.

In 1854, John Tyndall demonstrated light guidance in water jets, duplicating but not acknowledging Babinet (see Ref. 23 for more details).

Heinrich Rudolf Hertz (February 22, 1857 – January 1, 1894) was a German physicist after whom the hertz, the SI unit of frequency, is named. To quote from Ref. 24:

> In 1888, in a corner of his physics classroom at the Karlsruhe Polytechnic in Berlin, Hertz generated electric waves using an electric circuit; the circuit contained a metal rod that had a small gap at its midpoint, and when sparks crossed this gap violent oscillations of high frequency were set up in the rod. Hertz proved that these waves were transmitted through air by detecting them with another similar circuit some distance away. He also showed that like light waves they were reflected and refracted and, most important, that they traveled at the same speed as light but had a much longer wavelength. These waves, originally called Hertzian waves but now known as radio waves, conclusively confirmed Maxwell's prediction on the existence of electromagnetic waves, both in the form of light and radio waves.

Hertz was a very modest person; after the discovery he said, "This is just an experiment that proves Maestro Maxwell was right, we just have these mysterious electromagnetic waves that we cannot see with the naked eye. But they are there." "So, what next?" asked one of his students at the University of Bonn. "Nothing, I guess." Hertz later said, "I do not think that the wireless waves I have discovered will have any practical application."

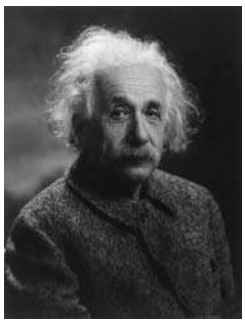
We should mention here that in 1842 (when Maxwell was only 11 years old) the U.S. physicist Joseph Henry had magnetized needles at a distance of over 30 ft (with two floors, each 14 in. thick) from a single spark. Thus, though Joseph Henry was not aware of it, he had produced and detected electromagnetic waves; for more details see e.g., the book by David Park (Ref. 25) and the original collection of Henry's papers referenced in Park's book.

Hertz was also the first scientist to observe the photoelectric effect. In 1887, while receiving the electromagnetic waves in a coil with a spark gap, he found that the maximum spark length was reduced when the apparatus was put in a black box (this is so because the box absorbed the ultraviolet radiation which helped the electrons to jump across the gap). Hertz reported the observations but did not pursue further and also did not make any attempt to explain them. In 1897, J. J. Thomson discovered electrons, and in 1899, he showed that electrons are emitted when light falls on a metal surface. In 1902, Philip Lenard observed that (1) the kinetic energy of the emitted electrons was independent of the intensity of the incident light and (2) the energy of the emitted electron increased when the frequency of the incident light was increased.

Alexander Graham Bell (March 3, 1847 – August 2, 1922) was born and raised in Edinburgh, Scotland he emigrated to Canada in 1870 and then to the United States in 1871. The photophone was invented jointly by Alexander Graham Bell and his assistant Charles Sumner Tainter on February 19, 1880. Bell believed the photophone was his most important invention.

Albert Abraham Michelson (December 19, 1852 – May 9, 1931) was born in Strelno, Prussia, and moved to the United States at the age of 2. Michelson built the famous interferometer which was later called the *Michelson interferometer*. He was awarded the 1907 Nobel Prize in Physics (the first American to receive the Nobel Prize in Science) for his optical precision instruments and the spectroscopic and metrological investigations carried out with their aid. In the presentation speech, the President of the Royal Swedish Academy of Sciences said, ". . . Your interferometer has rendered it possible to obtain a non-material standard of length, possessed of a degree of accuracy never hitherto attained. By its means we are enabled to ensure that the prototype of the meter has remained unaltered in length, and to restore it with absolute infallibility, supposing it were to get lost." In 1887, he and Edward Morley carried out the famous Michelson–Morley experiment, which proved that ether did not exist. David Park (Ref. 25) has written: "He (Michelson) was 34 when he established that ether cannot be found; he made delicate optical measurements for 44 more years and to the end of his days did not believe there could be a wave without some material substance to do the waving."

Maurice Paul Auguste Charles Fabry (June 11, 1867 – July 9, 1945) and **Jean-Baptiste Alfred Pérot** (November 3, 1863 – November 28, 1925) were French physicists. In 1897, Fabry and Pérot published their important article on what we now call the *Fabry–Pérot interferometer*. For more details about them see Ref. 26.


Library of Congress

Albert Einstein (March 14, 1879 –April 18, 1955) was an outstanding theoretical physicist. Einstein is best known for his theory of relativity and specifically mass-energy equivalence $E = mc^2$. Einstein in 1905 put forward that light consists of quanta of energy; this eventually led to the development of quantum theory. In 1917, in a paper entitled "On the Quantum Theory of Radiation," Einstein, while rederiving Planck's law, was able to predict the process of stimulated emission, and almost 40 years later, this prediction led to the development of the laser. He received the 1921 Nobel Prize in Physics for his services to Theoretical Physics, and especially for his explanation of the photoelectric effect. Some of Einstein's early papers can be found in the website in Ref. 27.

Geoffrey Ingram Taylor (March 7, 1886 – June 27, 1975) in 1909 demonstrated interference fringes by using an extremely feeble light source; this led the Nobel Prize–winning physicist P. A. M. Dirac to make the famous statement, "Each photon then interferes only with itself." Taylor has often been described as one of the great physical scientists of the twentieth century. For more details, see Ref. 28.

William Henry Bragg (July 2, 1862 – March 10, 1942) and **William Lawrence Bragg** (March 31, 1890 – July 1, 1971). **William Lawrence Bragg** (the son) discovered the most famous *Bragg's law*, which makes it possible to calculate the positions of the atoms within a crystal from the way in which an X-ray beam is diffracted by the crystal lattice. He made this discovery in 1912, during his first year as a research student in Cambridge. He discussed his ideas with his father (William Henry Bragg), who developed the X-ray spectrometer in Leeds. In 1915, father and son were jointly awarded the Nobel Prize in Physics for their services in the analysis of crystal structure by means of X-rays. The collaboration between father and son led many people to believe that the father was the inventor of *Bragg's law*, a fact that upset the son!

Arthur Holly Compton (September 10, 1892 – March 15, 1962) in 1922 found that the energy of an X-ray or gamma ray photon decreases due to scattering by free electrons. This discovery, known as the *Compton effect*, demonstrates the corpuscular nature of light. Compton received the 1927 Nobel Prize in Physics for his discovery of the effect named after him. The research papers of Compton can be found in the website given in Ref. 29.

Louis de Broglie (August 15, 1892 – March 19, 1987) was a French physicist. In 1924, de Broglie (pronounced in French as *de Broy*) formulated the *de Broglie hypothesis*, claiming that *all* matter, not just light, has a wavelike nature; he related wavelength to the momentum. De Broglie's formula was confirmed three years later for electrons with the observation of electron diffraction in two independent experiments. De Broglie received the 1929 Nobel Prize in Physics for his discovery of the wave nature of electrons. In the presentation speech it was mentioned:

> Louis de Broglie had the boldness to maintain that . . . matter is, by its nature, a wave motion. At a time when no single known fact supported this theory, Louis de Broglie asserted that a stream of electrons which passed through a very small hole in an opaque screen must exhibit the same phenomena as a light ray under the same conditions.

Paul Adrien Maurice Dirac (August 8, 1902 – October 20, 1984) and **Werner Karl Heisenberg** (December 5, 1901 –

February 1, 1976) were both celebrated theoretical physicists. Heisenberg was one of the founders of quantum mechanics and is also well known for discovering one of the central principles of modern physics, the Heisenberg uncertainty principle, which he developed in an essay published in 1927. The uncertainty principle (which can be derived directly from the axioms of quantum mechanics) can be used to explain the diffraction of a photon (or an electron). Dirac can be considered as the creator of the complete theoretical formulation of quantum mechanics. Albert Einstien said that it was "Dirac to whom in my opinion we owe the most logically perfect presentation of quantum mechanics." Dirac, in his famous book *Principles of Quantum Mechanics*, wrote:

> Some time before the discovery of quantum mechanics people realized that the connection between light waves and photons must be of a statistical character. What they did not clearly realize, however, was that the wave function gives information about the probability of one photon being in a particular place and not the probable number of photons in that place. The importance of the distinction can be made clear in the following way. Suppose we have a beam of light consisting of a large number of photons split up into two components of equal intensity. On the assumption that the beam is connected with the probable number of photons in it, we should have half the total number going into each component. If the two components are now made to interfere, we should require a photon in one component to be able to interfere with one in the other. Sometimes these two photons would have to annihilate one another and other times they would have to produce four photons. This would contradict the conservation of energy. The new theory, which connects the wave function with probabilities for one photon gets over the difficulty by making each photon go partly into each of the two components. Each photon then interferes only with itself. Interference between two different photons never occurs.

Dirac is widely regarded as one of the greatest physicists of all time.

Chandrasekhara Venkata Raman (November 7, 1888 – November 21, 1970) and **Kariamanikkam Srinivasa Krishnan** (December 4, 1898 – June 13, 1961) on February 28, 1928, observed the *Raman effect* in several organic vapors such as pentane, which they called "the new scattered radiation." Raman made a newspaper announcement on February 29, and on March 8, 1928 he communicated a paper entitled "A Change of Wavelength in Light Scattering to Nature," which was published on April 21, 1928. Although in the paper he acknowledged that the observations were made by K. S. Krishnan and himself, the paper had Raman as *the* author, and therefore the phenomenon came to be known as the Raman effect although many scientists (particularly in India) kept on referring to it as the Raman–Krishnan effect. Subsequently, there were several papers written by Raman and Krishnan. Raman got the 1930 Nobel Prize in Physics for "his work on the scattering of light and for the discovery of the effect named after him." At about the same time, Landsberg and Mandel'shtam (in Russia) were also working on light scattering, and according to Mandel'shtam, they observed the Raman lines on February 21, 1928. But the results were presented in April 1928, and it was only on May 6, 1928, that Landsberg and Mandel'shtam communicated their results to the journal *Naturwissenschaften*. But by then it was too late! Much later, scientists from Russia kept calling Raman scattering as Mandel'shtam–Raman scattering. For a nice historical account of Raman effect, see Ref. 30. In 1928, the Raman effect was discovered; 70 years later it has become an important mechanism for signal amplification in optical communication systems. Today we routinely talk about Raman amplification in optical fibers.

Dennis Gabor (June 5, 1900, Budapest – February 9, 1979, London). In 1947, while working in the area of electron optics at British Thomson-Houston Co. in the United Kingdom, Dennis Gabor invented holography. He was awarded the 1971 Nobel Prize in Physics for his invention and development of the holographic method. However, the field of holography advanced only after development of the laser in 1960. The first holograms that recorded 3D objects were made by Emmett Leith and Juris Upatnieks in Michigan, United States, in 1963 and Yuri Denisyuk in the Soviet Union.

Charles Hard Townes July 28, 1915) and (his sister's husband) **Arthur Leonard Schawlow** (May 5, 1921 – April 28, 1999) are both U.S. physicists. **Nikolay Gennadiyevich Basov** (December 14, 1922 – July 1, 2001) was a Russian physicist and educator. **Aleksandr Mikhailovich Prokhorov** (July 11, 1916–January 8, 2002) was a Russian physicist born in Australia. **Gordon Gould** (July 17, 1920 – September 16, 2005), was a U.S. physicist. The most important concept in the development of the laser is that of stimulated emission, which was introduced by Einstein in 1917. It took over 35 years to realize amplification through stimulated emission primarily because stimulated emission was long regarded as

a purely theoretical concept which never could be observed, because under normal conditions absorption would always dominate over emission. According to Townes, he conceived the idea of amplification through population inversion in 1951 (see Ref. 31); and in early 1954, Townes, Gordon, and Zeiger (at the Physics Department of Columbia University) published a paper on the amplification and generation of electromagnetic waves by stimulated emission. They coined the word *maser* for this device, which is an acronym for *m*icrowave *a*mplification by *s*timulated *e*mission of *r*adiation. Around the same time, Basov and Prochorov at the Lebedev Institute in Moscow independently published papers about the maser. In 1958, Schawlow and Townes published a paper entitled "Infrared and Optical Masers" in *Physical Review* showing how stimulated emission would work with much shorter wavelengths and describing the basic principles of the optical maser (later to be renamed a *laser*), initiating this new scientific field. Townes, Basov, and Prokhorov shared the 1964 Nobel Prize in Physics for their fundamental work in the field of quantum electronics, which has led to the construction of oscillators and amplifiers based on the maser-laser principle. Half of the prize was awarded to Townes and the other half jointly to Basov and Prokhorov. Schawlow got the Nobel Prize much later; he shared the 1981 Nobel Prize in Physics with Nicolaas Bloembergen and Kai Siegbahn for their contributions to the development of *laser spectroscopy*. However, many people believe that Gordon Gould (while a graduate student at Columbia University) is the inventor of the laser. On the first page of Gould's laser notebook (written in November 1957) he coined the acronym *LASER* and described the essential elements for constructing the laser. In fact, the term *laser* was first introduced to the public in Gould's 1959 conference paper "The LASER, Light Amplification by Stimulated Emission of Radiation." Gould for 30 years fought the United States Patent and Trademark Office for recognition as the inventor of the laser; see, e.g., the books by Taylor (Ref. 32) and by Bertolotti (Ref. 33). A portion of Bertolotti's book can be read at the website given in Ref. 34.

Theodore Harold Maiman (July 11, 1927 – May 5, 2007). In *A Century of Nature: Twenty-One Discoveries that Changed Science and the World*, (Ref. 35), C. H. Townes wrote an article entitled "The First Laser" (Ref. 36). In this article, Townes wrote:

> Theodore Maiman made the first laser operate on 16 May 1960 at the Hughes Research Laboratory in California, by shining a high-power flash lamp on a ruby rod with silver-coated surfaces. He promptly submitted a short report of the work to the journal *Physical Review Letters*, but the editors turned it down. Some have thought this was because *Physical Review* had announced that it was receiving too many papers on masers—the longer-wavelength predecessors of the laser—and had announced that any further papers would be turned down. But Simon Pasternack, who was an editor of *Physical Review Letters* at the time, has said that he turned down this historic paper because Maiman had just published, in June 1960, an article on the excitation of ruby with light, with an examination of the relaxation times between quantum states, and that the new work seemed to be simply more of the same. Pasternack's reaction perhaps reflects the limited understanding at the time of the nature of lasers and their significance. Eager to get his work quickly into publication, Maiman then turned to *Nature*, usually even more selective than *Physical Review Letters*, where the paper was better received and published on 6 August.

On December 12, 1960, **Ali Javan**, **William Bennett**, and **Donald Herriott** produced, for the first time, a continuous laser light (at 1.15 μm) from a gas laser. (Refs. 37 and 38).

In 1961, within one year of the development of the first laser, **Elias Snitzer** and his coworkers developed the first fiber-optic laser. Snitzer also invented both neodymium- and erbium-doped laser glass; see, e.g., Refs. 39 and 40.

C. Kumar N. Patel developed the carbon dioxide laser in 1963; it is now widely used in industry for cutting and welding and also in surgery. See Ref. 41.

In 1966, in a landmark theoretical paper (published in *Proceedings of IEE*), **Charles Kuen Kao** and **George Hockham** of Standard Telecommunications Laboratories in the United Kingdom pointed out that the loss in glass fibers was primarily caused by impurities and therefore it was not a fundamental property of the fiber itself. They said that if the impurities could be removed, the loss could be brought down to about few decibels per kilometer—or may be even less. If this could be achieved, then (to quote from Ref. 42) "the new form of communication medium . . . compared with existing co-axial cable and radio systems, has a larger information capacity and possible advantages in basic material cost." After this paper, scientists in the United States, United Kingdom, France, Japan, and Germany started working on purifying glass, and the first breakthrough was reported in 1970.

In 1970, Corning Glass Works scientists **Donald Keck**, **Robert Maurer**, and **Peter Schultz** successfully prepared the first batch of optical fiber with sufficiently low loss as to make fiber-optic communication a reality—this breakthrough was

the starting point of the fiber-optic revolution; see Ref. 43. In about 10 more years, as research continued, optical fibers became so transparent that more than 95% of the signal power would pass after propagating through 1 km of the optical fiber.

Semiconductor lasers that operate continuously at room temperature were first fabricated in May 1970 by Zhorev Alferov and his group in Leningrad, and in June 1970, by Izuo Hayashi and Morton Panish at Bell Labs (Ref. 44). This was a major turning point toward the development of the fiber-optic communication system. Alferov shared the 2000 Nobel Prize in Physics.

In 1978, the photosensitivity of germanium-doped-core optical fibers was discovered by **Kenneth Hill** while working at the Communications Research Centre in Ottawa, Canada. He also demonstrated the first in-fiber Bragg grating (see Ref. 45).

The erbium-doped fiber amplifier (usually abbreviated as EDFA) was invented in 1987 by a group including **David Payne**, **R. Mears**, and **L. Reekie**, from the University of Southampton, and a group from AT&T Bell Laboratories, including **E. Desurvire**, **P. Becker**, and **J. Simpson**. The EDFA brought about a revolution in fiber-optic communication systems.

REFERENCES

1. http://www.ibnalhaytham.net/custom.em?pid=673913
2. http://micro.magnet.fsu.edu/optics/timeline/people/witelo.html
3. http://www.polybiblio.com/watbooks/2915.html
4. http://members.aol.com/WSRNet/D1/hist.htm
5. http://www.greenlion.com/cgi-bin/SoftCart.100.exe/optics.html?E+scstore
6. http://www.ece.umd.edu/~taylor/optics3.htm
7. V. Ronchi, *The Nature of Light* (translated by V. Barocas), Heinemann, London, 1972.
8. W. B. Joyce and Alice Joyce, "Descartes, Newton and Snell's Law," *J. Opt. Soc. Am.*, Vol. 66, p. 1, 1976.
9. http://www.faculty.fairfield.edu/jmac/sj/scientists/grimaldi .htm
10. http://www.polarization.com/history/bart.html
11. C. Huygens, *Treatise on Light*, Dover Publications, 1962.
12. http://www.gutenberg.org/files/14725/14725-h/14725-h.htm
13. Isaac Newton, *Opticks*, Dover Publications, 1952.
14. http://books.google.com/books?id=GnAFAAAAQAAJ&pg=PA381&dq=Newton%27s+book+OPTICKS#PPA219,M1
15. http://physics.nad.ru/Physics/English/top_ref.htm
16. *Great Experiments in Physics*, pp. 96–101, Holt, Reinhart and Winston, New York, 1959.
17. http://www.cavendishscience.org/phys/tyoung/tyoung.htm
18. http://www.manhattanrarebooks-science.com/young.htm
19. http://www.phy.hr/~dpaar/fizicari/xmaxwell.html
20. http://vacuum-physics.com/Maxwell/maxwell_oplf.pdf
21. http://www.tyndall.ac.uk/general/history/john_tyndall_biography.shtml
22. http://books.google.com/books?id=gWMSAAAAIAAJ&printsec=frontcover&dq=scientific+papers+of+rayleigh#PPR3,M1
23. J. Hecht, *City of Light*, Oxford, 1999.
24. http://phisicist.info/hertz.html
25. David Park, *The Fire within the Eye: A Historical Essay on the Nature and Meaning of Light*, Princeton University Press, 1997.
26. J. F. Mulligan, "Who Were Fabry and Pérot?" *Am. J. Phys.*, Vol. 66, No. 9, pp. 798–802, September 1998 [available at the website http://www.physics.rutgers.edu/ugrad/387/Mulligan98.pdf].
27. http://books.google.com/books?id=WcDPrrf1-oQC&pg=PA56&ots=FNeFkfdlW_&dq=Papers+of+Albert+Einstein&sig=OADkr1YehpFPm9xCR2lpKlYNsz0#PPA110,M1
28. G. Batchelor, *The Life and Legacy of G. I. Taylor*, Cambridge University Press, 1994.
29. http://books.google.com/ooks?id=98sCh99YIJsC&dq=Papers+of+Arthur+H+Compton
30. G. Venkataraman, *Journey into Light: Life and Science of C. V. Raman*, Penguin Books, 1994.
31. http://www.greatachievements.org/?id=3717
32. Nick Taylor, *Laser: The Inventor, the Nobel Laureate, and the Thirty-Year Patent War*, Simon & Schuster Publishers, New York, 2001.
33. Mario Bertolotti, *The History of the Laser*, Institute of Physics Publishing, Philadelphia, 2005.
34. http://books.google.com/books?id=JObDnEtzMJUC&pg=PA241&lpg=PA241&dq=javan+bennett+and+herriot&source=web&ots=tvP2kA2aNb&sig=K7KDtOYXjgHjTPaYwpOxXquumT4#PPP9,M1
35. http://www.llnl.gov/nif/library/aboutlasers/how.html
36. http://www.press.uchicago.edu/Misc/Chicago/284158_townes.html
37. http://www.farhangsara.com/laser.htm
38. http://www.bell-labs.com/history/laser/contrib.html
39. http://www.zeiss.com/C125716F004E0776/0/D94B4F1F28466E2AC125717100513E1A/$File/Innovation_7_35.pdf
40. http://www.ieee.org/organizations/foundation/donors.html
41. http://www.bell-labs.com/history/laser/contrib.html
42. C. K. Kao and G. A. Hockham, "Dielectric-Fibre Surface Waveguides for Optical Frequencies," *Proc. IEE*, Vol. 113, No. 7, 1151, 1966.
43. http://www.beyonddiscovery.org/content/view.page.asp? I=448
44. http://www.bell-labs.com/history/laser/contrib.html
45. http://www.cap.ca/awards/press/1998-hill.html

Chapter Two

WHAT IS LIGHT?

For the rest of my life, I will reflect on what light is.

—Albert Einstein, Circa 1917

All the fifty years of conscious brooding have brought me no closer to the answer to the question, 'What are light quanta?' Of course today every rascal thinks he knows the answer, but he is deluding himself.

—Albert Einstein, 1951

2.1 INTRODUCTION

Humans have always been interested to know what light is. In the early days, a light beam was thought to consist of particles. Later, the phenomena of interference and diffraction were demonstrated which could be explained only by assuming a wave model of light. Much later, it was shown that phenomena such as the photoelectric effect and the Compton effect could be explained only if we assume a particle model of light. Now, as we know, the values of the mass and charge of electrons, protons, alpha particles, etc., are known to a tremendous degree of accuracy—approximately one part in a billion! Their velocities can also be changed by the application of electric and magnetic fields. Thus, we usually tend to visualize them as tiny particles. However, they also exhibit diffraction and other effects which can be explained only if we assume them to be *waves*. Thus, the answers to the questions such as "What is an electron" or "What is light?" are very difficult. Indeed electrons, protons, neutrons, photons, alpha particles, etc., are *neither particles nor waves*. The modern quantum theory describes them in a very abstract way which cannot be connected with everyday experience. To quote Feynman (from Ref. 1):

> Newton thought that light was made up of particles, but then it was discovered that it behaves like a wave. Later, however (in the beginning of the twentieth century), it was found that light did indeed sometimes behave like a particle. Historically, the electron, for example, was thought to behave like a particle, and then in many respects it behaved like a wave. So it really behaves like neither. Now we have given up. We say: 'it is like neither.' There is one lucky break, however—electrons behave just like light. The quantum behaviour of atomic objects (electrons, protons, neutrons, photons, and so on) is the same for all, they are all 'particle–waves,' or whatever you want to call them.

In this chapter, we will make a brief historical survey of the important experiments which led to models regarding the nature of light. Near the end of the chapter, we will qualitatively discuss how the wave and the particle aspects of radiation can be explained on the basis of the uncertainty principle and the probabilistic interpretation of matter waves.

2.2 THE CORPUSCULAR MODEL

The corpuscular model is perhaps the simplest model of light. According to it, a luminous body emits a stream of particles in all directions. Isaac Newton, in his book *Opticks* (Ref. 2), wrote, "Are not the ray of light very small bodies emitted from shining substance?" The particles are assumed to be very tiny so that when two light beams overlap, a collision between the two particles rarely occurs. Using the corpuscular model, one can explain the laws of reflection and refraction in the following manner.

The reflection law follows by considering the elastic reflection of a particle by a plane surface. To understand refraction, we consider the incidence of a particle at a plane surface ($y = 0$) as shown in Fig. 2.1; we are assuming that the motion is confined to the xy plane. The trajectory of the

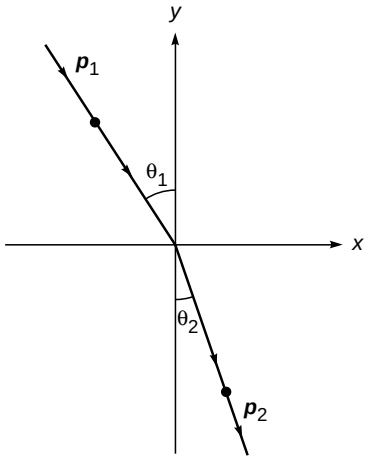


Fig. 2.1 Refraction of a corpuscle.

particle is determined by the conservation of the x component of the momentum ($= p \sin\theta$) where θ is the angle that the direction of propagation makes with the y axis. The conservation condition leads to equation

$$p_1 \sin\theta_1 = p_2 \sin\theta_2 \qquad \textbf{(1)}$$

where the angles θ_1 and θ_2 are defined in Fig. 2.1. The above equation directly gives Snell's law:

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{p_2}{p_1} = \frac{v_2}{v_1} \qquad \textbf{(2)}$$

To understand the explanation of Snell's law of refraction using the corpuscular model, we consider a simple experiment in which a ball moving with a certain speed on a horizontal surface moves down to a lower horizontal surface through a slope. Two stroboscopic pictures of the motion of the ball are shown in Fig. 2.2. The component of the momentum parallel to the edge of the slope does not change; however, the component perpendicular to the edge increases in value, resulting in an increased speed of the ball. Conversely, if the ball initially moves on the lower surface approaching the slope, the speed decreases as it goes up the slope; this is consistent with the reversibility of rays undergoing refraction. The slope can be approximately assumed to represent the interface between the two media.

Although the simple corpuscular model of light explains Snell's law of refraction satisfactorily, it predicts that if the ray moved toward the normal (i.e., if the refraction occurs at a denser medium), its speed would become higher which, as we shall see later, is not consistent with experimental observations. The wave theory does make the correct prediction of the ratio of velocities of waves in the two media.

We may mention here that in 140 AD**,** Claudius Ptolemy measured the angle of refraction in water for different angles of incidence in air and made a table of it. Using this table,

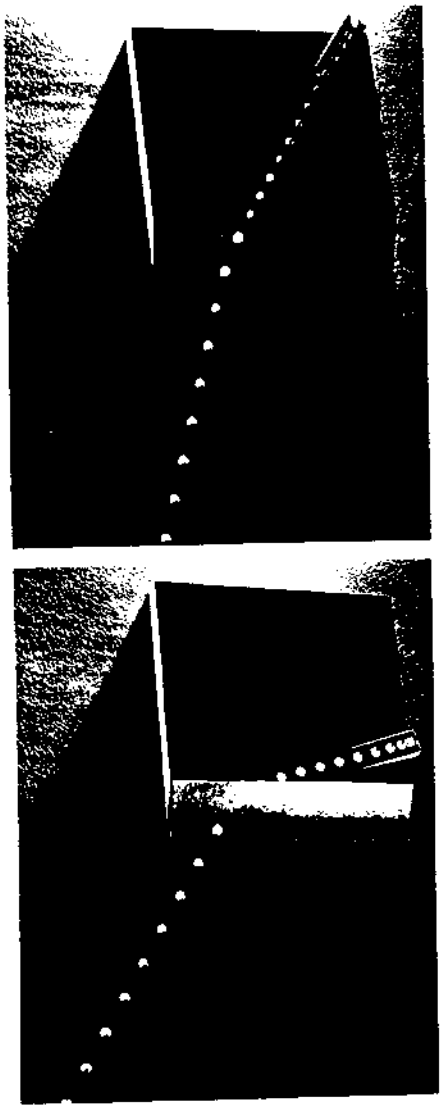


Fig. 2.2 Stroboscopic pictures of a ball with a certain speed on a horizontal surface moving down to a lower horizontal surface through a slope [Adapted from *PSSC, Physics,* D.C. Heath & Co., Boston, Mass., 1965; used with permission].

Willebrord Snell, in 1621, discovered the law of refraction which is known as Snell's law. In 1637, René Descartes derived Snell's law; this derivation is equivalent to the corpuscular derivation which is usually attributed to Newton. Newton (1642–1727) was only about 8 years old when Descartes (1596–1650) died, and therefore Descartes did not get the corpuscular model from Newton! The first edition of Newton's *Opticks* (in which Newton had discussed the corpuscular model) was published in 1704 (Ref. 2). It is probably because of the popularity of Newton's *Opticks* that the corpuscular theory is usually attributed to Newton; an English translation

of Descartes' original paper appears in a paper by Joyce and Joyce (Ref. 3). Descartes' theory remained undisputed until about 1662 when Fermat enunciated the principle of least time. Using this principle, Fermat derived Snell's law (see Chap. 3) and showed that if the velocity of light in the second medium is less, the ray will bend toward the normal, contrary to what is predicted by the corpuscular theory.

Finally, according to Newton, corpuscles of different sizes give rise to the sensation of different colors at the retina of the eye. He explained the prismatic spectrum by assuming that particles of different sizes refract at different angles. Commenting on Isaac Newton's *Opticks*, Einstein wrote:

> . . . this new edition of his [Newton's] work on Optics is nevertheless to be welcomed with warmest thanks, because it alone can afford us the enjoyment of a look at the personal activity of this unique man. . . . in one person he [Newton] combined the experimenter, the theorist, the mechanic and, not the least, the artist in exposition. He stands before us strong, certain and alone: his joy in creation and his minute precision are evident in every word and in every figure.

Perhaps the two most important experimental facts which led to the early belief in the corpuscular model of light were (1) the rectilinear propagation of light which results in the formation of sharp shadows and (2) that light could propagate through vacuum. The domain of optics in which light is assumed to travel in straight lines is known as **geometrical optics,** which can easily be explained on the basis of the corpuscular model of light. However, as careful experiments later showed, shadows are not perfectly dark; some light does enter the geometrical shadow which is due to the phenomenon of diffraction. This phenomenon is essentially due to the wave character of light and cannot be explained on the basis of the simple corpuscular model. As we shall see in later chapters, diffraction effects are usually difficult to observe because the wavelengths associated with light waves are extremely small. We should mention here that if we are under the shade of a building, then under the shade we can always read a book—the light that enters the shadow is due *not* to diffraction but to scattering of light by air molecules. This phenomenon of scattering is also responsible for the blue color of the sky and the red color of the setting Sun. If the Earth did not have an atmosphere, then the shadows would have been extremely dark, which is indeed the case on the surface of the Moon (see Fig. 2.3). Since the Moon does not have an atmosphere, the shadows would be extremely dark and we would never be able to read a book in our own shadow! Also on the surface of the Moon, the sky appears perfectly dark (see Fig. 2.3). However, even on the surface of the Moon, a small amount of light does enter the geometrical shadow because of diffraction.

(a)

(b)

Fig. 2.3 Photographs on the Moon. Because the Moon does not have any atmosphere, the sky and shadows are very dark. In (a) we can also see the Earth; color photographs appear in the insert at the back of the book.

2.3 THE WAVE MODEL

Although the corpuscular model explains the propagation of light through free space and can be made to predict the correct forms of the laws of reflection and refraction, a large

number of experimental observations (such as interference, diffraction, and polarization) arose which could not be explained on the basis of the corpuscular model of light. Historically, *Newton's rings*, which are a beautiful manifestation of the wave character of light, were observed by Hooke around the middle of the seventeenth century; the rings are named after Newton because he had given an explanation of their formation using the corpuscular model, which was later found to be quite unsatisfactory. The explanation of Newton's rings on the basis of the wave model is discussed in Chap. 15; Newton's explanation of the rings can be found in many places (see, e.g., Ref. 4).

Around 1665, Francesco Grimaldi, an Italian physicist, was probably the first person to observe the phenomenon of diffraction of white light as it passed through small apertures; Grimaldi concluded that—to quote from the Internet—"light is a fluid that exhibits wave-like motion." Later, Hooke also observed this phenomenon. As will be discussed in later chapters, a satisfactory explanation of the diffraction phenomenon can be given only if one assumes a wave model of light. This model was first put forward by Huygens in 1678 (Ref. 5). Using the wave model, Huygens could explain the laws of reflection and refraction (see Chap. 12), and he could also interpret the phenomenon of double refraction (see Chap. 22) discovered in 1669 by the Danish physicist Erasmus Bartholinus. However, so compelling was Newton's authority that it is said that *people around Newton had more faith in his corpuscular theory than Newton himself* and no one believed in Huygens' wave theory until 1801 when Thomas Young performed the famous interference experiment which could only be explained on the basis of a wave model of light (see Chap. 14). In addition, at the time of Huygens, light was thought to travel in straight lines, and Huygens tried to invoke unrealistic assumptions to explain the rectilinear propagation of light using his wave theory. This drawback was also one of the reasons for the immediate nonacceptance of the wave model. Young showed that the wavelength of light waves was about 6×10^{-5} cm. Because of the smallness of the wavelength, the diffraction effects are small and therefore light approximately travels in straight lines. Indeed, the branch of optics in which one completely neglects the finiteness of the wavelength is called geometrical optics, and a *ray* is defined as the path of energy propagation in the limit of $\lambda \to 0$.

In 1802, Young gave a satisfactory explanation of the formation of Newton's rings. In 1808, Malus observed the polarization of light, but he did not try to interpret this phenomenon. In 1816, Fresnel gave a satisfactory explanation of the diffraction phenomenon by means of a wave theory and calculated the diffraction patterns produced by various types of apertures and edges. In 1816, Fresnel along with Arago performed the famous experiment on the superposition of linearly polarized light waves, which was explained by Young, by assuming that light waves were transverse in character.

As mentioned above, both interference and diffraction phenomena could only be explained by assuming a wave model of light. It was believed that a wave would always require a medium and since light could propagate through vacuum, the presence of an "all pervasive" medium called the ether was assumed. In 1832 Fresnel derived the expressions for the reflection and transmission coefficients[1] by using models for ether vibrations. Poisson, Navier, Cauchy, and many other physicists contributed to the development of the ether theory which also necessitated the development of the theory of elasticity. There were considerable difficulties in the explanation of the models, and since we now know that ether does not exist, we will not go into the details of the various theories.

The nineteenth century also saw the development of electricity and magnetism. In 1820, Öersted discovered that currents caused magnetic effects. Soon after, Ampere found that two parallel wires carrying currents attract each other. Around 1830, Faraday carried out experiments which showed that a varying magnetic field induces an electromotive force; similar experiments were also carried out by Henry around the same time, and the law is also referred to as the Faraday-Henry law.

Soon afterward, Maxwell generalized Ampere's law by stating that a changing electric field can also set up a magnetic field. He summed up all the laws of electricity and magnetism in the form of equations [which are now referred to as Maxwell's equations (see Chap. 23)]. From these equations, he derived a wave equation and predicted the existence of electromagnetic waves (see Chap. 23). From the wave equation so derived, he showed that the velocity of the electromagnetic waves can be calculated from experiments in which a certain quantity of electric charge is measured by two different methods. These measurements were carried out in 1856 by Kohlrausch and Weber, and from their data, Maxwell found that the speed of the electromagnetic waves in air should be about 3.107×10^8 m s^{-1}. He found that this value was very close to the measured value of the speed of light, which according to the measurement of Fizeau in 1849 was 3.14858×10^8 m s^{-1}. The sole fact that the two values were

[1] The derivation of the Fresnel laws on the basis of electromagnetic theory will be discussed in Chap. 24. A derivation similar to the original derivation is given in Ref. 6.

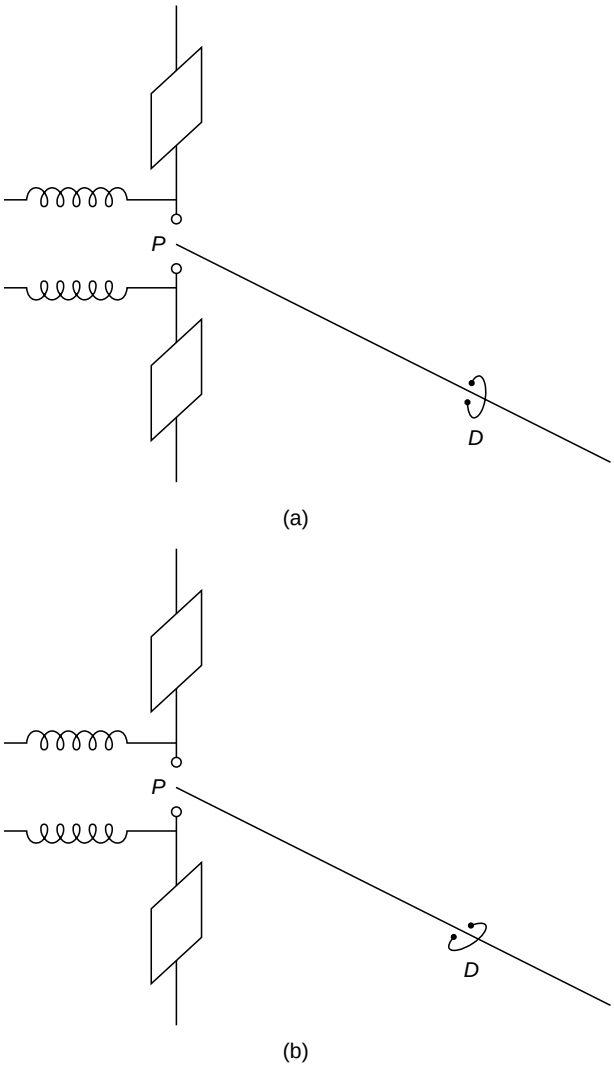


Fig. 2.4 Schematic of Hertz's experiment for generation and detection of electromagnetic waves. Sparks across the gap *P* produce electromagnetic waves whose frequency depends on the inductance and capacitance of the circuit. The electromagnetic waves can be detected by means of a detector *D*, which is nothing but a short wire bent in the form of a circle with a small gap. A signal is detected if the gap in the detector is parallel to the line joining the knobs of the spark gap *P* as shown in (a); if the gap is at right angles as shown in (b), no signal is received.

very close to each other led Maxwell to propound (around 1865) his famous electromagnetic theory of light, according to which

Light waves are electromagnetic waves.

Associated with a light wave are changing electric and magnetic fields; the changing magnetic field produces a time- and space-varying electric field; and the changing electric field produces a time- and space-varying magnetic field, and this results in the propagation of the electromagnetic wave even in free space. In 1888, Heinrich Hertz carried out experiments which could produce and detect electromagnetic waves of frequencies smaller than those of light. These waves were produced by discharging electrically charged plates through a spark gap. The frequency of the emitted electromagnetic waves depended on the values of the inductance and capacitance of the circuit. The electromagnetic waves could be detected by means of a detector, and it was found that a signal was not received when the detector was placed parallel to the source[2] (see Fig. 2.4).

Hertz also produced standing electromagnetic waves by getting them reflected by a metal sheet (see Figs. 13.3 and 13.4). He could calculate the wavelength of the waves, and knowing the frequency, he showed that the speed of the electromagnetic waves was the same as that of light. Using a collimated electromagnetic wave and getting it reflected by a metal sheet, he could demonstrate the laws of reflection. Hertz's experimental results provided a dramatic confirmation of Maxwell's electromagnetic theory. In addition, there were so many other experimental results, which were quantitatively explained by using Maxwell's theory, that toward the end of the nineteenth century, physicists thought that one had finally understood what light really is.

2.4 THE PARTICLE NATURE OF RADIATION

In 1887 Heinrich Hertz, while carrying out his experiments on electromagnetic waves, found that if the light emitted from one spark gap were blocked, it would reduce the maximum spark length in the other gap. After carrying out a series of experiments, he concluded that it was the ultraviolet radiation from the first spark that was helping the electrons to jump across the gap. Hertz reported the observations but did not pursue further and also did not make any attempt to explain them. In 1897, J. J. Thomson discovered electrons, and in 1899, he showed that electrons are emitted when light falls on a metal surface. In 1902, Philip Lenard observed that (1) the kinetic energy of the emitted electrons was independent of the intensity of the incident light and (2) that the energy of the emitted electron increased when the frequency of the incident light was increased. As will be discussed in Sec. 25.2, this phenomenon

[2] This follows directly from the dipole radiation pattern—see Sec. 22.4.1; in Fig. 22.4, the dipole is oscillating along the *z* axis, and the electric field on the *y* axis is along the *z* axis.

cannot be explained by a theory based on the wave model of light. In 1905, Einstein interpreted the photoelectric effect by putting forward his famous photon theory according to which light consisted of quanta of energy

$$E = h\nu \tag{3}$$

where ν is the frequency and h ($\approx 6.626 \times 10^{-34}$ J s) is Planck's constant; and that the emission of a photoelectron was the result of the interaction of a single quantum (i.e., of the photon) with an electron. In his 1905 paper (Ref. 7 and reprinted in Ref. 8), Einstein wrote that

> Monocromatic radiation behaves as if it consists of mutually independent energy quanta of magnitude $h\nu$.

In the same paper he also wrote that

> According to Maxwell's theory, energy is considered to be a continuous spatial function for all purely electromagnetic phenomena, hence also for light. . . . The wave theory of light, which operates with continuous spatial functions, has proved itself superbly . . .

In 1909, Einstein wrote (quoted from Ref. 9)

> It is undeniable that there is an extensive group of data concerning radiation which shows that light has certain fundamental properties that can be understood much more readily from the standpoint of the Newton emission (particle) theory than from the standpoint of the wave theory. It is my opinion, therefore, that the next phase of the development of theoretical physics will bring us a theory of light that can be interpreted as a kind of fusion of the wave and emission theories.

We may note the prediction of Einstein. To quote from Ref. 10:

> Owing to Einstein's paper of 1905, it was primarily the photoelectric effect to which physicists referred as an irrefutable demonstration of the existence of photons and which thus played an important part in the conceptual development of quantum mechanics.

It was only in 1926 that Gilbert Lewis, a U.S. chemist, coined the word *photon* to describe Einstein's *localized energy quanta*. Einstein received the 1921 Nobel Prize in Physics for his services to theoretical physics and especially for his explanation of the photoelectric effect. Einstein's photon theory predicted that if the frequency ν of the incident radiation were greater than the critical frequency ν_c, then the kinetic energy of the emitted electron would be $h(\nu - \nu_c)$, which was later verified by Millikan for visible light, by de Broglie for X-rays, and by Thibaud and Ellis for γ-rays. Einstein also showed that the photons, in addition to having an energy equal to $h\nu$, should have a momentum given by

$$p = \frac{h\nu}{c} = \frac{h}{\lambda} \tag{4}$$

which was verified experimentally in 1923 by Compton. This experiment is known as the Compton effect and will be discussed in detail in Sec. 26.3. Compton received the 1927 Nobel Prize in Physics for his discovery of the effect named after him.

In the year 1900, Max Planck put forward his famous theory of blackbody radiation, the derivation of which presupposed that energy can be absorbed and emitted by an individual resonator only in *quanta* of magnitude $h\nu$. According to Einstein:

> . . . I could nevertheless see to what kind of consequences this law [i.e., Planck's law] of temperature-radiation leads for the photoelectric effect and for other related phenomena of the transformation of radiation-energy, as well as for the specific heat of [especially] solid bodies. All my attempts however, to adapt the theoretical foundation of physics to this [new type of] knowledge failed completely. It was as if the ground had been pulled out from under one, with no firm foundation to be seen anywhere, upon which one could have built. That this insecure and contradictory foundation was sufficient to enable a man of Bohr's unique instinct and tact to discover the major laws of the spectral lines and of the electron shells of the atoms together with their significance for chemistry appeared to me like a miracle—and appears to me as a miracle even today. This is the highest form of musicality in the sphere of thought. (Quoted from the autobiographical notes by Einstein in *Albert Einstein: Philosopher, Scientist*, edited by P. A. Schilpp, Tudor Publishing Co., New York, 1951.)

In making this transition from Planck's *quantized oscillators* to *quanta of radiation*, Einstein made a very important conceptual transition; namely, he introduced the idea of corpuscular behavior of radiation. Although Newton had described light as a stream of particles, this view had been completely superseded by the wave picture of light, a picture that culminated in the electromagnetic theory of Maxwell. The revival of the particle picture now posed a severe conceptual problem, one of reconciling wavelike and particlelike behavior of radiation. It also soon became apparent that matter also exhibited both types of behavior. For example, an electron with

an accurately measured value of mass and charge could undergo diffraction in a manner similar to that of light waves. We will now give a brief account of some of the other important experimental evidence showing wave-particle duality that led to the development of the quantum theory.

2.5 WAVE NATURE OF MATTER

Experiments by Wilson with his cloud chamber had clearly shown the particlelike behavior of alpha and beta particles. These are emitted by radioactive elements, and when they pass through supersaturated vapor, they form tracks of condensed droplets. For alpha particles, these tracks are nearly straight lines; however, for electrons, the tracks are irregularly curved. The existence of continuous tracks suggests that the emissions from the radioactive substance can be regarded as minute particles moving with high speed. Further, the fact that electrons could be deflected by electric and magnetic fields and also the fact that one could accurately determine the ratio of their charge to mass suggest very strongly that electrons are particles. This view remained unchallenged for a number of years; C. T. R. Wilson was awarded the 1927 Nobel Prize in Physics for his method of making the paths of electrically charged particles visible by condensation of vapor.

At this stage, one could ask if matter may not show wavelike behavior also just as light exhibited corpuscular and wavelike behavior. In 1925, de Broglie proposed just such a hypothesis and argued that the relation given by Eq. (4) between wavelength and momentum applied for electrons as well. In his 1925 paper, he wrote that

> The basic idea of quantum theory is, of course, the impossibility of considering an isolated fragment of energy without assigning a certain frequency to it . . .

De Broglie was awarded the 1929 Nobel Prize in Physics for his discovery of the wave nature of electrons. In the presentation speech (on December 12, 1929), the chairman of the Nobel Committee for Physics said that

> Louis de Broglie had the boldness to maintain that not all the properties of matter can be explained by the theory that it consists of corpuscles. . . . At a time when no single known fact supported this theory, Louis de Broglie asserted that a stream of electrons which passed through a very small hole in an opaque screen must exhibit the same phenomena as a light ray under the same conditions. . . . The experimental results obtained have fully substantiated Louis de Broglie's theory. Hence there are not two worlds, one of light and waves, one of matter and corpuscles. There is only a single universe.

Later, de Broglie wrote (quoted from p. 58 of Ref. 9):

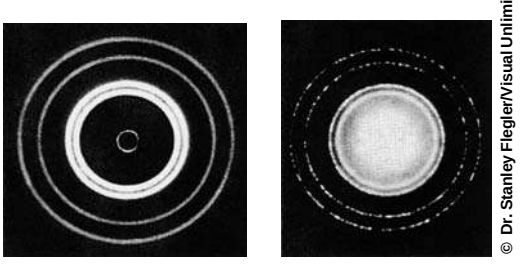
> I was convinced that the wave-particle duality discovered by Einstein in his theory of light quanta was absolutely general and extended to all of the physical world, and it seemed certain to me, therefore, that the propagation of a wave is associated with the motion of a particle of any sort—photon, electron, proton or any other.

In 1927, Davisson and Germer studied the diffraction of electrons from single crystals of nickel and showed that the diffraction patterns could be explained if the electrons were assumed to have a wavelength given by the de Broglie relation

$$\lambda = \frac{h}{p} \tag{5}$$

where p is the momentum of the electron. Shortly afterward, in 1928, G. P. Thomson carried out electron diffraction experiments by passing electrons through thin polycrystalline metal targets (see Sec. 18.10 for more details). The diffraction pattern consisted of concentric rings similar to the Debye-Scherrer rings obtained in the X-ray diffraction pattern. By measuring the diameters of the rings and from the known structure of the crystals, Thomson calculated the wavelength associated with the electron beam which was in agreement with the de Broglie relation [Eq. (5)]. In 1937, Davisson and Thomson shared the Nobel Prize for their experimental discovery of the diffraction of electrons by crystals. Max Jammer (Ref. 10) has written, "Thomson, the father, was awarded the Nobel Prize for having shown that the electron is a particle, and Thomson, the son, for having shown that the electron is a wave."

In Fig. 2.5 we have shown Debye-Scherrer rings produced by scattering of X-rays [see (a)] and by scattering of electrons [see (b)] by an aluminum foil. The two figures clearly show the similarity in the wavelike properties of X-rays and of electrons.



(a) (b)

Fig. 2.5 The diffraction pattern of aluminum foil produced (a) by X-rays and (b) by electrons; notice the similarity in the diffraction patterns.

2.6 THE UNCERTAINTY PRINCIPLE

The reconciliation of the corpuscular nature with the wave character of light (and also of the electron) has been brought about through the modern quantum theory; and perhaps the best known consequence of wave-particle duality is the uncertainty principle[3] of Heisenberg, which can be stated as follows:

> If the x coordinate of the position of a particle is known to an accuracy Δx, then the x component of the momentum cannot be determined to an accuracy better than $\Delta p_x \approx h/\Delta x$, where h is Planck's constant.

Alternatively, one can say that if Δx and Δp_x represent the accuracies with which the x coordinate of the position and the x component of the momentum can be determined, then the following inequality must be satisfied:

$$\Delta x \Delta p_x \geq h \tag{6}$$

We do not feel the effect of this inequality in our everyday experience because of the smallness of the value of Planck's constant ($\approx 6.6 \times 10^{-27}$ erg s). For example, for a tiny particle of mass 10^{-6} g, if the position is determined within an accuracy of about 10^{-6} cm, then according to the uncertainty principle, its velocity cannot be determined within an accuracy better than $\Delta v \approx 6 \times 10^{-16}$ cm s^{-1}. This value is much smaller than the accuracies with which one can determine the velocity of the particle. For a particle of a greater mass, Δv will be even smaller. Indeed, had the value of Planck's constant been much larger, the world would have been totally different. In a beautifully written book, Gamow (Ref. 13) has discussed what our world would be like if the effect of the uncertainty principle were perceivable by our senses.

2.7 THE SINGLE-SLIT DIFFRACTION EXPERIMENT

We will now show how the diffraction of a light beam (or an electron beam) can be explained on the basis of the corpuscular nature of radiation and the uncertainty principle. Consider a long narrow slit of width b as shown in Fig. 2.6. Now, one can always choose the distance between the source and the slit large enough that p_x can be assumed to have an arbitrarily small value. For example, for the source at a distance d, the maximum value of p_x of the photons approaching the slit will be

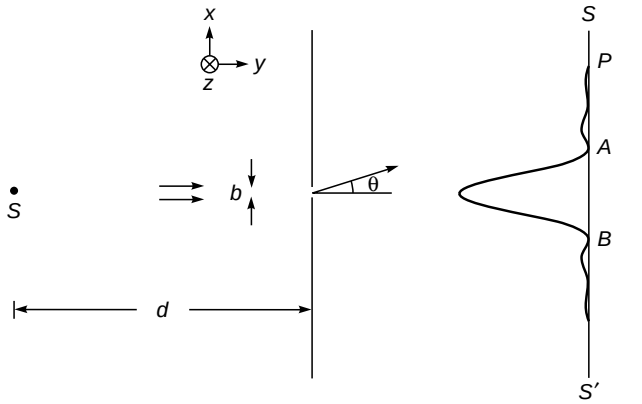


Fig. 2.6 Diffraction of a photon (or an electron) beam by a narrow slit of width b.

$$p_x \approx p\frac{b}{d} = \frac{h\nu}{c}\cdot\frac{b}{d}$$

which can be made arbitrarily small by choosing a large enough value of d. Thus we may assume the light source to be sufficiently far away from the slit that the photons approaching the slit can be assumed to have momentum in only the y direction. Now, according to the particle model of radiation, the number of particles reaching the point P (which lies in the geometrical shadow) will be extremely small; further, if we decrease the width of the slit, the intensity should decrease, which is quite contrary to the experimental results because we know that the beam undergoes diffraction and the intensity at a point such as P would normally increase if the width of the slit were made smaller. Thus, the classical corpuscular model is quite incapable of explaining the phenomenon of diffraction. However, if we use the uncertainty principle in conjunction with the corpuscular model, the diffraction phenomenon can be explained in the following manner: When a photon (or an electron) passes through the slit, we can say that

$$\Delta x \approx b$$

which implies that we can specify the position of the photon to an accuracy b. If we now use the uncertainty principle, we have

$$\Delta p_x \approx \frac{h}{b} \tag{7}$$

[3] The uncertainty principle can be derived from the Schrödinger equation (see, e.g., Chap. 5 of Ref. 12). And as mentioned by Richard Feynman, "Where did we get that [the Schrödinger equation] from? Nowhere. It is not possible to derive it from anything you know. It came out of the mind of Schrödinger."

i.e., just by making the photon pass through a slit of width b, the slit imparts a momentum in the x direction which is $\approx h/b$. It may be pointed out that *before* the photon entered the slit, p_x (and hence Δp_x) can be made arbitrarily small by putting the source sufficiently far away. Thus we may write $\Delta p_x \approx 0$. It would, however, be wrong to say that by making the photon pass through the slit, $\Delta p_x \, \Delta x$ is zero; this is so because of the fact that $\Delta p_x \approx 0$ *before* the photon entered the slit. After the photon has entered the slit, it is confined within a distance b in the x direction, and hence $\Delta p_x \approx h/b$. Further, since before the photon entered the slit $p_x \approx 0$, we will therefore have

$$|p_x| \approx \Delta p_x \approx \frac{h}{b}$$

But $p_x = p \sin \theta$, where θ is the angle that the photon coming out of the slit makes with the y axis (see Fig. 2.6). Thus

$$p \sin \theta \approx \frac{h}{b}$$

or

$$\sin \theta \approx \frac{h}{pb} \qquad \textbf{(8)}$$

The above equation predicts that the possibility of a photon traveling at an angle θ with the y direction is inversely proportional to the width of the slit; i.e., the smaller the value of b, the greater the value of θ and the greater the possibility of the photon to reach deep inside the geometrical shadow. This is indeed the diffraction phenomenon. Now, the momentum of a photon is given by

$$p = \frac{h}{\lambda}$$

Thus Eq. (8) becomes

$$\sin\theta \approx \frac{\lambda}{b} \qquad \textbf{(9)}$$

which is the familiar diffraction theory result, as will be discussed in Sec. 18.3. We can therefore say that the wave-particle duality is a consequence of the uncertainty principle and the uncertainty principle is a consequence of the wave-particle duality. To quote Max Born (Ref. 14),

> Physicists of today have learnt that not every question about the motion of an electron or a photon can be answered, but only those questions which are compatible with the uncertainty principle.

De Broglie suggested that the equation $\lambda = h/p$ is valid not only for photons but also for all particles such as electrons, protons, and neutrons. Indeed, the de Broglie relation has been verified by studying the diffraction patterns produced when electrons, neutrons, etc., pass through a single crystal; the patterns can be analyzed in a manner similar to X-ray diffractions (see Sec. 18.10). In Fig. 2.7, we show the experimental data of Shull (Ref. 15) who studied the Fraunhofer diffraction of neutrons by a single slit, and his experimental results agree with the intensity distribution as predicted by the wave theory with λ given by Eq. (5).

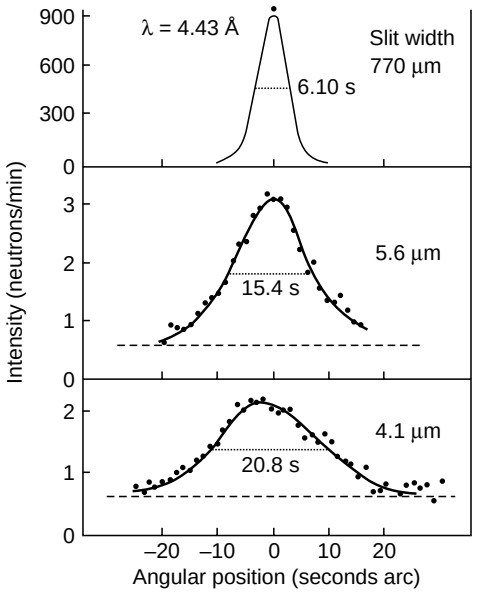


Fig. 2.7 Angular broadening of a neutron beam by small slits [After Ref. 15].

2.8 THE PROBABILISTIC INTERPRETATION OF MATTER WAVES

In Sec. 2.7 we saw that if a photon passes through a slit of width b, then the momentum imparted in the x direction (which is along the width of the slit) is $\approx h/b$. The question arises whether we can predict the trajectory of an individual photon. The answer is no. We cannot say where an individual photon will land up on the screen; we can only predict the probabilities of arrival of the photon in a certain region of the screen. We may, for example, say that the probability for the arrival of the photon in the region lying between the points A and B (see Fig. 2.6) is 0.85. This would imply that if the experiment were carried out with a large number of photons, about 85% of them would land up in region AB; but the fate of an individual photon can never be predicted. This is in contrast to Newtonian mechanics where the trajectories are always predetermined. Also, if we place a light detector on the screen, then it will always record one photon or none and never one-half of a photon. This essentially implies the corpuscular nature of the radiation. However, the probability

distribution is the same as predicted by the wave theory, and therefore if one performs an experiment with a large number of photons (as is indeed the case in most experiments); the intensity distribution recorded on the screen is the same as that predicted by the wave theory.

To explicitly show that diffraction is not a many-photon phenomenon, Taylor in 1909 carried out a beautiful experiment which consisted of a box with a small lamp that casts the shadow of a needle on a photographic plate (see Fig. 2.8). The intensity of light was so weak that between the slit and the photographic plate, it was almost impossible to find two photons (see Example 2.1). In fact, to get a good fringe pattern, Taylor made an exposure lasting several months. The diffraction pattern obtained on the photographic plate was the same as that predicted by the wave theory.

The corpuscular nature of radiation and the fact that one cannot predict the trajectory of an individual photon can be seen from Fig. 2.9, which consists of series of photographs showing the quality of pictures obtainable from various numbers of photons (Ref. 16). The photograph clearly shows that the picture is built up by the arrival of concentrated packets of energy, and the point at which a particular photon will arrive is entirely a matter of chance. The figure also shows that the photograph is featureless when a small number of photons are involved; and as the number of photons reaching the photographic plate increases, the intensity distribution becomes the same as would be predicted by the wave theory. To quote Feynman:

> . . . it would be impossible to predict what would happen. We can only predict the odds! This would mean, if it were true, that physics has given up on the problem of trying to predict exactly what will happen in a definite circumstance. Yes! physics has given up. We do not know how to predict what would happen in a given circumstance, and we believe now that it is impossible—that the only thing that can be predicted is the probability of different events. It must be recognized that this is a retrenchment in our earlier idea of understanding nature. It may be a backward step, but no one has seen a way to avoid it.

A somewhat similar situation arises in radioactivity. Consider a radioactive nucleus having a half-life of say 1 h. If we start with 1000 such nuclei, then on an average 500 of them will undergo radioactive decay in 1 h and about 250 of them in the next 1 h and so on. Thus, although to start with, all nuclei are identical, some nuclei would decay in the very first minute and other nuclei can survive for hours without

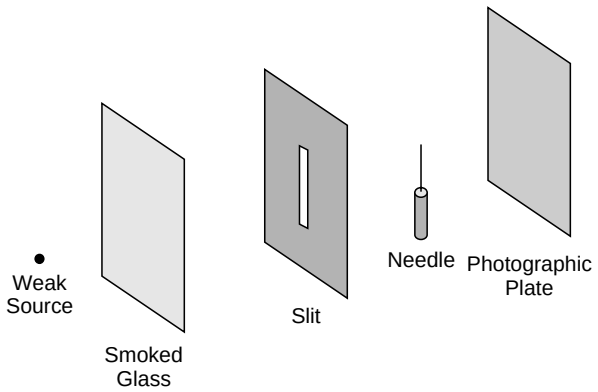


Fig. 2.8 Schematic diagram of the experimental arrangement of Taylor to study the diffraction pattern produced by a weak source. The whole apparatus was placed inside a box.

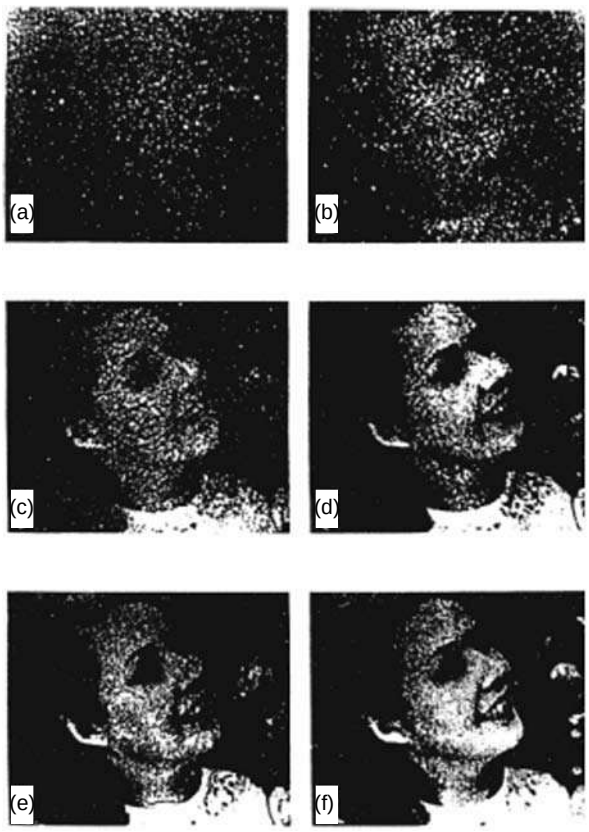


Fig. 2.9 Photographs showing the quality of a picture obtainable from various numbers of photons: (a), (b), (c), (d), (e), and (f) correspond to 3×10^3 photons, 1.2×10^4 photons, 9.3×10^4 photons, 7.6×10^5 photons, 3.6×10^6 photons, and 2.8×10^7 photons respectively. [From Ref. 16; reprinted with permission].

undergoing radioactive decay. Thus, one can never predict which nucleus will undergo decay in a specified period; one can predict only the probability of its undergoing decay in a certain interval of time. This is indeed a manifestation of quantum mechanics. To quote Feynman again:

> A philosopher once said it is necessary for the very existence of science that the same conditions always produced the same results. Well they don't!

2.9 AN UNDERSTANDING OF INTERFERENCE EXPERIMENTS

We consider the two-hole interference experiment similar to that performed by Young (see Secs. 14.4 and 14.5). The experimental arrangement is shown in Fig. 2.10 where a weak light source S_0 illuminates the hole S and the light emerging from the holes S_1 and S_2 produces the interference pattern on the screen PP'. The intensity is assumed to be so weak that in the region between the planes AB and PP' there is almost never more than one photon (see Example 2.1). Individual photons are also counted by a detector on the screen PP', and one finds that the intensity distribution has a $\cos^2$ pattern similar to that shown in Fig. 14.9. The corpuscular nature of the radiation is evident from its detection in the form of a single photon and never a fraction of a photon. The appearance of the interference pattern is due to the fact that a photon interferes with itself. Quantum theory tells us that a photon partially passes through the hole S_1 and partially through S_2. This is not the splitting of the photon into two halves, but only implies that if we wish to find out through which hole the photon passed, then one-half the time it will be found to have passed through the hole S_1 and one-half the time through S_2. The photon is in a state which is a superposition of two states, one corresponding to the wave emanating from hole S_1 and the other to the one emanating from hole S_2. The superposed state will give rise to an intensity distribution similar to that obtained by considering the superposition of two waves. Had we employed a device (such as a microscope) that determined which hole the photon had passed through, then the interference pattern on the screen would have been washed out. This is a consequence of the fact that a measurement always disturbs the system. This is very nicely discussed in Ref. 1. Thus we may say that the photons would arrive as packets of energy, but the probability distribution (on the screen) will be proportional to the intensity distribution predicted by using a wave model.

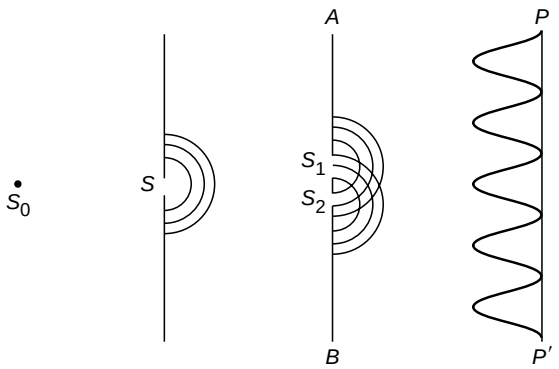


Fig. 2.10 Young's double-hole experimental arrangement for obtaining the interference pattern. Here S_0 represents a point source.

In a recent paper, Tonomura and his co-workers (Ref. 17) demonstrated the single electron buildup of an interference pattern. Their results are shown in Fig. 2.11. It may be seen that when there are very few electrons, they arrive randomly; however, when a large number of electrons are involved, one obtains an intensity distribution similar to the one predicted by wave theory.

We next consider the interference experiment involving the Michelson interferometer in which a light beam is partially reflected by a beam splitter and the resulting beams are made to interfere (see Fig. 2.12); the interference pattern produced by a Michelson interferometer is discussed in Chap. 15. According to Dirac (Ref.18),

> Some time before the discovery of quantum mechanics people realized that the connection between light waves and photons must be of a statistical character. What they did not clearly realize, however, was that the wave function gives information about the probability of one photon being in a particular place and not the probable number of photons in that place. The importance of the distinction can be made clear in the following way. Suppose we have a beam of light consisting of a large number of photons split up into two components of equal intensity. On the assumption that the beam is connected with the probable number of photons in it, we should have half the total number going into each component. If the two components are now made to interfere, we should require a photon in one component to be able to interfere with one in the other. Sometimes these two photons would have to annihilate one another and other times they would have to produce four photons. This would contradict the conservation of energy. The new theory, which connects the wave function

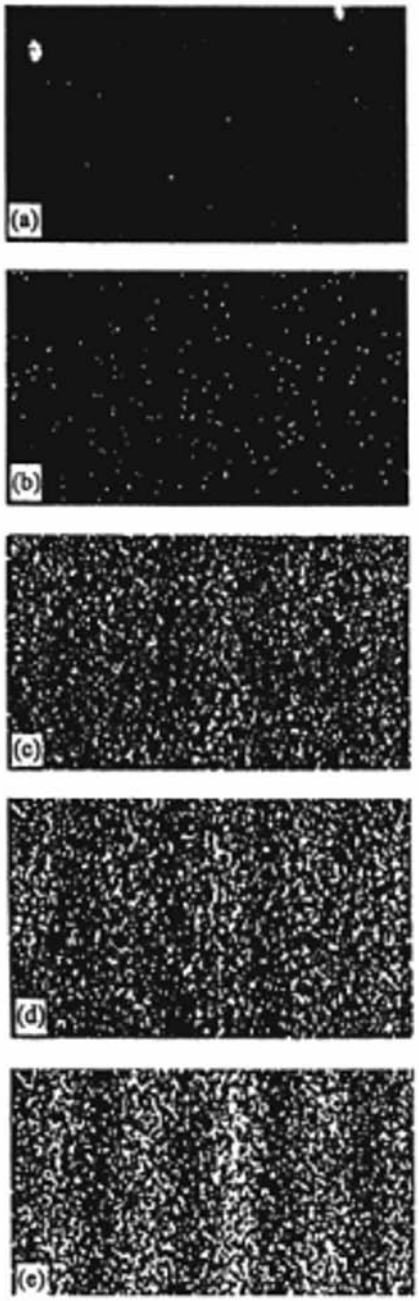


Fig. 2.11 Buildup of the electron interference pattern. Number of electrons in (a), (b), (c), (d), and (e) is 10, 100, 3000, 20,000, and 70,000, respectively. [Photographs reprinted with permission from A. Tonomura, J. Endo, T. Matsuda, T. Kawasaki, and H. Ezawa, "Demonstration of Single-Electron Build up of an Interference Pattern," American Journal of Physics, Vol. 57, No. 2, p. 117, 1989; copyright 1989, American Association of Physics Teachers.]

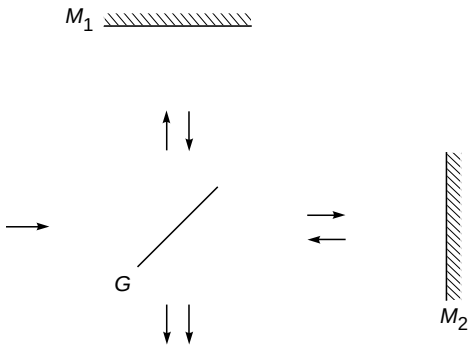


Fig. 2.12 Schematic of the setup of the Michelson interferometer. Here G represents a beam splitter and M_1 and M_2 represent plane mirrors.

with probabilities for one photon gets over the difficulty by making each photon go partly into each of the two components. Each photon then interferes only with itself. Interference between two different photons never occurs.

In the Michelson interferometer experiment, Dirac argues:

> . . . we describe the photon as going partly into each of the two components into which the incident beam is split. The photon is then, as we may say, in a translational state given by the superposition of the two translational states associated with the two components. . . . For a photon to be in a definite translational state it need not be associated with one single beam of light, but may be associated with two or more beams of light, which are the components into which one original beam has been split. In the accurate mathematical theory each translational state is associated with one of the wave functions of ordinary wave optics, which may describe either a single beam or two or more beams into which one original beam has been split.

These translational states can be superposed in a manner similar to the one employed while considering the interference of two beams. Thus, each photon goes partly into each of the two components and interferes only with itself. If we try to determine the fate of a single photon by measuring the energy in one of the components, then Dirac argues:

> The result of such a determination must be either a whole photon or nothing at all. Thus the photon must change suddenly from being partly in one beam and partly in the other to be entirely in one of the beams. This sudden change is due to the disturbance in the

translational state of the photon which the observation necessarily makes. It is impossible to predict in which of the two beams the photon will be found. Only the probability of either result can be calculated. . . . Our description of the photon allows us to infer that, after such an energy measurement, it would not be possible to bring about any interference effects between the two components. So long as the photon is partly in one beam and partly in the other, interference can occur when the two beams are superposed, but this probability disappears when the photon is forced entirely into one of the beams by an observation.

2.10 THE POLARIZATION OF A PHOTON

Let us consider the incidence of a plane electromagnetic wave on a polaroid whose pass axis is along the *y* direction (see Fig. 2.13); obviously, the electric vector of the transmitted wave would be along the *y* direction (see Chap. 22). Thus, if the electric vector associated with the incident wave oscillates along the *x* axis, the wave will be absorbed by the polaroid. On the other hand, if the electric vector oscillates along the *y* axis, it will just pass through the polaroid. Further, if the electric vector makes an angle θ with the pass axis, then the intensity of the transmitted beam will be $I_0 \cos^2\theta$, where I_0 represents the intensity of the incident beam (this is known as Malus' law which will be discussed in Sec. 22.3).

In the photon theory also one can associate a certain state of polarization with every photon. One can argue that if the electric vector associated with the photon is along the *y* (or the *x*) axis, then the photon will pass through or get absorbed by the polaroid. The question now arises as to what will happen to a single photon if the electric vector makes an angle θ with the pass axis. The answer is that the probability for the photon to pass through the polaroid is $\cos^2\theta$, and if the experiment is conducted with *N* photons (and if *N* is very large), then about $N \cos^2\theta$, photons will pass through; one cannot predict the fate of an individual photon.

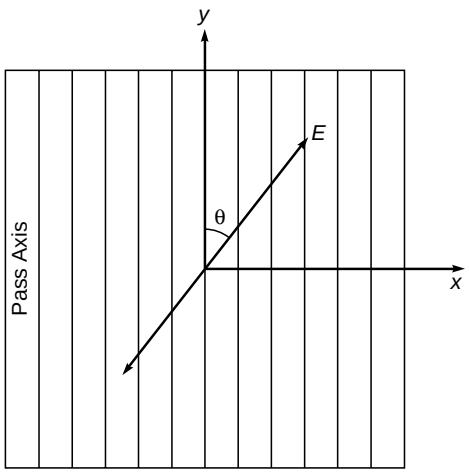


Fig. 2.13 The incidence on a polaroid of a linearly polarized light beam whose electric vector makes an angle θ with the *y* axis; the pass axis of the polaroid is along the *y* axis.

Example 2.1 Let a source (with $\lambda = 5 \times 10^{-5}$ cm) of power 1 W be used in the experimental arrangement shown in Fig. 2.10.

(a) Calculate the number of photons emitted by the source per second.

(b) Assume the radii of the holes *S*, S_1, and S_2 to be 0.02 cm and $S_0S = SS_1 = SS_2 = 100$ cm and the distance between the planes *AB* and *PP′* to be also 100 cm. Show that in the region between the planes *AB* and *PP′* one can almost never find two photons.

Solution:

(a) The energy of each photon will be

$$h\nu = \frac{hc}{\lambda} = \frac{6.6\times10^{-34}\,\text{J s}\times3\times10^{8}\,\text{m s}^{-1}}{5\times10^{-7}\,\text{m}}$$
$$\approx 4 \times 10^{-19}\ \text{J}$$

Thus the number of photons emitted per second will be

$$\frac{1\,\text{W}}{4\times10^{-19}\,\text{J}} = 2.5 \times 10^{18}$$

(b) The number of photons passing through the hole *S* will be approximately

$$\frac{2.5\times10^{18}\times\pi\times(0.02)^2}{4\times\pi\times(100)^2} = 2.5 \times 10^{10} \text{ per second}$$

Similarly, the number of photons passing through either S_1 or S_2 will be approximately

$$\frac{2.5\times10^{10}\times2\times\pi\times(0.02)^2}{2\pi\times(100)^2} = 1000 \text{ per second}$$

where we have assumed that after passing through *S*, the photons are evenly distributed in the hemisphere. This is strictly not correct because the diffraction pattern is actually an Airy pattern (see Chap. 18); nevertheless, the above calculations are qualitatively correct. The distance between the planes *AB* and *PP′* is 100 cm which will be traversed by a photon in time $\approx 3 \times 10^{-9}$ s. Thus, approximately every thousandth of a second a photon enters the region and the space is traversed much before the second photon enters. Therefore, in the region between *AB* and *PP′* one will (almost) never find two photons. This is somewhat similar to the case where, on average, 100 persons pass through a room in 1 yr and the time that each person takes to cross the room is ≈ 1 s; thus it will be highly improbable to have two persons simultaneously in the room.

Example 2.2 In this example we will use the uncertainty principle to determine the size of the hydrogen atom.[4] Although this example is not directly related to optics, it demonstrates the far-reaching consequences of the uncertainty principle.

We consider the hydrogen atom which consists of a proton and an electron. Since the proton is very much heavier than the electron, we consider only the motion of the electron. Let the electron be confined to a region of linear dimension $\approx a$. Thus according to the uncertainty principle;

$$p \approx \Delta p \approx \frac{\hbar}{a} \tag{10}$$

where $\hbar = h/(2\pi)$; the reason for using $\hbar$ rather than h will be mentioned later. The kinetic energy KE of the electron will be given by

$$\text{KE} = \frac{p^2}{2m} \approx \frac{\hbar^2}{2ma^2} \tag{11}$$

Now, there exists an electrostatic attraction between the two particles; the corresponding potential energy PE given by

$$\text{PE} = -\frac{q^2}{4\pi\varepsilon_0 a} \tag{12}$$

where q ($\approx 1.6 \times 10^{-19}$ C) represents the magnitude of the charge of the electron and ε_0 ($\approx 8.854 \times 10^{-12}$ C N^{-2} m^{-2}) represents the permittivity of free space. Thus the total energy E is given by

$$\begin{aligned} E &= \text{KE} + \text{PE} \\ &\approx \frac{\hbar^2}{2ma^2} - \frac{q^2}{4\pi\varepsilon_0 a} \end{aligned} \tag{13}$$

The system would settle to a state of lowest energy; thus we must set dE/da equal to zero:

$$0 = \frac{dE}{da} = -\frac{\hbar^2}{ma^3} + \frac{q^2}{4\pi\varepsilon_0 a^2}$$

implying

$$a = a_0 = \frac{\hbar^2}{m\left[q^2/(4\pi\varepsilon_0)\right]} \tag{14}$$

If we substitute the values of $\hbar$ ($\approx 1.055 \times 10^{-34}$ J s), m ($\approx 9.11 \times 10^{-31}$ kg), ε_0, and q, we obtain

$$a = a_0 \approx 0.53 \times 10^{-10} \text{ m} = 0.53 \text{ Å} \tag{15}$$

Thus we get the remarkable result that the size of the hydrogen atom is a direct consequence of the uncertainty principle. To quote Feynman:

> So we now understand why we do not fall through the floor.... In order to squash the atoms close together, the electrons would be confined to smaller space and by the uncertainty principle, their momenta would have to be higher on the average, and that means high energy; the resistance to atomic compression is a quantum mechanical effect ...

[4] The analysis is adapted from Ref. 1.

We next substitute the value of a from Eq. (13) into Eq. (12) to obtain

$$\begin{aligned} E &= \frac{\hbar^2}{2m}\left(\frac{m}{\hbar^2}\frac{q^2}{4\pi\varepsilon_0}\right)^2 - \frac{q^2}{4\pi\varepsilon_0}\left(\frac{m}{\hbar^2}\frac{q^2}{4\pi\varepsilon_0}\right) \\ &= -\frac{m}{2\hbar^2}\left(\frac{q^2}{4\pi\varepsilon_0}\right)^2 \end{aligned} \tag{16}$$

Substituting the values of $\hbar$, m, q, and ε_0, we get

$$\begin{aligned} E &\approx -2.17 \times 10^{-19} \text{ J} \\ &\approx -13.6 \text{ eV} \end{aligned} \tag{17}$$

which is nothing but the ground-state energy of the hydrogen atom. Thus, that the ionization potential of hydrogen atom is ≈ 13.6 eV follows from the uncertainty principle. We may point out that the uncertainty principle can be used to give only an order of magnitude of the size of the hydrogen atom or its ionization potential; we had intentionally chosen the constants in such a way that the ground-state energy comes out to be correct. It is for this reason that we had chosen $\hbar$ instead of h in Eqs. (10) and (11).

2.11 THE TIME-ENERGY UNCERTAINTY RELATION

When an atom makes a transition from an energy state E_2 to an energy state E_1 ($E_2 > E_1$), a photon of frequency $\nu = (E_2 - E_1)/h$ is emitted. This emission is essentially a pulse of a duration $\approx 10^{-9}$ s; this duration is usually denoted by τ. This leads to a frequency width $\Delta\nu$, and as will be shown in Chap. 17,

$$\tau\,\Delta\nu \geq 1 \tag{18}$$

Multiplying both sides by h, we get the time-energy uncertainty principle:

$$\Delta t\,\Delta E \geq h \tag{19}$$

which can be interpreted as follows: If Δt represents the uncertainty in the time at which a time-dependent process takes place, then the uncertainty ΔE in the energy of this process will be $\geq h/\Delta t$. Assuming $\Delta t \approx 10^{-9}$ s,

$$\Delta E \approx \frac{h}{\Delta t} \approx 4 \times 10^{-6} \text{ eV}$$

Summary

- The corpuscular model of light is due to Descartes rather than to Newton. The law of refraction was discovered experimentally in 1621 by Snell. In 1637, using corpuscular model, Descartes derived Snell's law of refraction.

- The wave model of light was first propounded by Huygens in 1678. Using the wave model, Huygens could explain the laws of reflection and refraction, and he could also interpret the phenomenon of double refraction.
- Around the middle of the nineteenth century, Maxwell generalized Ampere's law by stating that a changing electric field can also produce a magnetic field. He summed up all the laws of electricity and magnetism in the form of equations which are now referred to as Maxwell's equations. From these equations, he derived a wave equation, predicted the existence of electromagnetic waves, and showed that the speed of the electromagnetic waves in air should be about 3.107×10^8 m s^{-1}, which was very close to the measured value of the speed of light. The sole fact that the two values were very close to each other led Maxwell to propound his famous electromagnetic theory of light, according to which *light waves are electromagnetic waves.*
- In 1905, Einstein interpreted the photoelectric effect by putting forward his famous photon theory, according to which the energy in a light beam of frequency ν was concentrated in corpuscles of energy $h\nu$, where h represents Planck's constant.
- The consequence of wave-particle duality is the uncertainty principle of Heisenberg according to which *if the x coordinate of the position of a particle is known to an accuracy* Δx*, then the x component of the momentum cannot be determined to an accuracy better than* $\Delta p_x \approx h/\Delta x$*, where h is Planck's constant.*
- The classical corpuscular model is quite incapable of explaining the diffraction of light by a single slit. However, if we use the uncertainty principle in conjunction with the corpuscular model, the diffraction phenomenon can be explained.
- In Young's double-hole interference pattern, the corpuscular nature of the radiation is evident from its detection in the form of single photons and never a fraction of a photon. The appearance of the interference pattern is due to the fact that a photon interferes with itself. The quantum theory tells us that a photon partially passes through the two holes. This is not the splitting of the photon into two halves, but only implies that the photon is in a state which is a superposition of two states, one corresponding to the wave emanating from the first hole and the other to the wave emanating from the second hole. The superposed state will give rise to an intensity distribution similar to that obtained by considering the superposition of two waves.

Problems

2.1 An electron of energy 200 eV is passed through a circular hole of radius 10^{-4} cm. What is the uncertainty introduced in the momentum and also in the angle of emergence?

[**Ans:** $\Delta p \sim 5 \times 10^{-24}$ g cm s^{-1}; $\Delta\theta \approx 6 \times 10^{-6}$ rad]

2.2 In continuation of the previous problem, what would be the corresponding uncertainty for a 0.1 g lead ball thrown with a velocity 10^3 cm s^{-1} through a hole 1 cm in radius?

[**Ans:** $\Delta\theta \approx 5 \times 10^{-30}$ rad]

2.3 A photon of wavelength 6000 Å is passed through a slit of width 0.2 mm.

(a) Calculate the uncertainty introduced in the angle of emergence.

(b) The first minimum in the single-slit diffraction pattern occurs at $\sin^{-1}(\lambda/b)$, where b is the width of the slit. Calculate this angle and compare with the angle obtained in part (a).

[**Ans:** $\Delta\theta \approx 3 \times 10^{-3}$]

2.4 A 50 W bulb radiates light of wavelength 0.6 μm. Calculate the number of photons emitted per second.

[**Ans:** $\approx 1.5 \times 10^{20}$ photons/s]

2.5 Calculate the uncertainty in the momentum of a proton which is confined to a nucleus of radius equal to 10^{-13} cm. From this result, estimate the kinetic energy of the proton inside the nucleus and the strength of the nuclear interaction. What would be the kinetic energy for an electron, if it had to be confined within a similar nucleus?

2.6 The lifetime of the 2P state of the hydrogen atom is about 1.6×10^{-9} s. Use the time-energy uncertainty relation to calculate the frequency width $\Delta\nu$.

[**Ans:** $\approx 6 \times 10^8$ s^{-1}]

2.7 A 1 W laser beam (of diameter 2 cm) falls normally on two circular holes each of diameter 0.05 cm, as shown in Fig. 2.14. Calculate the average number of photons that will

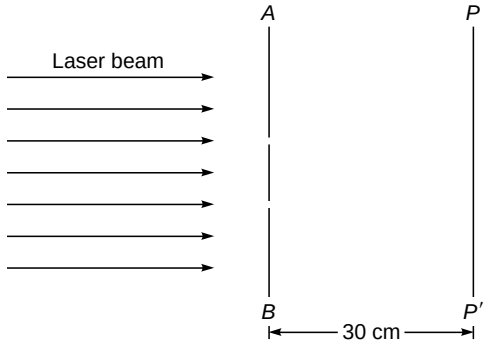


Fig. 2.14

be found between planes *AB* and *PP′*. Assume $\lambda = 6 \times 10^{-5}$ cm and the distance between planes *AB* and *PQ* is 30 cm.

[**Ans:** $\approx 4 \times 10^6$ photons]

Solutions

2.5 The proton is confined within a sphere of radius $r_0 \approx 10^{-13}$ cm. Thus the uncertainty in the momentum must be at least of the order of $\hbar/r_0$, or

$$p \sim \frac{\hbar}{r_0}$$

Therefore, the kinetic energy of the proton will be given by

$$E = \frac{p^2}{2m_p} \sim \frac{\hbar^2}{2m_p r_0^2}$$

where m_p is the mass of the proton. On substitution, we get

$$E \sim \frac{(1.05\times10^{-27}\ \text{erg s})^2}{2\times1.67\times10^{-24}\ \text{g}\times(10^{-13}\ \text{cm})^2}$$

$$\approx 3\times10^{-5}\ \text{erg} \approx 20\ \text{MeV}$$

Since the proton is bound inside the nucleus, the average of the potential energy $\langle V \rangle$ must be negative and greater in magnitude than the kinetic energy. Therefore

$$-\langle V \rangle \geq 20\ \text{MeV}$$

which indeed gives the correct order of the potential energy. The uncertainty in momentum for the electron is again ∇/r_0; however, since the rest mass of the electron is very much smaller than that of the proton, the velocity of the electron is very close to c and we have to use the extreme relativistic formula for the energy

$$E = cp = \frac{c\hbar}{r_0}$$

$$\approx \frac{(3\times10^{10})\,(1.05\times10^{-27})}{10^{-13}\times1.6\times10^{-6}}\,\text{MeV} \sim 200\,\text{MeV}$$

Although electrons do emerge from nuclei in β decay, they seldom have energies exceeding a few million electronvolts. Thus one does not expect the electron to be a basic constituent of the nucleus; the rare occasions when β decay occurs may be attributed to the transformation of a neutron into a proton and an electron (and the neutrino) so that the electron is in fact created at the instant the decay occurs.

REFERENCES AND SUGGESTED READINGS

1. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. III, Addison-Wesley Publishing Co., Reading, Mass., 1965.
2. Isaac Newton, *Opticks*, Dover Publications, 1952.
3. W. B. Joyce and A. Joyce, *"Descartes, Newton and Snell's law," J. Opt. Soc. America*, Vol. 66, p. 1, 1976.
4. A. W. Barton, *A Text Book on Light*, Longmans Green & Co., London, 1939.
5. C. Huygens, *Treatise on Light*, Dover Publications, 1962.
6. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. II, Addison-Wesley Publishing Co., Reading, Mass., 1965.
7. A. Einstein, *"On a Heuristic Point of View concerning the Production and Transformation of Light," Annalen der Physik*, Vol. 17, p. 132, 1905.
8. J. Stachel (ed.), *Einstein's Miraculous Year: Five Papers that changed the Face of Physics*, Princeton University Press, Princeton, 1998; reprinted by Shrishti Publishers, New Delhi.
9. W. H. Cropper, *The Quantum Physicists and an Introduction to Their Physics*, Oxford University Press, New York, 1970.
10. M. Jammer, *The Conceptual Development of Quantum Mechanics*, McGraw-Hill, New York, 1965.
11. R. Eisberg and R. Resnick, *Quantum Physics*, John Wiley & Sons, New York, 1974.
12. A. Ghatak and S. Lokanathan, *Quantum Mechanics,: Theory and Applications*, Macmillan India, New Delhi, 2005; reprinted by Kluwer Academic Publishers, Dordrecht, 2004.
13. G. Gamow, *Mr. Tompkins in Wonderland*, Cambridge University Press, Cambridge, 1940.
14. M. Born, *Atomic Physics*, Blackie & Son, London, 1962.
15. C. G. Shull, "Neutron Diffraction: A General Tool in Physics in Current Problems in Neutron Scattering," *Proceedings of the Symposium held at CNEN Casaccia Center in September 1968*, CNEN, Rome, 1970.
16. A. Rose, *"Quantum Effects in Human Vision," Advances in Biological and Medical Physics*, Vol. V, Academic Press, 1957.
17. Reprinted with permission from A. Tonomura, J. Endo, T. Matsuda, T. Kawasaki, and H. Ezawa, "Demonstration of Single-Electron Build up of an Interference Pattern" *American Journal of Physics*, Vol. 57, Issue 2, p. 117, 1989. Copyright © 1989, American Association of Physics Teachers.
18. P. A. M. Dirac, *The Principles of Quantum Mechanics*, Oxford University Press, Oxford, 1958.

PART 1
Geometrical Optics

This part (consisting of four chapters) is entirely based on geometrical optics and includes

- Ray tracing through graded-index media, explaining in detail the phenomena of mirage and looming, and also reflection from the ionosphere.
- Ray tracing through a system of lenses, leading to various concepts used in the design of optical instruments.
- A detailed description of the matrix method in paraxial optics, which is extensively used in the industry.
- A study of aberrations of optical systems.

Chapter Three

FERMAT'S PRINCIPLE AND ITS APPLICATIONS

Now in the further development of science, we want more than just a formula. First we have an observation, then we have numbers that we measure, then we have a law which summarizes all the numbers. But the real *glory* of science is that *we can find a way of thinking* such that the law is *evident*. The first way of thinking that made the law about the behavior of light evident was discovered by Fermat in about 1650, and it is called the *principle of least time*, or *Fermat's principle*.

—Richard Feynman, *Feynman Lectures on Physics*, Vol. I

Important Milestones

AD 140	*Greek physicist Claudius Ptolemy measured the angle of refraction in water for different angles of incidence in air and made a table of it.*
1621	*Although the above-mentioned numerical table was made in AD 140, it was only in 1621 that Willebrord Snel, a Dutch mathematician, discovered the law of refraction which is now known as Snell's law.*
1637	*Descartes derived Snell's law; his derivation assumed the corpuscular model of light.*
1657	*Pierre de Fermat enunciated his principle of "least time" and derived Snell's law of refraction and showed that if the velocity of light in the second medium is less, the ray will bend toward the normal, contrary to what is predicted by the corpuscular theory.*

3.1 INTRODUCTION

The study of the propagation of light in the realm of geometrical optics employs the concept of rays. To understand what a ray is, consider a circular aperture in front of a point source P as shown in Fig. 3.1. When the diameter of the aperture is quite large (~1 cm), then on the screen SS', one can see a patch of light with well-defined boundaries. When we start decreasing the size of the aperture, then at first the size of the patch starts decreasing; but when the size of the aperture becomes very small ($\lesssim 0.1$ mm), then the pattern obtained on SS' ceases to have well-defined boundaries. This phenomenon is known as *diffraction* and is a direct consequence of the finiteness of the wavelength (which is denoted by λ). In Chaps. 18 and 20 we will discuss the phenomenon of diffraction in great detail and will show that the diffraction effects become smaller with the decrease in wavelength. Indeed, in the limit of $\lambda \to 0$, the diffraction effects will be absent, and even for extremely small diameters of the aperture, we will obtain a well-defined shadow on the screen SS'; and therefore, in the zero wavelength limit one can obtain an infinitesimally thin pencil of light—this is called a *ray*. Thus, a ray defines the path of propagation of the energy in the limit of the wavelength going to zero. Since light has a wavelength of the order of 10^{-5} cm, which is small compared to the dimensions of normal optical instruments such as lenses and mirrors, one can, in many applications, neglect the finiteness of the wavelength. The field of optics under such an approximation (i.e., the neglect of the finiteness of the wavelength) is called *geometrical optics*.

The field of geometrical optics can be studied by using Fermat's principle which determines the path of the rays. According to this principle *the ray will correspond to that path for which the time taken is an extremum in comparison*

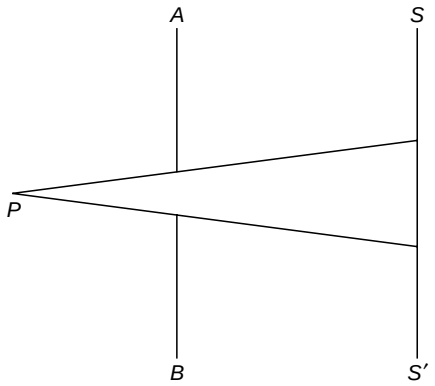


Fig. 3.1 The light emitted by the point source P is allowed to pass through a circular hole and if the diameter of the hole is very large compared to the wavelength of light then the light patch on the screen SS' has well-defined boundaries.

to nearby paths, i.e., it is either a minimum or a maximum or stationary[1]. Let $n(x, y, z)$ represent the position-dependent refractive index. Then

$$\frac{ds}{c/n} = \frac{n\,ds}{c}$$

will represent the time taken to traverse the geometric path ds in a medium of refractive index n. Here, c represents the speed of light in free space. Thus, if τ represents the total time taken by the ray to traverse the path AB along the curve C (see Fig. 3.2), then

$$\tau = \frac{1}{c}\sum_i n_i\ ds_i = \frac{1}{c}\int_{A\xrightarrow[C]{}B} n\,ds \qquad \textbf{(1)}$$

where ds_i represents the ith arc length and n_i the corresponding refractive index; the symbol $A\xrightarrow[C]{}B$ below the integral represents the fact that the integration is from point A to point B through curve C. Let τ' be the time taken along the nearby path $AC'B$ (shown as the dashed curve in Fig. 3.2), and if ACB indeed represents the path of a ray, then τ will be less than, greater than, or equal to τ' for *all* nearby paths like $AC'B$. Thus according to Fermat's principle, out of the many paths connecting the two points, the light ray would follow that path for which the time taken is an extremum. Since c is a constant, one can alternatively define a ray as the path for which

$$\int_{A\xrightarrow[C]{}B} n\,ds \qquad \textbf{(2)}$$

is an extremum[2]. The above integral represents the optical path from A to B along C; i.e., the ray would follow the path for which

$$\delta \int_{A\to B} n\,ds = 0 \qquad \textbf{(3)}$$

where the left-hand side represents the change in the value of the integral due to an infinitesimal variation of the ray path. According to the original statement of Fermat,

> The actual path between two points taken by a beam of light is the one which is traversed in the least time.

The above statement is incomplete and slightly incorrect. The correct form is:

> The actual ray path between two points is the one for which the optical path length is stationary with respect to variations of the path.

This is expressed by Eq. (3), and in this formulation, the ray paths may correspond to maxima, minima, or stationary.

From the above principle one can immediately see that in a homogeneous medium (i.e., in a medium whose refractive index is constant at each point), the rays will be straight lines because a straight line will correspond to a minimum value of the optical path connecting two points in the medium. Thus referring to Fig. 3.3, if A and B are two points in a homogeneous

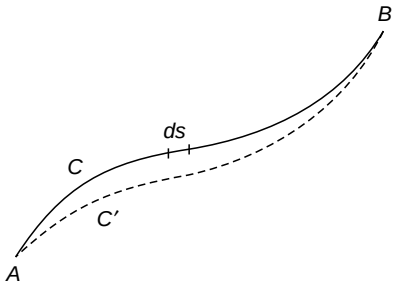


Fig. 3.2 If path ACB represents the actual ray path, then the time taken in traversing path ACB will be an extremum in comparison to any nearby path $AC'B$.

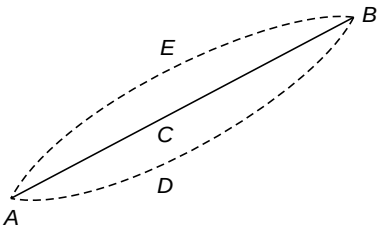


Fig. 3.3 Since the shortest distance between two points is along a straight line, light rays in a homogeneous medium are straight lines; all nearby paths like AEB or ADB will take longer times.

[1] The entire field of classical optics (both geometrical and physical) can be understood from Maxwell's equations, and of course, Fermat's principle can be derived from Maxwell's equations (see Refs. 1 and 2).

[2] A nice discussion on the extremum principle is given in Chap. 26 of Ref. 3.

medium, then the ray path will be along the straight line *ACB* because any nearby path such as *ADB* or *AEB* will correspond to a longer time.

3.2 LAWS OF REFLECTION AND REFRACTION FROM FERMAT'S PRINCIPLE

We will now obtain the laws of reflection and refraction from Fermat's principle. Consider a plane mirror *MN* as shown in Fig. 3.4. To obtain the laws of reflection, we have to determine the path from *A* to *B* (via the mirror) which has the minimum optical path length. Since the path would lie completely in a homogeneous medium, we need to minimize only the path length. Thus we have to find that path *APB* for which *AP* + *PB* is a minimum. To find the position of *P* on the mirror, we drop a perpendicular from *A* on the mirror and let *A′* be a point on the perpendicular such that *AR* = *RA′*; thus *AP* = *PA′* and *AQ* = *A′Q*, where *AQB* is another path adjacent to *APB*. Thus we have to minimize the length *A′PB*. Clearly, for *A′PB* to be a minimum, *P* must be on the straight line *A′B*. Thus points *A*, *A′*, *P*, and *B* will be in the same plane, and if we draw a normal *PS* at *P*, then this normal will also lie in the same plane. Simple geometric considerations show that

$$\angle APS = \angle SPB$$

Thus for minimum optical path length, the angle of incidence *i* (= ∠*APS*) and the angle of reflection *r* (= ∠*SPB*) must be equal; and the incident ray, the reflected ray, and the normal to the surface at the point of incidence on the mirror must be in the same plane. These form the laws of reflection. Actually, in the presence of the mirror there will be two ray paths which will connect points *A* and *B*; the two paths will be *AB* and *APB*. Fermat's principle tells us that whenever the optical path length is an extremum, we will have a ray, and thus, in general, there may be more than one ray path connecting two points.

To obtain the laws of refraction, let *PQ* be a surface separating two media of refractive indices n_1 and n_2, as shown in Fig. 3.5. Let a ray starting from point *A* intersect the interface at *R* and proceed to *B* along *RB*. Clearly, for minimum optical path length, the incident ray, the refracted ray, and the normal to the interface must all lie in the same plane. To determine that point *R* for which the optical path length from *A* to *B* is a minimum, we drop perpendiculars *AM* and *BN* from *A* and *B*, respectively, on the interface *PQ*. Let $AM = h_1$, $BN = h_2$, and $MR = x$. Then since *A* and *B* are fixed, $RN = L - x$, where $MN = L$ is a fixed quantity. The optical path length from *A* to *B*, by definition, is

$$L_{op} = n_1 AR + n_2 RB$$
$$= n_1 \sqrt{x^2 + h_1^2} + n_2 \sqrt{(L-x)^2 + h_2^2} \qquad \textbf{(4)}$$

To minimize this, we must have

$$\frac{dL_{op}}{dx} = 0$$

i.e.,
$$\frac{n_1 x}{\sqrt{x^2 + h^2}} - \frac{n_2 (L-x)}{\sqrt{(L-x)^2 + h^2}} = 0 \qquad \textbf{(5)}$$

Further, as can be seen from Fig. 3.5,

$$\sin\theta_1 = \frac{x}{\sqrt{x^2 + h_1^2}}$$

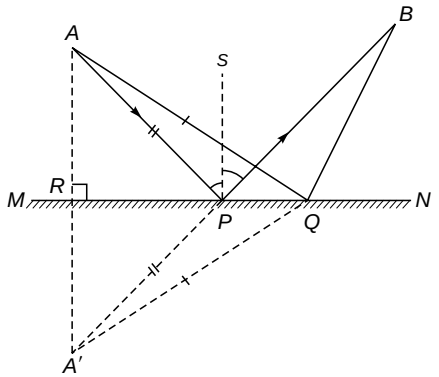


Fig. 3.4 The shortest path connecting two points *A* and *B* via the mirror is along path *APB,* where point *P* is such that *AP*, *PS,* and *PB* are in the same plane and ∠*APS* = ∠*SPB*, with *PS* being normal to the plane of the mirror. The straight-line path *AB* is *also* a ray.

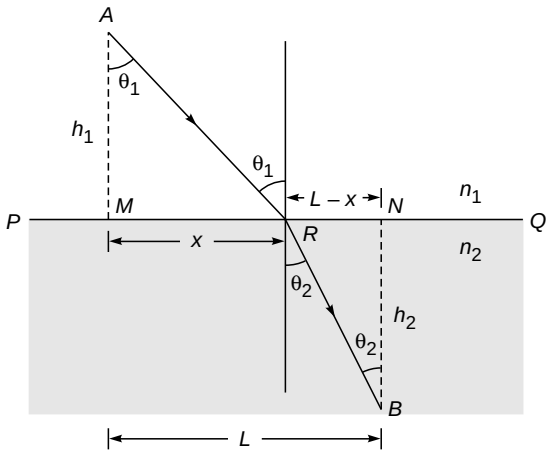


Fig. 3.5 Points *A* and *B* are two points in the media of refractive indices n_1 and n_2. The ray path connecting *A* and *B* will be such that $n_1 \sin\theta_1 = n_2 \sin\theta_2$.

and $$\sin\theta_2 = \frac{L-x}{\sqrt{(L-x)^2 + h_2^2}}$$

Thus Eq. (5) becomes

$$n_1 \sin\theta_1 = n_1 \sin\theta_2 \qquad \textbf{(6)}$$

which is Snell's law of refraction.

The laws of reflection and refraction form the basic laws for tracing light rays through simple optical systems, such as a system of lenses and mirrors.

Example 3.1 Consider a set of rays, parallel to the axis, incident on a paraboloidal reflector (see Fig. 3.6). Show, by using Fermat's principle, that all the rays will pass through the focus of the paraboloid; a paraboloid is obtained by rotating a parabola about its axis. This is the reason why a paraboloidal reflector is used to focus parallel rays from a distant source, as in radio astronomy (see Figs. 3.7 and 3.8).

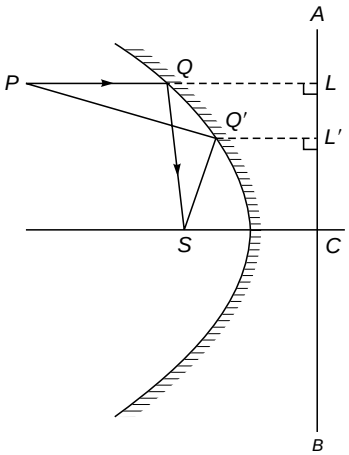


Fig. 3.6 All rays parallel to the axis of a paraboloidal reflector pass through the focus after reflection (the line *ACB* is the directrix). It is for this reason that antennas (for collecting electromagnetic waves) or solar collectors are often paraboloidal in shape.

Solution: Consider a ray *PQ*, parallel to the axis of the parabola, incident at the point *Q* (see Fig. 3.6). To find the reflected ray, one has to draw a normal at the point *Q* and then draw the reflected ray. It can be shown from geometric considerations that the reflected ray *QS* will always pass through the focus *S*. However, this procedure will be quite cumbersome, and as we will show, the use of Fermat's principle leads us to the desired results immediately.

To use Fermat's principle, we try to find out the ray connecting the focus *S* and an arbitrary point *P* (see Fig. 3.6). Let the ray path be *PQ′S*. According to Fermat's principle, the ray path will correspond to a minimum value of $PQ' + Q'S$. From the point Q' we drop a perpendicular $Q'L'$ on the directrix *AB*. From the definition of the parabola it follows that $Q'L' = Q'S$. Thus

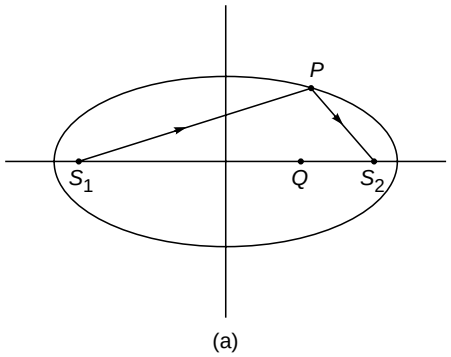
$$PQ' + Q'S = PQ' + Q'L'$$

Fig. 3.7 A paraboloidal satellite dish. A color photograph appears in the insert at the back of the book.

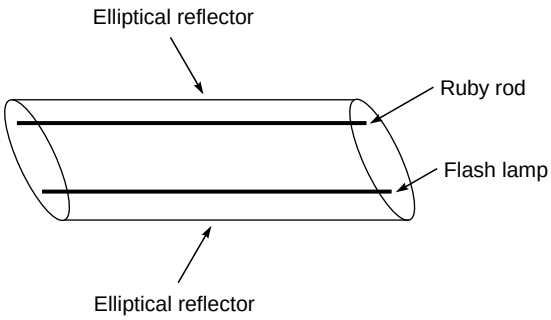
Fig. 3.8 Fully steerable 45 m paraboloidal dishes of the Giant Metrewave Radio Telescope (GMRT) in Pune, India. The GMRT consists of 30 dishes of 45 m diameter with 14 antennas in the central array. Photograph courtesy: Professor Govind Swarup, GMRT, Pune. A color photograph appears in the insert at the back of the book.

Let *L* be the foot of the perpendicular drawn from point *P* on *AB*. Then for $PQ' + Q'L'$ to be a minimum, point *Q* should lie on the straight line *PQL*; and thus the actual ray which connects the points *P* and *S* will be $PQ + QS$, where *PQ* is parallel to the axis. Therefore, all rays parallel to the axis will pass through *S*, and conversely, all rays emanating from point *S* will become parallel to the axis after suffering a reflection.

Example 3.2 Consider an elliptical reflector whose foci are the points S_1 and S_2 (see Fig. 3.9). Show that all rays emanating from point S_1 will pass through point S_2 after undergoing a reflection.



(a)



(b)

Fig. 3.9 Fig. 3.9 (a) All rays emanating from one of the foci of an ellipsoidal reflector will pass through the other focus. (b) In a ruby laser one may have a configuration in which the laser rod and the flash lamp coincide with the focal lines of a cylindrical reflector with elliptical cross section.

Solution: Consider an arbitrary point P on the ellipse (see Fig. 3.9). It is well known that $S_1P + S_2P$ is a constant, and therefore, all rays emanating from point S_1 will pass through S_2. Notice that here we have an example where the time taken by the ray is stationary; i.e., it is neither a maximum nor a minimum but has a constant value for all points lying on the mirror. As a corollary, we may note the following two points:

1. Excepting the rays along the axis, no other ray (emanating from either of the foci) will pass through an arbitrary point Q which lies on the axis.
2. The above considerations will remain valid even for an ellipsoid of revolution obtained by rotating the ellipse about its major axis.

Because of the above-mentioned property of elliptical reflectors, they are often used in laser systems. For example, in a ruby laser (see Chap. 26) one may have a configuration in which the laser rod and the flash lamp coincide with the focal lines of a cylindrical reflector of elliptical cross section [see Fig. 3.9(b)]; such a configuration leads to an efficient transfer of energy from the lamp to the ruby rod.

Example 3.3 Consider a spherical refracting surface SPM separating two media of refractive indices n_1 and n_2 (see Fig. 3.10). Point C represents the center of the spherical surface SPM.

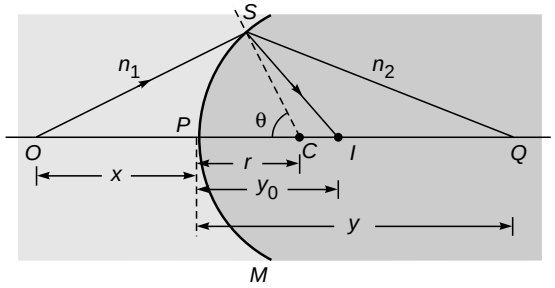


Fig. 3.10 SPM is a spherical refracting surface separating two media of refractive indices n_1 and n_2. And C represents the center of the spherical surface.

Consider two points O and Q such that the points O, C, and Q are in a straight line. Calculate the optical path length OSQ in terms of the distances x, y, r and the angle θ (see Fig. 3.10). Use Fermat's principle to find the ray connecting the two points O and Q. Also, assuming the angle θ to be small, determine the paraxial image of the point O.

[*Note*: We reserve the symbol R to represent the radius of curvature of a spherical surface which will be positive (or negative) depending upon whether the center of curvature lies on the right (or left) of the point P. The quantity r represents the magnitude of the radius of curvature which, for Fig. 3.10, happens to be R. Similarly, the quantities x and y are the magnitudes of the distances; the sign convention is discussed later in this problem.]

Solution: From the triangle SOC we have

$$OS = [(x + r)^2 + r^2 - 2(x + r)r \cos\theta]^{1/2}$$

$$\approx \left[x^2 + 2rx + 2r^2 - 2(xr + r^2)\left(1 - \frac{\theta^2}{2}\right)\right]^{1/2}$$

$$\approx x\left(1 + \frac{rx + r^2}{x^2}\theta^2\right)^{1/2} \approx x + \frac{1}{2}r^2\left(\frac{1}{r} + \frac{1}{x}\right)\theta^2$$

where we have assumed θ (measured in radians) to be small so that we may use the expression

$$\cos\theta \approx 1 - \frac{\theta^2}{2}$$

and also make a binomial expansion. Similarly, by considering the triangle SCQ, we would have

$$SQ \approx y - \frac{1}{2}r^2\left(\frac{1}{r} - \frac{1}{y}\right)\theta^2$$

Thus the optical path length OSQ is given by

$$L_{op} = n_1 OS + n_2 SQ$$

$$\approx (n_1 x + n_2 y) + \frac{1}{2}r^2\left(\frac{n_1}{x} + \frac{n_2}{y} - \frac{n_2 - n_1}{r}\right)\theta^2 \qquad \textbf{(7)}$$

For the optical path to be an extremum, we must have

$$\frac{dL_{op}}{d\theta} = 0 = r^2\left(\frac{n_1}{x}+\frac{n_2}{y}-\frac{n_2-n_1}{r}\right)\theta \quad \textbf{(8)}$$

Thus, unless the quantity inside the parentheses is zero, we must have $\theta = 0$, implying that the *only* ray connecting points *O* and *Q* will be the straight-line path *OPQ*, which also follows from Snell's law because the ray *OP* hits the spherical surface normally and should proceed undeviated.

On the other hand, if the value of *y* was such that the quantity inside the parentheses were zero, i.e., if *y* were equal to y_0 such that

$$\frac{n_2}{y_0}+\frac{n_1}{x} = \frac{n_2-n_1}{r} \quad \textbf{(9)}$$

then $dL_{op}/d\theta$ would vanish for *all* values of θ; of course, Now θ is assumed to be small—which is the paraxial approximation. Now, if point *I* corresponds to $PI = y_0$ (see Fig. 3.10), then *all* paths such as *OSI* are allowed ray paths, implying that *all* (paraxial) rays emanating from *O* will pass through *I* and *I* will therefore represent the paraxial image point. Obviously, all rays like *OSI* (which start from *O* and pass through *I*) take the *same* amount of time in reaching point *I*.

Equation (9) is a particular form of the equation determining the paraxial image point

$$\frac{n_2}{v}-\frac{n_1}{u} = \frac{n_2-n_1}{R} \quad \textbf{(10)}$$

with the sign convention that all distances measured to the right of point *P* are positive and those to its left negative. Thus $u = -x$, $v = +y$, and $r = +R$.

To determine whether the ray path *OPQ* corresponds to minimum time or maximum time or stationary, we must determine the sign of $d^2L_{op}/d\theta^2$ which is given by

$$\frac{d^2L_{op}}{d\theta^2} = r^2\left(\frac{n_1}{x}+\frac{n_2}{y}-\frac{n_2-n_1}{r}\right)$$

$$= r^2 n_2\left(\frac{1}{y}-\frac{1}{y_0}\right)$$

Obviously, if $y > y_0$ (i.e., point *Q* is on the right of the paraxial image point *I*), then $d^2L_{op}/d\theta^2$ is negative and the ray path *OPQ* corresponds to *maximum* time in comparison with nearby paths and conversely. On the other hand, if $y = y_0$, then $d^2L_{op}/d\theta^2$ will vanish, implying that the extremum corresponds to stationarity. Thus, in the paraxial approximation, all rays emanating from point *O* will take the same amount of time to reach point *I*.

Alternatively, one can argue that if *I* is the paraxial image point of *P*, then

$$n_1OP + n_2PI = n_1OS + n_2SI$$

Thus, when *Q* lies on the right of point *I*, we have

$$\begin{aligned} n_1OP + n_2PQ &= n_1OS + n_2(SI - PI + PQ) \\ &= n_1OS + n_2(SI + IQ) \\ &> n_1OS + n_2SQ \end{aligned}$$

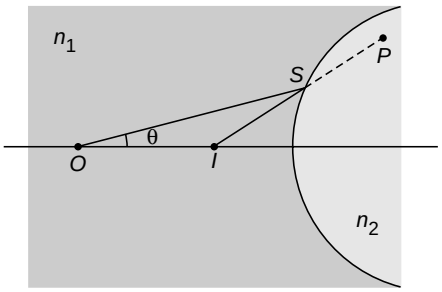


Fig. 3.11 The refracted ray is assumed to diverge away from the principal axis.

implying that the ray path *OPQ* corresponds to a maximum. Similarly, when *Q* lies on the left of point *I*, then the ray path *OPQ* corresponds to a minimum; and when *Q* coincides with *I*, we have the stationarity condition.

Example 3.4 We again consider refraction at a spherical surface; however, the refracted ray is assumed to diverge away from the principal axis (see Fig. 3.11). Let us consider paraxial rays, and let *I* be a point (on the axis) such that $n_1OS - n_2SI$ is independent of point *S*. Thus, for paraxial rays, the quantity

$$n_1OS - n_2SI \text{ is independent of } \theta \quad \textbf{(11)}$$

and is an extremum. Let *P* be an arbitrary point in the second medium, and we wish to find the ray path connecting points *O* and *P*. For *OSP* to be an allowed ray path

$L_{op} = n_1OS + n_2SP$ should be an extremum

or

$L_{op} = (n_1OS - n_2SI) + n_2(IS + SP)$ should be an extremum

where we have added and subtracted n_2SI. Now, the point *I* is such that the first quantity is already an extremum; thus, the quantity $SP + SI$ should be an extremum, and therefore it should be a straight line. Thus the refracted ray must appear to come from the point *I*. We may therefore say that for a virtual image we must make the quantity

$$n_1OS - n_2SI \quad \textbf{(12)}$$

an extremum.

3.3 RAY PATHS IN AN INHOMOGENEOUS MEDIUM

In an inhomogeneous medium, the refractive index varies in a continuous manner and, in general, the ray paths are curved. For example, on a hot day, the air near the ground has a higher temperature than the air which is much above the surface. Since the density of air decreases with increase in temperature, the refractive index increases continuously as we go above the ground. This leads to the phenomenon known as *mirage*. We will use Snell's law (or Fermat's principle) to determine the ray paths in an inhomogeneous

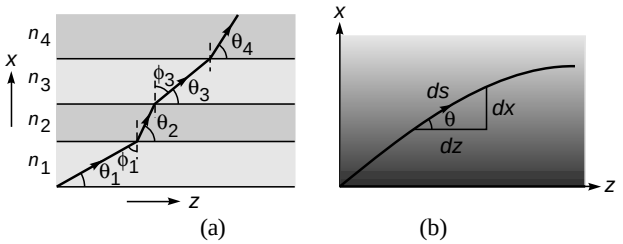


Fig. 3.12 (a) In a layered structure, the ray bends in such a way that the product $n_i \cos \theta_i$ remains constant. (b) For a medium with a continuously varying refractive index, the ray path bends in such a way that the product $n(x) \cos \theta(x)$ remains constant.

medium. We will restrict ourselves to the special case when the refractive index changes continuously along one direction only; we assume this direction to be along the x axis.

The inhomogeneous medium can be thought of as a limiting case of a medium consisting of a continuous set of thin slices of media of different refractive indices; see Fig. 3.12(a). At each interface, the light ray satisfies Snell's law, and one obtains [see Fig. 3.12(a)]

$$n_1 \sin \phi_1 = n_2 \sin \phi_2 = n_3 \sin \phi_3 = \cdots \quad \textbf{(13)}$$

Thus, the product

$$n(x) \cos \theta(x) = n(x) \sin \phi(x) \quad \textbf{(14)}$$

is an invariant of the ray path; we will denote this invariant by $\tilde{\beta}$. The value of this invariant may be determined from the fact that if the ray initially makes an angle θ_1 (with the z axis) at a point where the refractive index is n_1, then the value of $\tilde{\beta}$ is $n_1 \cos \theta_1$. Thus, in the limiting case of a continuous variation of refractive index, the piecewise straight lines shown in Fig. 3.12(a) form a continuous curve which is determined from the equation

$$n(x) \cos \theta(x) = n_1 \cos \theta_1 = \tilde{\beta} \quad \textbf{(15)}$$

implying that as the refractive index changes, the ray path bends in such a way that the product $n(x) \cos \theta(x)$ remains constant [see Fig. 3.12(b)]. Equation (15) can be used to derive the ray equation (see Sec. 3.4).

3.3.1 The Phenomenon of Mirage[3]

We are now in a position to qualitatively discuss the formation of a mirage. As mentioned earlier, on a hot day the refractive index continuously decreases as we go near the ground. Indeed, the refractive index variation can be approximately assumed to be of the form

$$n(x) \approx n_0 + kx \quad 0 < x < \text{few meters} \quad \textbf{(16)}$$

[3] For more details, see Refs. 4–8.

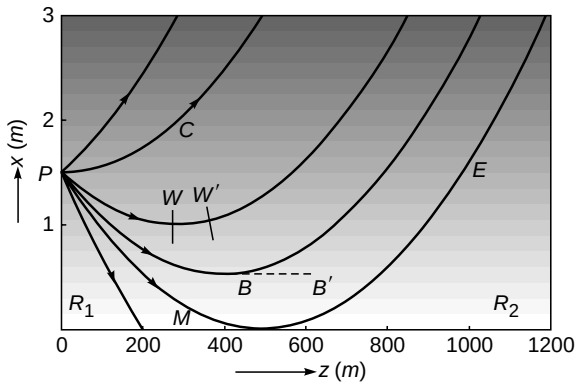


Fig. 3.13 Ray paths in a medium characterized by a linear variation of refractive index [see Eq. (16)] with $k \approx 1.234 \times 10^{-5}$ m^{-1}. The object point is at a height of 1.5 m, and the curves correspond to +0.2°, 0°, –0.2°, –0.28°, –0.3486°, and –0.5° The shading shows that the refractive index increases with x.

where n_0 is the refractive index of air at $x = 0$ (i.e., just above the ground) and k is a constant. The exact ray paths (see Example 3.8) are shown in Fig. 3.13.

We consider a ray which becomes horizontal at $x = 0$. At the eye position E ($x = x_e$), if the refractive index is n_e, and if at that point the ray makes an angle θ_e with the horizontal, then

$$\tilde{\beta} = n_0 = n_e \cos \theta_e \quad \textbf{(17)}$$

Usually $\theta_e << 1$ so that

$$\frac{n_0}{n_e} = \cos \theta_e \approx 1 - \frac{1}{2}\theta_e^2$$

$$\Rightarrow \quad \theta_e \approx \sqrt{2\left(1 - \frac{n_0}{n_e}\right)} \quad \textbf{(18)}$$

At constant air pressure

$$(n_0 - 1)T_0 \approx (n_e - 1)T_e \quad \textbf{(19)}$$

From Eq. (19) we get

$$1 - \frac{n_0 - 1}{n_e - 1} = 1 - \frac{T_e}{T_0}$$

or

$$\frac{n_e - n_0}{n_e} = \frac{n_e - 1}{n_e}\left(1 - \frac{T_e}{T_0}\right)$$

so that

$$\theta_e \approx \sqrt{2\left(1 - \frac{1}{n_e}\right)\left(1 - \frac{T_e}{T_0}\right)} \quad \textbf{(20)}$$

On a typical hot day the temperature near road surface $T_0 \approx 323$ K (= 50°C), and, about 1.5 m above the ground, $T_e \approx 303$ K (= 30°C). Now, at 30°C, $n_e \approx 1.00026$, giving $\theta_e \approx 5.67 \times 10^{-3}$ rad $\approx 0.325°$. In Fig. 3.13 we have shown rays emanating (at different angles) from a point P which is 1.5 m above the ground; thus each ray has a specified value of the invariant $\tilde{\beta}$ (= $n_1 \cos \theta_1$). The figure shows that when the object point P and the observation point E are close to the ground, the only ray path connecting points P and E will be along the curve PME, and that a ray emanating horizontally from the point P will propagate in the upward direction as PC as shown in figure. Thus, in such a condition, the eye at E will see the mirage and *not see* the object directly at P. We also find that there is a region R_2 where none of the rays (emanating from the point P) reaches; thus, an eye in this region cannot see neither the object or its image. This is therefore called the *shadow region*. Furthermore, there is also a region R_1 where only the object is directly visible and the virtual image is not seen.

The *bending up* of the ray after it becomes parallel to the z axis cannot be directly inferred from Eq. (15) because at such a point, $\theta = 0$ and one may expect the ray to proceed horizontally beyond the turning point, as shown by a dotted line in Fig. 3.13; the point at which $\theta = 0$ is known as the *turning point*. However, from considerations of symmetry and from the reversibility of ray paths, it immediately follows that the ray path should be symmetric about the turning point and hence bend up. Physically, the bending of the ray can be understood by considering a small portion of a wave front such as W (see Fig. 3.13); the upper edge will travel with a smaller speed in comparison with the lower edge, and this will cause the wave front to tilt (see W'), causing the ray to bend. Furthermore, a straight-line path such as BB' does not correspond to an extremum value of the optical path.

We next consider a refractive index variation which saturates to a constant value as $x \to \infty$:

$$n^2(x) = n_0^2 + n_2^2(1 - e^{-\alpha x}) \quad x > 0 \qquad \textbf{(21)}$$

where n_0, n_2, and α are constants and once again x represents the height above the ground. The refractive index at $x = 0$ is n_0, and for large values of x, it approaches $(n_0^2 + n_2^2)^{1/2}$. The exact ray paths are obtained by solving the ray equation (see Example 3.10) and are shown in Figs 3.14 and 3.15; they correspond to the following values of various parameters:

$$n_0 = 1.000233 \qquad n_2 = 0.45836 \qquad \alpha = 2.303\ \text{m}^{-1} \qquad \textbf{(22)}$$

The actual values of the refractive index for parameters given by the above equation are not very realistic—nevertheless, it allows us to understand qualitatively the ray paths in a graded index medium. Figures 3.14 and 3.15 show the ray paths emanating from points that are 0.43 and 2.8 m above the ground, respectively. In Fig. 3.14, the point P corresponds to a value of the refractive index equal to 1.06455 (= n_1), and different rays correspond to different values of θ_1, the angle that the ray makes with the z axis at the point P. From Fig. 3.14 we again see that when the object point P and the observation point E are close to the ground, the only ray path connecting points P and E will be along the curve PME, and that a ray emanating horizontally from the point P will propagate in the upward direction, shown as PC in Fig. 3.14.

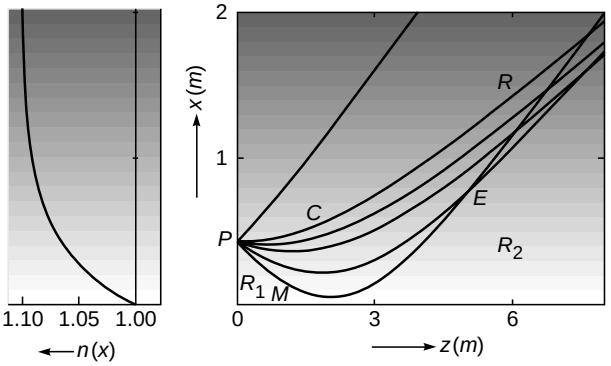


Fig. 3.14 Ray paths in a medium characterized by Eqs. (21) and (22). The object point is at a height of $1/\alpha$ (≈0.43 m), and the curves correspond to θ_1 (the initial launch angle) = $+\pi/10$, 0, $-\pi/60$, $-\pi/30$, $-\pi/15$, and $-\pi/10$. The shading shows that the refractive index increases with x.

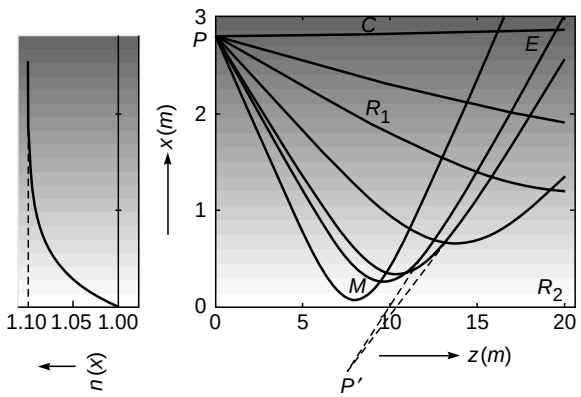


Fig. 3.15 Ray paths in a medium characterized by Eqs. (21) and (22). The object point is at a height of 2.8 m, and the curves correspond to θ_1 (the initial launch angle) = 0, $-\pi/60$, $-\pi/30$, $-\pi/16$, $-\pi/11$, $-\pi/10$, and $-\pi/8$. The shading shows that the refractive index increases with x.

Thus, in such a condition, the eye at E will see the mirage and *not see* the object directly at P. However, if points P and E are much above the ground (see Fig. 3.15), the eye will see the object almost directly (because of rays like PCE) and will also receive rays appearing to emanate from such points as P'. It may be readily seen that different rays do not appear to come from the same point, and hence the reflected image seen will have considerable aberrations. Once again, there is a *shadow region* R_2 where none of the rays (emanating from the point P) reaches there; thus, an eye in this region cannot see either the object or its image. The actual formation of mirage is shown in Figs. 3.16 and 3.17.

Fig. 3.16 A typical mirage as seen on a hot road on a warm day. The photograph was taken by Professor Piotr Pieranski of Poznan University of Technology in Poland; used with permission from Professor Pieranski. A color photograph appears in the insert at the back of the book.

Fig. 3.17 This is actually *not* a reflection in the ocean, but the miraged (inverted) image of the Sun's lower edge. A few seconds later (notice the motion of the bird to the left of the Sun!), the reflection fuses with the erect image. The photographs were taken by Dr. George Kaplan of the U.S. Naval Observatory and are on the website http://mintaka.sdsu.edu/GF/explain/simulations/infmir/Kaplan_photos.html created by Dr. A Young; photographs used with permissions from Dr. Kaplan and Dr. A Young. Color photographs appear in the insert at the back of the book.

Example 3.5 As an example, for an object shown in Fig. 3.14, let us calculate the angle at which the ray should be launched so that it becomes horizontal at $x = 0.2$ m. Now,

$$\text{at } x = 0.2 \text{ m}, \qquad n(x) = 1.03827$$

Thus, if θ_1 represents the angle that the ray makes with the z axis at the point P (see Fig. 3.14), then

$$n_1 \cos \theta_1 = 1.03827 \times \cos 0$$

implying

$$\theta_1 \approx 13°$$

Further, for the ray which becomes horizontal at $x = 0.2$ m the value of the invariant is given by

$$\tilde{\beta} \approx 1.03827$$

Example 3.6 In Fig. 3.15, the object point corresponds to $x = 2.8$ m where $n(x) \approx 1.1$. Thus for a ray launched with $\theta_1 = -\pi/8$,

$$\tilde{\beta} = 1.1 \cos \theta_1 = 1.01627$$

Thus if the ray becomes horizontal at $x = x_2$, then

$$n(x_2) = \tilde{\beta} = 1.01627$$

and

$$x_2 = -\frac{1}{\alpha} \ln \left[1 - \frac{n^2(x_2) - n_0^2}{n_2^2} \right]$$
$$\approx 0.073 \text{ m}$$

3.3.2 The Phenomenon of Looming

The formation of mirage discussed above occurs due to an increase in the refractive index of air above the hot surface. On the other hand, above cold seawater, the air near the water surface is colder than the air above it, and hence there is an opposite temperature gradient. A suitable refractive index variation for such a case can be written as

$$n^2(x) = n_0^2 + n_2^2 e^{-\alpha x} \qquad \textbf{(23)}$$

The equation describing the ray path is discussed in Prob. 3.13. We assume the values of n_0, n_2, and α to be given by Eq. (22). For an object point P at a height of 0.5 m, the ray paths are shown in Fig. 3.18. If the eye is at E, then it will receive rays appearing to emanate from P'. Such a phenomenon in which the object appears to be above its actual position is known as *looming*; it is commonly observed in viewing ships over cold seawaters (see Figs. 3.19 and 3.20). Moreover, since no other rays emanating from P reach A, the object cannot be observed directly.

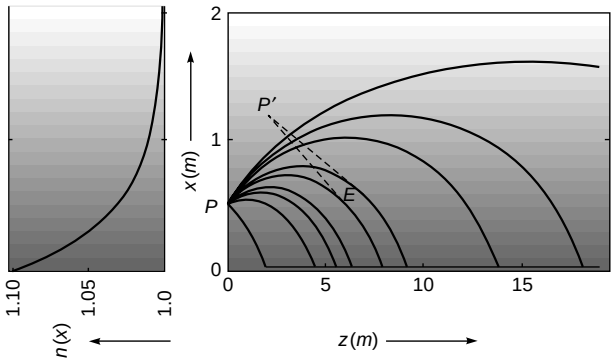


Fig. 3.18 Ray paths corresponding to the refrective index distribution given by Eq. (23) for an object at a height of 0.5 m; the values of n_0, n_2, and α are given by Eq. (22).

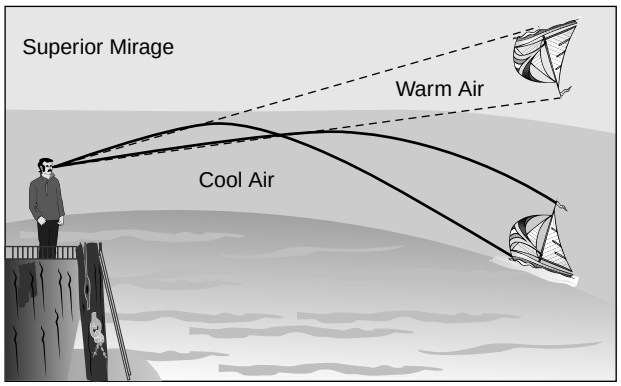


Fig. 3.19 If we are looking at the ocean on a cold day, then the air near the surface of the water is cold and gets warmer as we go up. Thus, as we go up, the refractive index decreases continuously; and because of curved ray paths, one will observe an inverted image of the ship as shown in the figure above. A color photograph appears in the insert at the back of the book.

Fig. 3.20 A house in the archipelago with a superior mirage. Figure adapted from http:/virtual.finland.fi/netcomm/news showarticle.asp?intNWSAID=25722. The photograph was taken by Dr. Pekka Parviainen in Turku, Finland; used with permission from Dr. Parviainen. A color photograph appears in the insert at the back of the book.

Fig. 3.21 The noncircular shape of the Sun at sunset. A color photograph appears in the insert at the back of the book.

3.3.3 The Graded Index Atmosphere

One of the interesting phenomena associated with imaging in a graded index medium is the noncircular shape of the setting or the rising Sun (see Fig. 3.21). This can easily be understood in the following manner. The refractive index of the air gradually decreases as we move outward. If we approximate the continuous refractive index gradient by a finite number of layers (each layer having a specific refractive index), then the ray will bend in a way similar to that shown in Fig. 3.22. Thus the Sun (which is actually at *S*) appears to be in the direction of *S*′. It is for this reason that the setting Sun appears flattened and

Fig. 3.22 The atmosphere is a graded index medium, and because of refraction, light from *S* appears to come from *S*9.

also leads to the fact that the days are usually about 5 min longer than they would have been in the absence of the atmosphere. Obviously, if we were on the surface of the Moon, the rising or the setting Sun would look not only white but also circular!

3.4 THE RAY EQUATION AND ITS SOLUTIONS

In this section, we will derive the ray equation, the solution of which will give the precise ray paths in an inhomogeneous medium. We will restrict ourselves to the special case when the refractive index changes continuously along only one direction, which we assume to be along the x axis. This medium can be thought of as the limiting case of a medium comprising a continuous set of thin slices of media of different refractive indices. As discussed earlier, for a continuously varying refractive index, the product $n(x)\cos\theta(x)$ is an invariant of the ray path which we denote by $\tilde{\beta}$:

$$n(x)\cos\theta(x) = \tilde{\beta} \tag{24}$$

Furthermore, for a continuous variation of refractive index, the piecewise straight lines shown in Fig. 3.12(a) form a continuous curve as in Fig. 3.12(b). If ds represents the infinitesimal arc length along the curve, then

$$(ds)^2 = (dx)^2 + (dz)^2$$

or
$$\left(\frac{ds}{dz}\right)^2 = \left(\frac{dx}{dz}\right)^2 + 1 \tag{25}$$

Now, if we refer to Fig. 3.12(b), we find that

$$\frac{dz}{ds} = \cos\theta = \frac{\tilde{\beta}}{n(x)} \tag{26}$$

Thus Eq. (25) becomes

$$\left(\frac{dx}{dz}\right)^2 = \frac{n^2(x)}{\tilde{\beta}^2} - 1 \tag{27}$$

For a given $n(x)$ variation, Eq. (27) can be integrated to give the ray path $x(z)$; however, it is often more convenient to put Eq. (27) in a slightly different form by differentiating it with respect to z:

$$2\frac{dx}{dz}\frac{d^2x}{dz^2} = \frac{1}{\tilde{\beta}^2}\frac{dn^2}{dx}\frac{dx}{dz}$$

or
$$\frac{d^2x}{dz^2} = \frac{1}{2\tilde{\beta}^2}\frac{dn^2}{dx} \tag{28}$$

Both Eqs. (27) and (28) represent rigorously correct ray equations when the refractive index depends only on the x coordinate.

Example 3.7 As a simple application of Eq. (28), let us consider a homogeneous medium for which $n(x)$ is a constant. In such a case, the right-hand side of Eq. (28) is zero and one obtains

$$\frac{d^2x}{dz^2} = 0$$

Integrating the above equation twice with respect to z, we obtain

$$x = Az + B$$

which is the equation of a straight line, as it ought to be in a homogeneous medium.

Example 3.8 We next consider the ray paths in a medium characterized by the refractive index variation

$$n(x) = n_0 + kx \tag{29}$$

For the above profile, the ray equation [Eq. (28)] takes the form

$$\frac{d^2x}{dz^2} = \frac{1}{2\tilde{\beta}^2}\frac{dn^2}{dx} = \frac{k}{\tilde{\beta}^2}(n_0 + kx)$$

or
$$\frac{d^2X}{dz^2} = \kappa^2 X(z) \tag{30}$$

where
$$X \equiv x + \frac{n_0}{k} \quad \text{and} \quad \kappa = \frac{k}{\tilde{\beta}} \tag{31}$$

Thus the ray path is given by

$$x(z) = -\frac{n_0}{k} + C_1 e^{\kappa z} + C_2 e^{-\kappa z} \tag{32}$$

where the constants C_1 and C_2 are to be determined from initial conditions. We assume that at $z = 0$, the ray is launched at $x = x_1$, making an angle θ_1 with the z axis; thus

$$x(z = 0) = x_1$$

and
$$\left.\frac{dx}{dz}\right|_{z=0} = \tan\theta_1$$

Elementary manipulations would give us

$$C_1 = \frac{1}{2}\left[x_1 + \frac{1}{k}(n_0 + n_1\sin\theta_1)\right] \tag{33}$$

and

$$C_2 = \frac{1}{2}\left[x_1 + \frac{1}{k}(n_0 - n_1\sin\theta_1)\right] \tag{34}$$

where $n_1 = n_0 + kx_1$ represents the refractive index at $x = x_1$ and we have used the fact that

$$\tilde{\beta} = n_1\cos\theta_1 \tag{35}$$

Figure 3.13 shows the ray paths as given by Eq. (32) with $x_1 = 1.5$ m, $n_1 = 1.00026$, and $k \approx 1.234 \times 10^{-5}$ m^{-1}.

3.4.1 Ray Paths in Parabolic Index Media

We consider a parabolic index medium characterized by the following refractive index distribution:

$$n^2(x) = n_1^2 - \gamma^2 x^2 \tag{36}$$

We will use Eq. (27) to determine the ray paths. Equation (27) can be written as

$$\int \frac{dx}{\sqrt{n^2(x) - \tilde{\beta}^2}} = \pm \frac{1}{\tilde{\beta}} \int dz \tag{37}$$

Substituting for $n^2(x)$, we get

$$\int \frac{dx}{\sqrt{x_0^2 - x^2}} = \pm \Gamma \int dz \tag{38}$$

where

$$x_0 = \frac{1}{\gamma} \sqrt{n_1^2 - \tilde{\beta}^2} \tag{39}$$

and

$$\Gamma = \frac{\gamma}{\tilde{\beta}} \tag{40}$$

Writing $x = x_0 \sin\theta$ and carrying out the straightforward integration, we get

$$x = \pm x_0 \sin[\Gamma(z - z_0)] \tag{41}$$

We can always choose the origin such that $z_0 = 0$ so that the general ray path would be given by

$$x = \pm x_0 \sin \Gamma z \tag{42}$$

We could have also used Eq. (28) to obtain the ray path. Now, in an optical waveguide the refractive index distribution is usually written in the form[4]:

$$n^2(x) = \begin{cases} n_1^2 \left[1 - 2\Delta \left(\frac{x}{a}\right)\right]^2 & |x| < a \text{ core} \\ n_2^2 = n_1^2(1 - 2\Delta) & |x| < a \text{ cladding} \end{cases} \tag{43}$$

The region $|x| < a$ is known as the *core* of the waveguide, and the region $|x| > a$ is usually referred to as the *cladding*. Thus

$$\gamma = \frac{n_1 \sqrt{2\Delta}}{a} \tag{44}$$

In a typical parabolic index fiber,

$$n_1 = 1.5 \quad \Delta = 0.01 \quad a = 20\ \mu\text{m} \tag{45}$$

giving

$$n_2 \approx 1.485$$

and

$$\gamma \approx 1.0607 \times 10^4\ \text{m}^{-1}$$

Typical ray paths for different values of θ_1 are shown in Fig. 3.23. Obviously, the rays will be guided in the core if

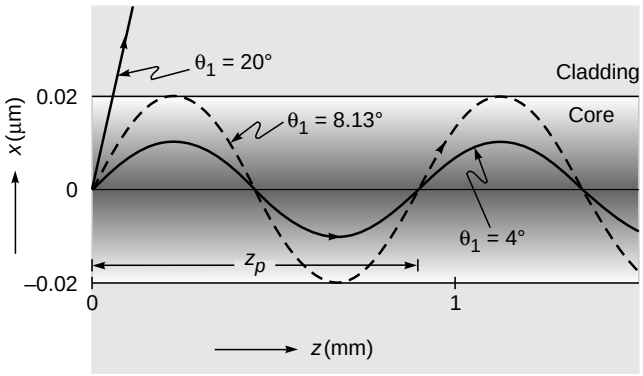


Fig. 3.23 Typical ray paths in a parabolic index medium for parameters given by Eq. (45) for $\theta_1 = 4°$, 8.13°, and 20°.

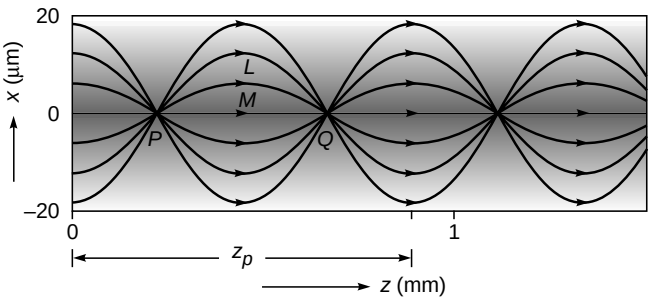


Fig. 3.24 Paraxial ray paths in a parabolic index medium. Notice the periodic focusing and defocusing of the beam.

$n_2 < \tilde{\beta} < n_1$. When $\tilde{\beta} = n_2$, the ray path will become horizontal at the core-cladding interface. For $\tilde{\beta} < n_2$, the ray will be incident at the core-cladding interface at an angle, and the ray will be refracted away. Thus, we may write

$$\begin{aligned} n_2 < \tilde{\beta} < n_1 \quad &\Rightarrow \quad \text{guided rays} \\ \tilde{\beta} < n_2 \quad &\Rightarrow \quad \text{refracting rays} \end{aligned} \tag{46}$$

In Fig. 3.23, the ray paths shown correspond to

$$z_0 = 0 \quad \text{and} \quad \theta_1 = 4°, 8.13°, \text{and } 20°$$

the corresponding values of $\tilde{\beta}$ are approximately 1.496 ($> n_2$), 1.485 ($= n_2$), and 1.410 ($< n_2$)—the last ray undergoes refraction at the core-cladding interface. The periodical length z_p of the sinusoidal path is given by

$$z_p = \frac{2\pi}{\Gamma} = \frac{2\pi a \cos\theta_1}{\sqrt{2\Delta}} \tag{47}$$

Thus for the two rays shown in Fig. 3.23 (with $\theta_1 = 4°$ and 8.13°) the values of z_p would be 0.8864 and 0.8796 mm, respectively. In the paraxial approximation, $\cos\theta_1 \approx 1$ and all rays have the same periodic length. In Fig. 3.24, we have plotted

[4] Ray paths in such media are of tremendous importance as they readily lead to very important results for parabolic index fibers which are extensively used in fiber-optic communication systems (see Sec. 27.7).

typical paraxial ray paths for rays launched along the z axis. Different rays (shown in the figure) correspond to different values of $\tilde{\beta}$.

Four interesting features may be noted:

1. In the paraxial approximation $\tilde{\beta} \approx n_1$, all rays launched horizontally come to a focus at a particular point. Thus the medium acts as a converging lens of focal length given by
$$f \approx \frac{\pi}{2}\frac{a}{\sqrt{2\Delta}} \qquad \textbf{(48)}$$
2. Rays launched at different angles with the axis (see, for instance, the rays emerging from point P) get trapped in the medium, and hence the medium acts as a "guide." Indeed such media are referred to as optical waveguides, and their study forms a subject of great contemporary interest.
3. Ray paths would be allowed only in the region where $\tilde{\beta}$ is less than or equal to $n(x)$ [see Eq. (26)]. Further, dx/dz would be zero (i.e., the ray would become parallel to the z axis) when $n(x)$ equaled $\tilde{\beta}$; this immediately follows from Eq. (27).
4. The rays periodically focus and defocus as shown in Fig. 3.24. In the paraxial approximation, all rays emanating from P will focus at Q; and if we refer to our discussion in Example 3.3, all rays must take the same time to go from P to Q. Physically, although the ray PLQ traverses a larger path in comparison to PMQ, it does so in a medium of "lower" average refractive index—thus the greater path length is compensated for by a greater "average speed" and hence *all* rays take the same time to propagate through a certain distance of the waveguide (see Sec. 3.4.2 for an exact calculation). It is for this reason that parabolic index waveguides are extensively used in fiber-optic communication systems (see Sec. 27.7).

We may mention here that gradient index (GRIN) lenses, characterized by parabolic variation of refractive index in the transverse direction, are now commercially available and find use in many applications (see Fig. 3.25). For example, a GRIN lens can be used to couple the output of a laser diode to an optical fiber; the length of such a GRIN lens would be $z_P/4$ (see Fig. 3.24); typically $z_P \approx$ few centimeters, and the diameter of the lens would be few millimeters. Such small size lenses find many applications. Similarly, a GRIN lens of length $z_P/2$ can be used to transfer collimated light from one end of the lens to the other.

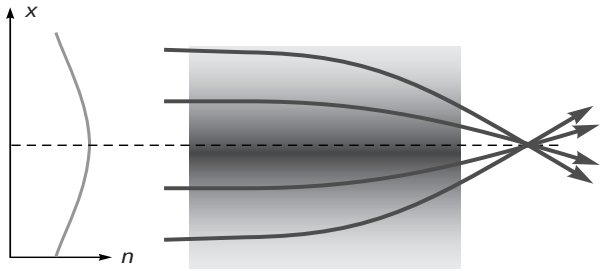


Fig. 3.25 A GRIN lens with a near parabolic refractive index variation will focus a light beam similar to a conventional lens. Because of this property, a GRIN lens (with properly chosen length) can be used to couple light from a laser diode to an optical fiber or can be used to transfer a collimated light beam from one end of the lens to the other.

3.4.2 Transit Time Calculations in a Parabolic Index Waveguide

In this section we will calculate the time taken by a ray to traverse a certain length through a parabolic index waveguide as described by Eq. (36). Such a calculation is of considerable importance in fiber-optic communication systems (see Sec. 27.7). As shown in Sec. 3.4.1, the ray path (inside the core) is given by

$$x = x_0 \sin \Gamma z \qquad \textbf{(49)}$$

where x_0 and Γ have been defined through Eqs. (39) and (40). Let $d\tau$ represent the time taken by a ray to traverse the arc length ds [see Fig. 3.12(b)]:

$$d\tau = \frac{ds}{c/n(x)} \qquad \textbf{(50)}$$

where c is the speed of light in free space. Since

$$n(x)\frac{dz}{ds} = \tilde{\beta}$$

[see Eq. (26)], we may write Eq. (50) as

$$d\tau = \frac{1}{c\tilde{\beta}}\, n^2(x)\, dz$$

$$= \frac{1}{c\tilde{\beta}}\left(n_1^2 - \gamma^2 x^2\right) dz$$

or
$$d\tau = \frac{1}{c\tilde{\beta}}\left(n_1^2 - \gamma^2 x_0^2 \sin^2 \Gamma z\right) dz \qquad \textbf{(51)}$$

where in the last step we have used Eq. (49). Thus if $\tau(z)$ represents the time taken by the ray to traverse a distance z

along the waveguide, then

$$\tau(z) = \frac{n_1^2}{c\tilde{\beta}}\int_0^z dz - \frac{\gamma^2 x_0^2}{c\tilde{\beta}}\int_0^z \frac{1-\cos 2\Gamma z}{2}\,dz$$

$$= \frac{1}{c\tilde{\beta}}\left(n_1^2 - \frac{1}{2}\gamma^2 x_0^2\right) z + \frac{\gamma^2 x_0^2}{2c\tilde{\beta}}\,\frac{1}{2\Gamma}\sin 2\Gamma z$$

or
$$\tau(z) = \frac{1}{2c\tilde{\beta}}\left(n_1^2 + \tilde{\beta}^2\right) z + \frac{n_1^2 - \tilde{\beta}^2}{4c\gamma}\sin 2\Gamma z \qquad \textbf{(52)}$$

where we have used Eq. (39). When $\tilde{\beta} = n_1$ (which corresponds to the ray along the z axis),

$$\tau(z) = \frac{z}{c/n_1} \qquad \textbf{(53)}$$

which is what we should have expected as the ray will *always* travel with speed c/n_1. For large values of z, the second term on the RHS of Eq. (52) would make a negligible contribution to $\tau(z)$, and we may write

$$\tau(z) \approx \frac{1}{2c}\left(\tilde{\beta} + \frac{n_1^2}{\tilde{\beta}}\right) z \qquad \textbf{(54)}$$

Now, if a pulse of light is incident on one end of the waveguide, it will in general excite all rays, and since different rays take different amounts of time, the pulse will get temporally broadened. Thus, for a parabolic index waveguide, this broadening will be given by

$$\Delta\tau = \tau(\tilde{\beta} = n_2) - \tau(\tilde{\beta} = n_1)$$

or
$$\Delta\tau = \frac{z}{2c}\,\frac{(n_1 - n_2)^2}{n_2} \approx \frac{z n_2}{2c}\,\Delta^2 \qquad \textbf{(55)}$$

where in the last step we have assumed

$$\Delta \equiv \frac{n_1^2 - n_2^2}{2n_1^2} \approx \frac{n_1 - n_2}{n_2} \qquad \textbf{(56)}$$

For the fiber parameters given by Eq. (45), we get

$$\Delta\tau \approx 0.25 \text{ ns km}^{-1} \qquad \textbf{(57)}$$

We will use this result in Chap. 27.

Example 3.9 We next consider the ray paths in a medium characterized by the refractive index variation

$$n^2(x) = \begin{cases} n_1^2 & x < 0 \\ n_1^2 - gx & x > 0 \end{cases} \qquad \textbf{(58)}$$

Thus, in the region $x > 0$, $n^2(x)$ decreases linearly with x and Eq. (28) takes the form

$$\frac{d^2x}{dz^2} = -\frac{g}{2\tilde{\beta}^2}$$

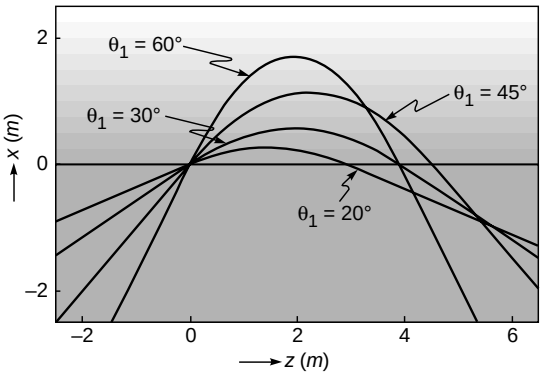


Fig. 3.26 Parabolic ray paths (corresponding to $\theta_1 = 20°$, 30°, 45°, and 60°) in a medium characterized by refractive index variation given by Eq. (58). The ray paths in the region $x < 0$ are straight lines.

the general solution of which is given by

$$x(z) = -\frac{g}{4\tilde{\beta}^2}\,z^2 + K_1 z + K_2 \qquad \textbf{(59)}$$

Consider a ray incident on the origin ($x = 0$, $z = 0$) as shown in Fig. 3.26. Thus

$$K_2 = 0 \qquad \text{and} \qquad \tilde{\beta} = n_1 \cos\theta_1 \qquad \textbf{(60)}$$

Further,

$$\left.\frac{dx}{dz}\right|_{z=0} = K_1 = \tan\theta_1 \qquad \textbf{(61)}$$

Thus the ray path will be given by

$$x(z) = \begin{cases} (\tan\theta_1)\, z & z < 0 \\ -\dfrac{gz}{4\tilde{\beta}^2}(z - z_0) & 0 < z < z_0 \\ -\dfrac{gz_0}{4\tilde{\beta}^2}(z - z_0) & z > z_0 \end{cases} \qquad \textbf{(62)}$$

where

$$z_0 = \frac{2n_1^2}{g}\sin 2\theta_1$$

Thus in the region $0 < z < z_0$, the ray path is a parabola. Typical ray paths are shown in Fig. 3.26, and the calculations correspond to

$$n_1 = 1.5, \quad g = 0.1 \text{ m}^{-1}$$

and different rays correspond to

$$\theta_1 = \frac{\pi}{9},\ \frac{\pi}{6},\ \frac{\pi}{4},\ \text{and}\ \frac{\pi}{3}$$

3.4.3 Reflections from the Ionosphere

The ultraviolet rays in the solar radiation result in the ionization of the constituent gases in the atmosphere, resulting in the formation of what is known as the ionosphere. The ionization is almost negligible below a height of about 60 km. Because of the presence of the free electrons (in the ionosphere), the refractive index is given by [see Eq. (76) of Chap. 7].

$$n^2(x) = 1 - \frac{N_e(x)q^2}{m\varepsilon_0\omega^2} \qquad \textbf{(63)}$$

where

$N_e(x)$ = number of electrons/unit volume, m^{-3}
x = height above the ground,
ω = angular frequency of electromagnetic wave
$q \approx 1.60 \times 10^{-19}$ C represents the charge of the electron
$m \approx 9.11 \times 10^{-31}$ kg represents the mass of the electron
$\varepsilon_0 \approx 8.854 \times 10^{-12} \times 10^{-12}$ C^2 N^{-1} m^{-2} represents the dielectric permittivity of vacuum

Thus as the electron density starts increasing from 0 (beyond the height of 60 km), the refractive index starts decreasing and the ray path would be similar to that described in Example 3.9.

If n_T represents the refractive index at the turning point (where the ray becomes horizontal), then (see Fig. 3.27)

$$\tilde{\beta} = \cos\theta_1 = n_T \qquad \textbf{(64)}$$

Thus if an electromagnetic signal sent from the point A (at an angle θ_1) is received at the point B, one can determine the refractive index (and hence the electron density) of the ionospheric layer where the beam has undergone the reflection. This is how the shortwave radio broadcasts ($\lambda \approx 20$ m) sent at a particular angle from a particular city (say, London) would reach another city (say, New Delhi) after undergoing reflection from the ionosphere. Further, for normal incidence, $\theta = \pi/2$ and $n_T = 0$, implying

$$N_e(x_T) = \frac{m\varepsilon_0\omega^2}{q^2} \qquad \textbf{(65)}$$

In a typical experiment, an electromagnetic pulse (of frequency between 0.5 and 20 MHz) is sent vertically upward,

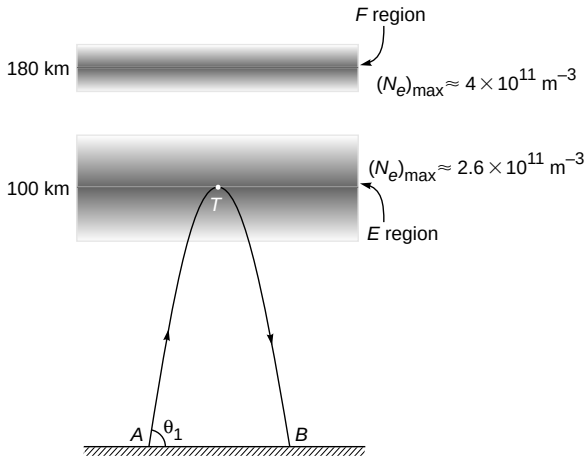


Fig. 3.27 Reflection from the E region of the ionosphere. The point T represents the turning point. The shading shows the variation of electron density.

and if the echo is received after a delay of Δt, then

$$\Delta t \approx \frac{2h}{c} \qquad \textbf{(66)}$$

where h represents the height at which it undergoes reflection. Thus if electromagnetic pulse is reflected from the E layer of the ionosphere (which is at a height of about 100 km), the echo will be received after about 670 μs. Alternatively, by measuring the delay Δt, one can determine the height (at which the pulse gets reflected) from the relation

$$h \approx \frac{c}{2}\Delta t \qquad \textbf{(67)}$$

In Fig. 3.28 we plotted the frequency dependence of the equivalent height of reflection (as obtained from the delay time of echo) from the E and F regions of the ionosphere.

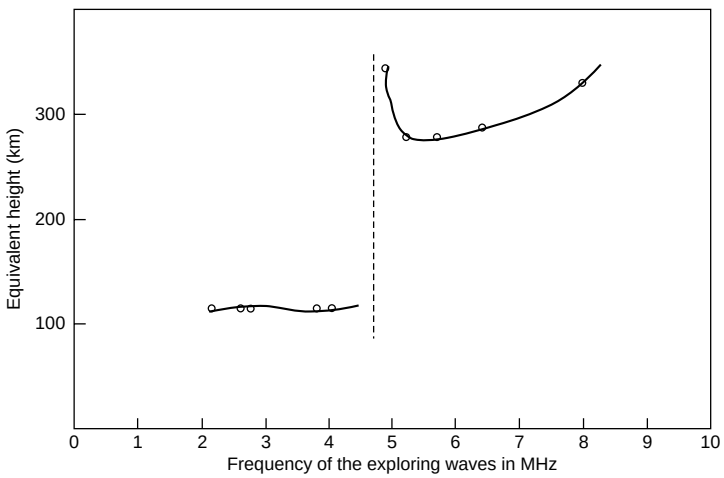


Fig. 3.28 Frequency dependence of the equivalent height of reflection from the E and F regions of the ionosphere [Adapted from Ref. 9].

From the figure we find that at $\nu = 4.6 \times 10^6$ Hz, echoes suddenly disappear from the 100 km height. Thus,

$$N_e(100\text{ km}) \approx \frac{m\varepsilon_0 (2\pi\nu)^2}{q^2}$$

$$\approx \frac{9.11\times10^{-31}\times 8.854\times10^{-12}\times(2\pi\times4.6\times10^6)^2}{(1.6\times10^{-19})^2}$$

$$\approx 2.6\times10^{11}\text{ electrons/m}^3$$

If we further increase the frequency, the echoes appear from the *F* region of the ionosphere. For more details of the studies on the ionosphere, the reader is referred to one of the most outstanding texts on the subject by Professor S. K. Mitra (Ref. 9).

Example 3.10 In this example we will obtain the solution of the ray equation for the refractive index variation given by

$$n^2(x) = n_0^2 + n_2^2(1 - e^{-\alpha x}) \quad \textbf{(68)}$$

Substituting in Eq. (27), we obtain

$$\pm dz = \frac{\tilde{\beta}\,dx}{\left[(n_0^2 + n_2^2 - \tilde{\beta}^2) - n_2^2 e^{-\alpha x}\right]^{1/2}}$$

$$= \frac{\tilde{\beta} e^{\alpha x/2}\,dx}{n_2\left(K^2 e^{\alpha x} - 1\right)^{1/2}}$$

or

$$\pm dz = \frac{2\tilde{\beta}}{K\alpha n_2}\frac{d\Phi}{(\Phi^2-1)^{1/2}} \quad \textbf{(69)}$$

where

$$K = \frac{1}{n_2}(n_0^2 + n_2^2 - \tilde{\beta}^2)^{1/2} \quad \textbf{(70)}$$

and

$$\Phi(x) = Ke^{\alpha x/2} \quad \textbf{(71)}$$

The $\pm$ sign in Eq. (69) corresponds to a ray going up and a ray going down, respectively. Further,

$$\tilde{\beta} = n_1 \cos\theta_1 \quad \textbf{(72)}$$

where θ_1 is the angle that the ray initially makes with the *z* axis at $x = x_1$, $z = 0$, and $n_1 = n(x_1)$. Carrying out the elementary integration, we get

$$x(z) = \frac{2}{\alpha}\ln\left[\frac{1}{K}\cosh\gamma(C \pm z)\right] \quad \textbf{(73)}$$

where

$$\gamma = \frac{\alpha K n_2}{2\tilde{\beta}} \quad \textbf{(74)}$$

which gives us the ray path. Since $x = x_1$ at $z = 0$ (the initial point),

$$C = \frac{1}{\gamma}\cosh^{-1}\left(Ke^{\alpha x_1/2}\right) \quad \textbf{(75)}$$

Further,

$$Ke^{\alpha x_1/2} = \left(\frac{n_0^2 + n_p^2 - \tilde{\beta}^2}{n_0^2 + n_p^2 - n_1^2}\right)^{1/2} \quad \textbf{(76)}$$

Thus for a ray launched horizontally at $x = x_1$, $C = 0$. Typically ray paths (for different values of θ_1) are shown in Figs. 3.14 and 3.15.

3.5 REFRACTION OF RAYS AT THE INTERFACE BETWEEN AN ISOTROPIC MEDIUM AND AN ANISOTROPIC MEDIUM

In this section we will use Fermat's principle to determine the direction of the refracted ray for a ray incident at the interface of an isotropic and an anisotropic medium.[5] We point out that in an isotropic medium the properties remain the same in all directions; typical examples are glass, water, and air. On the other hand, in an anisotropic medium, some of the properties (such as speed of light) may be different in different directions. In Chap. 22, we will consider anisotropic media in greater detail; we mention here that when a light ray is incident on a crystal such as calcite, (in general) it splits into two rays known as ordinary and extraordinary rays. The velocity of the ordinary ray is the same in all directions. Thus the ordinary ray obeys Snell's laws, but the extraordinary ray does not. We will now use Fermat's principle to study the refraction of a ray when it is incident from an isotropic medium into an anisotropic medium—both media are assumed to be homogeneous.

In a uniaxial medium, the refractive index variation for the extraordinary ray is given by [see Eq. (121) of Chap. 22]

$$n^2(\theta) = n_o^2 \cos^2\theta + n_e^2 \sin^2\theta \quad \textbf{(77)}$$

where n_o and n_e are constants of the crystal and θ represents the angle that the ray makes with the optic axis. Obviously, when the extraordinary ray propagates parallel to the optic axis (i.e., when $\theta = 0$), its speed is c/n_o and when it propagates perpendicular to the optic axis ($\theta = \pi/2$), its speed is c/n_e.

[5] A proof for the applicability of Fermat's principle in anisotropic media has been given by Newcomb (Ref. 10); the proof, however, is quite complicated. Ray paths in biaxial media are discussed in Ref. 11.

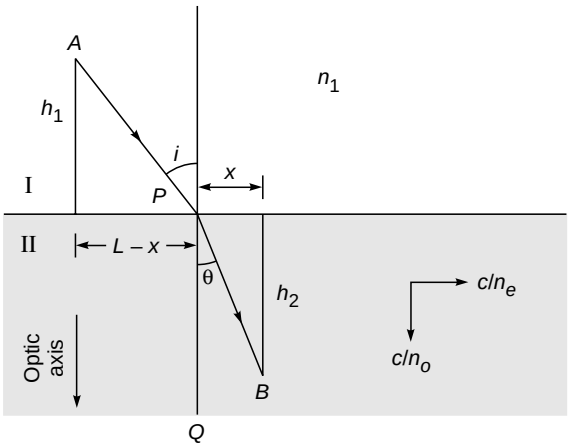


Fig. 3.29 The direction of the refracted extraordinary ray when the optic axis (of the uniaxial crystal) is normal to the surface.

3.5.1 Optic Axis Normal to the Surface

We first consider the particularly simple case of the optic axis being normal to the surface. Referring to Fig. 3.29, the optical path length from A and B is given by

$$L_{\text{op}} = n_1[h_1^2 + (L - x)^2]^{1/2} + n(\theta)\left(h_2^2 + x^2\right)^{1/2} \qquad \textbf{(78)}$$

where n_1 is the refractive index of medium I and we have assumed the incident ray, the refracted ray, and the optic axis to lie in the same plane. Since

$$\cos\theta = \frac{h_2}{(h_2^2 + x^2)^{1/2}} \quad \text{and} \quad \sin\theta = \frac{x}{(h_2^2 + x^2)^{1/2}}$$

we have

$$L_{\text{op}} = n_1[h_1^2 + (L - x)]^{1/2} + \left(n_o^2 h_2^2 + n_e^2 x^2\right)^{1/2} \qquad \textbf{(79)}$$

For the actual ray path, we must have

$$\frac{dL_{\text{op}}}{dx} = 0$$

implying

$$\frac{n_1 (L - x)}{[h_1^2 + (L - x)^2]^{1/2}} = \frac{n_e^2 x}{\left(n_o^2 h_2^2 + n_e^2 x^2\right)^{1/2}}$$

or

$$n_1 \sin i = \frac{n_e^2 \tan r}{\left(n_o^2 + n_e^2 \tan^2 r\right)^{1/2}} \qquad \textbf{(80)}$$

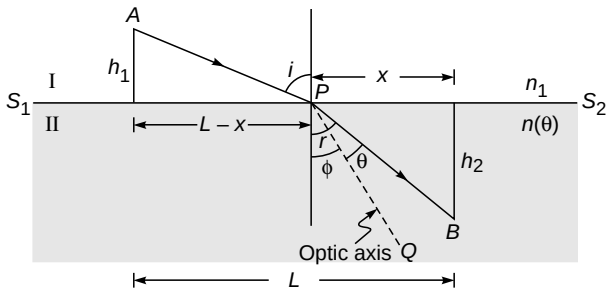


Fig. 3.30 The direction of the refracted extraordinary ray when the optic axis (of the uniaxial crystal) lies in the plane of incidence, making an angle ϕ with the normal to the interface.

where we have used the fact that

$$\text{Angle of refraction } r = \theta \qquad \text{and} \qquad \tan r = \frac{x}{h_2}$$

Simple manipulations give us

$$\tan r = \frac{n_o n_1 \sin i}{n_e \sqrt{n_e^2 - n_1^2 \sin^2 i}} \qquad \textbf{(81)}$$

Using this, we can calculate the angle of refraction for a given angle of incidence (when the optic axis is normal to the surface). As a an example, we assume the first medium to be air so that $n_1 = 1$. Then

$$\tan r = \frac{n_o \sin i}{n_e \sqrt{n_e^2 - \sin^2 i}} \qquad \text{when } n_1 = 1 \qquad \textbf{(82)}$$

If we assume the second medium to be calcite, then

$$n_o = 1.65836 \qquad \text{and} \qquad n_e = 1.48641$$

Thus for $i = 45°$, we readily get

$$r \approx 31.1°$$

If $n_o = n_e = n_2$ (say), then Eq. (80) simplifies to

$$n_1 \sin i = n_2 \sin r \qquad \textbf{(83)}$$

which is nothing but Snell's law.

3.5.2 Optic Axis in the Plane of Incidence[6]

We next consider a more general case of the optic axis making an angle ϕ with the normal; however, the optic axis is assumed to lie in the plane of incidence as shown in Fig. 3.30. We may mention here that in general, in an anisotropic medium, the refracted ray does not lie in the plane of incidence. However, it can be shown that if the optic axis lies in the plane of incidence, then the refracted ray also lies in

[6] May be skipped in the first reading.

the plane of incidence. In the present calculation, we are assuming this and finding the direction of the refracted ray for a given angle of incidence. Now, the optical path length from A to B (see Fig. 3.30) is given by

$$L_{\text{op}} = n_1[h_1^2 + (L - x)^2]^{1/2} + n(\theta)\left(h_2^2 + x^2\right)^{1/2} \quad \textbf{(84)}$$

Since $\theta = r - \phi$, we have

$$\begin{aligned} n^2(\theta) &= n_o^2 \cos^2(r - \phi) + n_e^2 \sin^2(r - \phi) \\ &= n_o^2 (\cos r \cos \phi + \sin r \sin \phi)^2 \\ &\quad + n_e^2(\sin r \cos \phi - \cos r \sin \phi)^2 \\ &= n_o^2 \left(\frac{h_2}{\sqrt{h_2^2 + x^2}} \cos \phi + \frac{x}{\sqrt{h_2^2 + x^2}} \sin \phi \right)^2 \\ &\quad + n_e^2 \left(\frac{x}{\sqrt{h_2^2 + x^2}} \cos \phi - \frac{h_2}{\sqrt{h_2^2 + x^2}} \sin \phi \right)^2 \end{aligned}$$

Thus

$$n(\theta) = \frac{1}{\sqrt{h_2^2 + x^2}} [n_o^2 (h_2 \cos \phi + x \sin \phi)^2 + n_e^2(x \cos \phi - h_2 \sin \phi)^2]^{1/2} \quad \textbf{(85)}$$

and

$$L_{\text{op}} = n_1[h_1^2 + (L - x)^2]^{1/2} + [n_o^2(h_2 \cos\phi + x \sin\phi)^2 + n_e^2(x \cos\phi - h_2 \sin\phi)^2]^{1/2} \quad \textbf{(86)}$$

For the actual ray path, we must have

$$\frac{dL_{\text{op}}}{dx} = 0$$

implying

$$\frac{n_1(L - x)}{[h_1^2 + (L - x)^2]^{1/2}} = \frac{n_o^2(h_2 \cos\phi + x \sin\phi) \sin\phi + n_e^2(x \cos\phi - h_2 \sin\phi) \cos\phi}{[n_o^2(h_2 \cos\phi + x \sin\phi)^2 + n_e^2(x \cos\phi - h_2 \sin\phi)^2]^{1/2}}$$

or
$$n_1 \sin i = \frac{n_o^2 \cos\theta \sin\phi + n_e^2 \sin\theta \cos\phi}{\left(n_o^2 \cos^2\theta + n_e^2 \sin^2\theta\right)^{1/2}} \quad \textbf{(87)}$$

For given values of the angles i and ϕ, the above equation can be solved to give the values of θ and hence the angle of refraction r $(= \theta + \phi)$.

Some interesting particular cases may be noted.

1. When $n_o = n_e = n_2$, the anisotropic medium becomes isotropic and Eq. (87) simplifies to

$$n_1 \sin i = n_2 \sin(\theta + \phi) = n_2 \sin r$$

which is nothing but Snell's law.

2. When $\phi = 0$, i.e., the optic axis is normal to the surface, Eq. (87) becomes

$$n_1 \sin i = \frac{n_e^2 \sin\theta}{\left(n_o^2 \cos^2\theta + n_e^2 \sin^2\theta\right)^{1/2}} = \frac{n_e^2 \sin r}{\left(n_o^2 \cos^2 r + n_e^2 \sin^2 r\right)^{1/2}} \quad \textbf{(88)}$$

where we have used the fact that $r = \theta$. The above equation is identical to Eq. (80).

3. Finally, we consider normal incidence, i.e., $i = 0$. Thus, Eq. (87) gives us

$$n_o^2 \cos\theta \sin\phi + n_e^2 \sin\theta \cos\phi = 0$$

or

$$n_o^2 \cos(r - \phi) \sin\phi + n_e^2 \sin(r - \phi) \cos\phi = 0$$

or

$$\cos r\,(n_o^2 \cos\phi \sin\phi - n_e^2 \sin\phi \cos\phi) + \sin r\,[n_o^2 \sin^2\phi + n_e^2 \cos^2\phi) = 0$$

or

$$\tan r = \frac{(n_e^2 - n_o^2) \sin\phi \cos\phi}{n_o^2 \sin^2\phi + n_e^2 \cos^2\phi} \quad \textbf{(89)}$$

Equation (89) shows that in general $r \neq 0$ (see Fig. 3.31). For normal incidence, the above analysis is valid for an arbitrary orientation of the optic axis; the refracted (extraordinary) ray lies in the plane containing the normal and the optic axis. Furthermore, for normal incidence, when the crystal is rotated about the normal, the refracted ray also rotates on the surface of a cone [see Fig. 22.18(b)].

Returning to Eq. (89), we note that when the optic axis is normal to the surface ($\phi = 0$) or when the optic axis is parallel to the surface but lying in the plane of incidence ($\phi = \pi/2$), $r = 0$ and the ray goes undeviated.

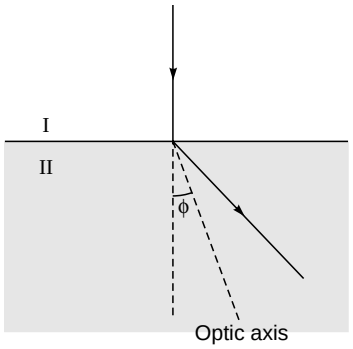


Fig. 3.31 For normal incidence, in general, the refracted extraordinary ray undergoes finite deviation. However, the ray proceeds undeviated when the optic axis is parallel or normal to the surface.

Summary

- The slightly modified version of Fermat's principle is that *the actual ray path between two points is the one for which the optical path length is stationary with respect to variations of the path.*
- Laws of reflections and Snell's law of refraction ($n_1 \sin \phi_1 = n_2 \sin \phi_2$, where ϕ_1 and ϕ_2 represent the angles of incidence and refraction) can be derived from Fermat's principle.
- For an inhomogeneous medium characterized by the refractive index variation $n(x)$, the ray paths $x(z)$ are such that the product $n(x) \cos \theta(x)$ remains constant, and here $\theta(x)$ is the angle that the ray makes with the z axis; this constant is denoted by $\tilde{\beta}$ which is known as the ray invariant. The exact ray paths are determined by solving either of the equations

$$\frac{dx}{dz} = \pm \frac{\sqrt{n^2(x) - \tilde{\beta}^2}}{\tilde{\beta}}$$

or

$$\frac{d^2x}{dz^2} = \frac{1}{2\tilde{\beta}^2} \frac{dn^2(x)}{dx}$$

  where the invariant $\tilde{\beta}$ is determined from the initial launching condition of the ray.
- Ray paths obtained by solving the ray equation can be used to study mirage, looming, and reflections from the ionosphere.
- In a parabolic index medium $n^2(x) = n_1^2 - \gamma^2 x^2$, the ray paths are sinusoidal:

$$x(z) = \pm x_0 \sin \Gamma z$$

  where $\Gamma = 1/\tilde{\beta}$, $x_0 = 1/\gamma\sqrt{n_1^2 - \tilde{\beta}^2}$, and we have assumed $z = 0$ where $x = 0$. Rays launched at different angles take approximately the same time in propagating through a large length of the medium.
- Fermat's principle can be used to study refraction of rays at the interface of an isotropic medium and an anisotropic medium.

Problems

3.1 In this and the following two problems we will use Fermat's principle to derive laws governing paraxial image formation by spherical mirrors.

Consider an object point O in front of a concave mirror whose center of curvature is at the point C. Consider an arbitrary point Q on the axis of the system, and using a method similar to that used in Example 3.3, show that the optical path length L_{op} (= $OS + SQ$) is approximately given by

$$L_{op} \approx x + y + \frac{1}{2} r^2 \left(\frac{1}{x} + \frac{1}{y} - \frac{2}{r} \right) \theta^2 \qquad \textbf{(90)}$$

where the distances x, y, and r and the angle θ are defined in Fig. 3.32; θ is assumed to be small. Determine the paraxial image point, and show that the result is consistent with the mirror equation

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{R} \qquad \textbf{(91)}$$

where u and v are the object and image distance and R is the radius of curvature with the sign convention that all distances to the right of P are positive and to its left negative.

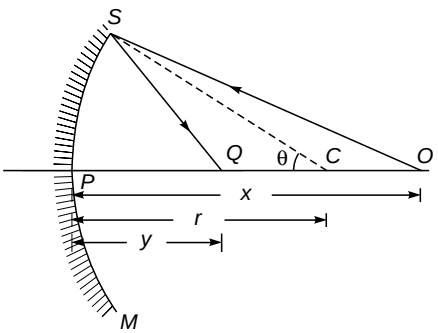


Fig. 3.32 Paraxial image formation by a concave mirror.

3.2 Fermat's principle can also be used to determine the paraxial image points when the object forms a virtual image. Consider an object point O in front of the convex mirror SPM (see Fig. 3.33). One should now assume the optical path length L_{op} to be $OS - SQ$; the minus sign occurs because the rays at S point away from Q (see Example 3.4). Show that

$$L_{op} \approx OS - SQ \approx x - y + \frac{1}{2} r^2 \left(\frac{1}{x} - \frac{1}{y} + \frac{2}{r} \right) \theta^2 \qquad \textbf{(92)}$$

where the distances x, y, and r and the angle θ are defined in Fig. 3.33. Show that the paraxial image is formed at $y = y_0$ which is given by

$$\frac{1}{x} - \frac{1}{y_0} = -\frac{2}{r} \qquad \textbf{(93)}$$

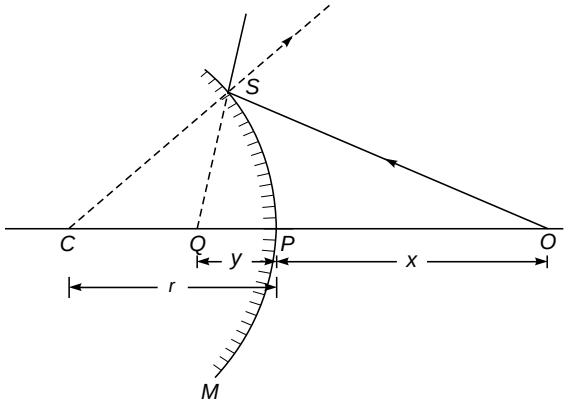


Fig. 3.33 Paraxial image formation by a convex mirror.

which is consistent with Eq. (91) because whereas the object distance u is positive, the image distance v and the radius of curvature R are negative since the image point and the center of curvature lie on the left of the point P.

3.3 Proceeding as in Prob. 3.2, use Fermat's principle to determine the mirror equation for an object point at a distance less than $R/2$ from a concave mirror of radius of curvature R.

3.4 We next consider a point object O in front of a concave refracting surface SPM separating two media of refracting indices n_1 and n_2 (see Fig. 3.34); C represents the center of curvature. In this case also one obtains a virtual image. Let Q represent an arbitrary point on the axis. We now have to consider the optical path length $L_{op} = n_1OS - n_2SQ$; show that it is given by

$$L_{op} = n_1OS - n_2SQ$$

$$\approx n_1x - n_2y - \frac{1}{2}r^2\left(\frac{n_2}{y} - \frac{n_1}{x} - \frac{n_2 - n_1}{r}\right)\theta^2 \qquad \textbf{(94)}$$

Also show that the above expression leads to the paraxial image point which is consistent with Eq. (10); we note that u, v, and R are all negative quantities because they are on the left of the refracting surface.

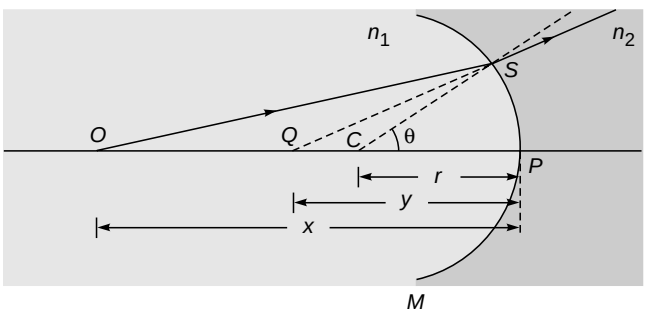


Fig. 3.34 Paraxial image formation by a concave refracting surface SPM.

3.5 If we rotate an ellipse about its major axis, we obtain what is known as an ellipsoid of revolution. Show by using Fermat's principle that all rays parallel to the major axis of the ellipse will focus to one of the focal points of the ellipse (see Fig. 3.35), provided the eccentricity of the ellipse equals n_1/n_2.

[*Hint*: Start with the condition that

$$n_2AC' = n_1QB + n_2BC$$

and show that the point B (whose coordinates are x and y) lies on the periphery of an ellipse.]

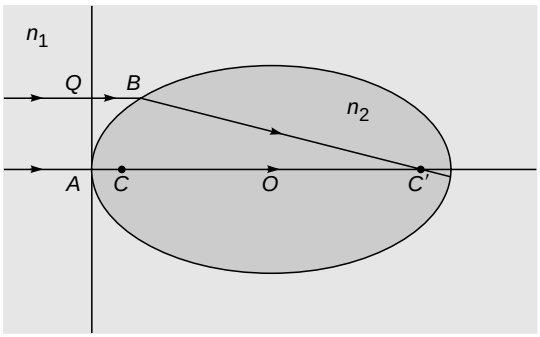


Fig. 3.35 All rays parallel to the major axis of the ellipsoid of revolution will focus to one of the focal points of the ellipse provided the eccentricity = n_1/n_2.

3.6 Point C is the center of the reflecting sphere of radius R (see Fig. 3.36). Points P_1 and P_2 are two points on a diameter equidistant from the center. **(a)** Obtain the optical path length $P_1O + OP_2$ as a function of θ, and **(b)** find the values of θ for which P_1OP_2 is a ray path from reflection at the sphere.

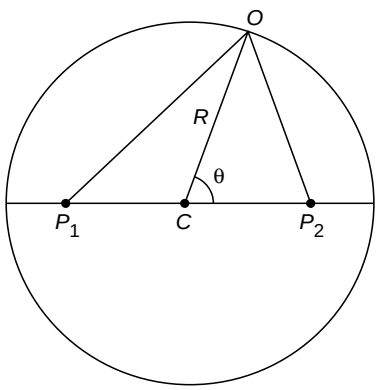


Fig. 3.36 A spherical reflector.

3.7 SPM is a spherical refracting surface separating two media of refractive indices n_1 and n_2 (see Fig. 3.37). Consider an

object point O forming a virtual image at the point I. We assume that *all* rays emanating from O appear to emanate from I so as to form a perfect image. Thus according to Fermat's principle, we must have

$$n_1 OS - n_2 SI = n_1 OP - n_2 PI$$

where S is an arbitrary point on the refracting surface. Assuming the right-hand side to be zero, show that the refracting surface is spherical, with the radius given by

$$r = \frac{n_1}{n_1 + n_2}\ OP \qquad \textbf{(95)}$$

Thus show that

$$n_1^2 d_1 = n_2^2 d_2 = n_1 n_2 r \qquad \textbf{(96)}$$

where d_1 and d_2 are defined in Fig. 3.37 (see also Fig. 4.12 and Sec. 4.10).

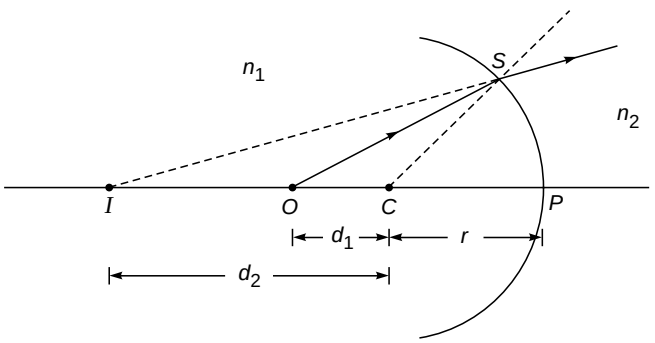


Fig. 3.37 All rays emanating from O and getting refracted by the spherical surface SPM appear to come from I.

[*Hint*: We consider a point C which is at a distance d_1 from the point O and d_2 from the point I. Assume the origin to be at O, and let (x, y, z) represent the coordinates of the point S. Thus

$$n_1(x^2 + y^2 + z^2)^{1/2} - n_2(x^2 + y^2 + \Delta^2)^{1/2} = n_1(r + d_1) - n_2(r + d_2) = 0$$

where $\Delta = d_2 - d_1$. The above equation would give the equation of a sphere whose center is at a distance of $n_2 r/n_1$ $(= d_1)$ from O.]

3.8 Referring to Fig. 3.38, if I represents a perfect image of the point O, show that the equation of the refracting surface (separating two media of refractive indices n_1 and n_2) is given by

$$n_1(x^2 + y^2 + z^2)^{1/2} + n_2[x^2 + y^2 + (z_2 - z)^2]^{1/2} = n_1 z_1 - n_2(z_2 - z_1) \qquad \textbf{(97)}$$

where the origin is assumed to be at the point O and the coordinates of P and I are assumed to be $(0, 0, z_1)$, and $(0, 0, z_2)$, respectively. The surface corresponding to Eq. (97) is known as a Cartesian oval.

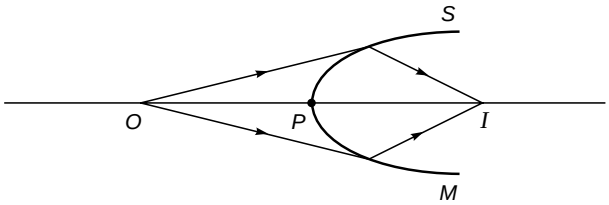


Fig. 3.38 The Cartesian oval. All rays emanating from O and getting refracted by SPM pass through I.

3.9 For the refractive index variation given by Eqs. (21) and (22), a ray is launched at $x = 0.43$ m, making an angle $-\pi/60$ with the z axis (see Fig. 3.14). Calculate the value of x at which it will become horizontal. [**Ans:** $x \approx 0.41$m]

3.10 For the refractive index variation given by Eqs. (21) and (22), a ray is launched at $x = 2.8$ m such that it becomes horizontal at $x = 0.2$ m (see Fig. 3.15). Calculate the angle that the ray will make with the z axis at the launching point. [**Ans:** $\theta_1 \approx 19°$]

3.11 Consider a parabolic index medium characterized by the following refractive index variation:

$$n^2(x) = \begin{cases} n_1^2\left[1 - 2\Delta\left(\dfrac{x}{a}\right)^2\right] & |x| < a \\ n_1^2(1 - 2\Delta) = n_2^2 & |x| > a \end{cases}$$

Assume $n_1 = 1.50$, $n_2 = 1.48$, and $a = 50$ μm. Calculate the value of Δ.

(a) Assume rays launched on the axis at $z = 0$ (i.e., $x = 0$ when $z = 0$) with

$\tilde{\beta}$ =1.495, 1.490, 1.485, 1.480, 1.475, and 1.470

In each case calculate the angle that the ray initially makes with the z axis (θ_1) and plot the ray paths. In each case find the height at which the ray becomes horizontal.

(b) Assume rays incident normally on the plane $z = 0$ at $x = 0$, ±10 μm, ±20 mm, ±30 μm, ±40 μm. Find the corresponding values of $\tilde{\beta}$, calculate the focal length for each ray, and qualitatively plot the ray paths.

3.12 In an inhomogeneous medium the refractive index is given by

$$n^2(x) = \begin{cases} 1 + \dfrac{x}{L} & \text{for } x > 0 \\ 1 & \text{for } x < 0 \end{cases}$$

Write down the equation of a ray (in the x/z plane) passing through the point (0, 0, 0) where its orientation with respect to x axis is 45°.

$$\left[\textbf{Ans: } x(z) = \frac{z^2}{4L\tilde{\beta}^2} + z\right]$$

3.13 For the refractive index profile given by Eq. (23), show that Eq. (27) can be written in the form

$$\pm \frac{\alpha K_1 n_2}{2\tilde{\beta}}\, dz = \frac{dG}{\sqrt{1-G^2}} \qquad \textbf{(98)}$$

where

$$K_1 = \frac{\sqrt{\tilde{\beta}^2 - n_0^2}}{n_2} \quad \text{and} \quad G(x) = K_1 e^{\alpha x/2} \qquad \textbf{(99)}$$

Integrate Eq. (98) to determine the ray paths.

$$\left[\textbf{Ans: } x(z) = \frac{2}{\alpha}\ \ln\left\{\frac{1}{K_1}\sin\left[\frac{n_2 K_1 \alpha}{2\tilde{\beta}}\left(z+z_0\right)\right]\right\}\right]$$

3.14 Consider a graded index medium characterized by the refractive index distribution

$$n^2(x) = n_1^2 \operatorname{sech}^2 gx \qquad \textbf{(100)}$$

Substitute in Eq. (27) and integrate to obtain

$$x(z) = \frac{1}{g}\ \sinh^{-1}\left(\frac{\sqrt{n_1^2 - \tilde{\beta}^2}}{\tilde{\beta}} \sin gz\right) \qquad \textbf{(101)}$$

Notice that the periodic length

$$z_p = \frac{2\pi}{g}$$

is independent of the launching angle (see Fig. 3.39), and *all* rays rigorously take the same amount of time in propagating through a distance z_p in the z direction.

[*Hint*: While carrying out the integration, make the substitution: $\zeta = \dfrac{\tilde{\beta}}{\sqrt{n_1^2 - \tilde{\beta}^2}}\ \sinh gx$.]

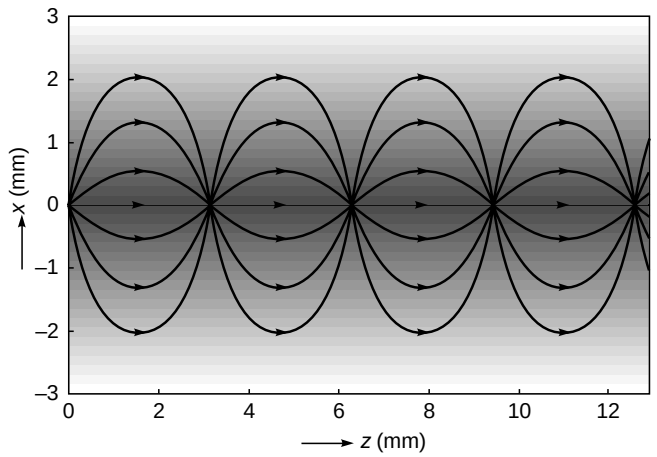


Fig. 3.39 Ray paths in a graded index medium characterized by Eq. (100).

3.15 For $z < 0$, $n = 1$

For $z > 0$,

$$n^2(x) = \begin{cases} n_1^2\left(1-\alpha|x|/a\right) & |x| < a \\ n_1^2\left[1-\alpha\right]; & |x| < a \end{cases}$$

$n_1 = 2.0$; $\alpha = 15/16$; $a = 30$ μm.

A ray is incident at the point A ($x = x_0 = 14$ μm, $z = 0$) as shown in Fig. 3.40. **(a)** Calculate $\tilde{\beta}$ for the ray inside the graded index medium. **(b)** Calculate the maximum height h of the ray. **(c)** Calculate the angle θ that the ray makes with the z axis at C. **(d)** Derive the equation of the ray path. **(e)** Calculate the time taken for the ray to traverse from B to C.

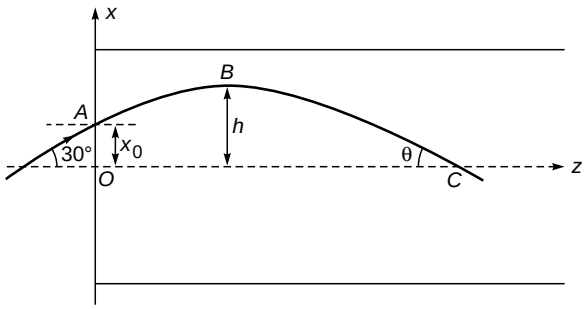


Fig. 3.40

3.16

$$n^2(x) = \begin{cases} 4 & \text{for } x < 0 \\ 4\left(1-\dfrac{x^2}{a^2}\right) & \text{for } 0 < x < \sqrt{3}\ \text{mm} \\ 1 & \text{for } x > \sqrt{3}\text{mm} \end{cases}$$

where $a = 2$ mm. A ray is launched at 45° as shown in Fig. 3.41.

(a) Determine the ray path.

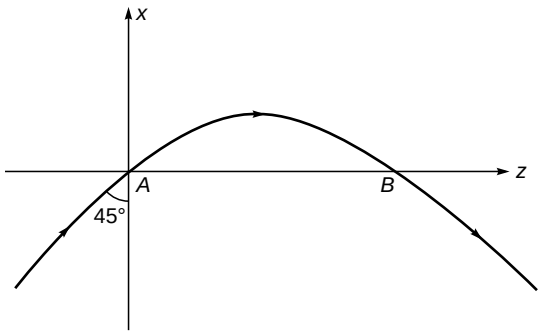


Fig. 3.41

(b) What is the time taken by the ray from A to B?

$$\left[\textbf{Ans: (a)}\ x = \frac{a}{\sqrt{2}}\sin\left(\frac{\sqrt{2}z}{a}\right);\ \textbf{(b)}\ \frac{3\pi a}{2c}\right]$$

REFERENCES AND SUGGESTED READINGS

1. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, United Kingdom, 1975.
2. A. K. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978. [Reprinted by Macmillan India, New Delhi.]
3. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley Publishing Co., Reading, Mass., 1965.
4. M. S. Sodha, A. K. Aggarwal, and P. K. Kaw, "Image Formation by an Optically Stratified Medium: Optics of Mirage and Looming," *British Journal of Applied Physics*, vol. 18, p. 503, 1967.
5. R. T. Bush and R. S. Robinson, "A Note Explaining the Mirage," *American Journal of Physics*, Vol. 42, p. 774, 1974.
6. A. B. Fraser and W. H. Mach, "Mirages," *Scientific American*, January, Vol. 234, p. 102, 1976.
7. E. Khular, K. Thyagarajan, and A. K. Ghatak, "A Note on Mirage Formation," *American Journal of Physics*, Vol. 45, p. 90, 1977.
8. W. J. Humphreys, *Physics of the Air*, McGraw-Hill Book Co., New York, 1920.
9. S. K. Mitra, *The Upper Atmosphere*, 2d ed., The Asiatic Society, Calcutta, India, 1952.
10. W. A. Newcomb, "Generalized Fermat's Principles," *American Journal of Physics*, Vol. 51, p. 338, 1983.
11. E. Khular, K. Thyagarajan, and A. K. Ghatak, "Ray Tracing in Uniaxial and Biaxial Media," *Optik*, Vol. 46, p. 297, 1976.
12. V. Lakshminarayanan, A. Ghatak, and K. Thyagarajan, *Lagrangian Optics*, Kluwer Academic Publishers, 2002.

Chapter Four

REFRACTION AND REFLECTION BY SPHERICAL SURFACES

The use of plane and curved mirrors and of convex and concave lenses were discovered independently in China and in Greece. References to burning mirrors go back almost to the start of history, and it is possible that Chinese and Greek knowledge were both derived from a common source in Mesopotamia, India or Egypt ... Pythagoras, Greek philosopher and mathematician (6^{th} century BC), suggested that light consists of rays that, acting like feelers, travel in straight lines from the eye to the object and that the sensation of sight is obtained when these rays touch the object. In this way, the more mysterious sense of sight is explained in terms of the intuitively accepted sense of touch. It is only necessary to reverse the direction of these rays to obtain the basic scheme of modern geometrical optics. The Greek mathematician Euclid (300 BC), who accepted the Pythagorean idea, knew that the angle of reflected light rays from a mirror equals the angle of incident light rays from the object to the mirror. The idea that light is emitted by a source and reflected by an object and then enters the eye to produce the sensation of sight was known to Epicurus, another Greek philosopher (300 BC). The Pythagorean hypothesis was eventually abandoned and the concept of rays traveling from the object to the eye was finally accepted about AD 1000 under the influence of an Arabian mathematician and physicist named Alhazen.

—*The New Encyclopedia Britannica*, Vol. 23

Alhazen had used spherical and parabolic mirrors and was aware of spherical aberration. He also investigated the magnification produced by lenses and atmospheric refraction. His work was translated into Latin and became accessible to later European scholars.

—From the Internet

4.1 INTRODUCTION

In this chapter we will study the formation of an image by simple optical systems. We will assume the optical system to be made up of a number of refracting surfaces such as a combination of lenses.[1] To trace a ray through such an optical system, it is necessary only to apply Snell's laws at each refracting surface which are as follows:

1. The incident ray, the refracted ray, and the normal (to the surface) lie in the same plane.
2. If ϕ_1 and ϕ_2 represent the angles of incidence and refraction, respectively, then

$$\frac{\sin \phi_1}{\sin \phi_2} = \frac{n_2}{n_1} \tag{1}$$

where n_1 and n_2 are the refractive indices of the two media (see Fig. 4.1). Although there is no additional physics involved (other than the Snell's laws) in the tracing of rays, the design of even a simple optical system involves tracing many rays and therefore considerable numerical computations. Nowadays, such numerical computations are usually done on a high-speed computer. In fact, optical designers were among the first to make use of electronic computers when they were introduced in the early 1950s.

[1] The optical system may also consist of mirrors, in which case the reflection of rays should also be taken into account (see Sec. 4.3).

4.2 REFRACTION AT A SINGLE SPHERICAL SURFACE

We will first consider refraction at a spherical surface *SPM* separating two media of refractive indices n_1 and n_2 [see Fig. 4.1(a)]. Let *C* represent the center of curvature of the spherical surface. We will consider a point object *O* emitting rays in all directions. We will use Snell's laws of refraction to determine the image of the point *O*. We mention that not all rays emanating from *O* converge to a single point; however, if we consider only those rays which make small angles with the line joining the points *O* and *C*, then all rays do converge to a single point *I* [see Fig. 4.1(a)]. This is known as the *paraxial approximation*, and according to Fermat's principle all paraxial rays take the *same* amount of time to travel from *O* to *I* (see Example 3.3).

Now, in terms of the angles defined in Fig. 4.1(a), we have

$$\phi_1 = \beta + \alpha_1 \qquad \text{and} \qquad \phi_2 = \beta - \alpha_2$$

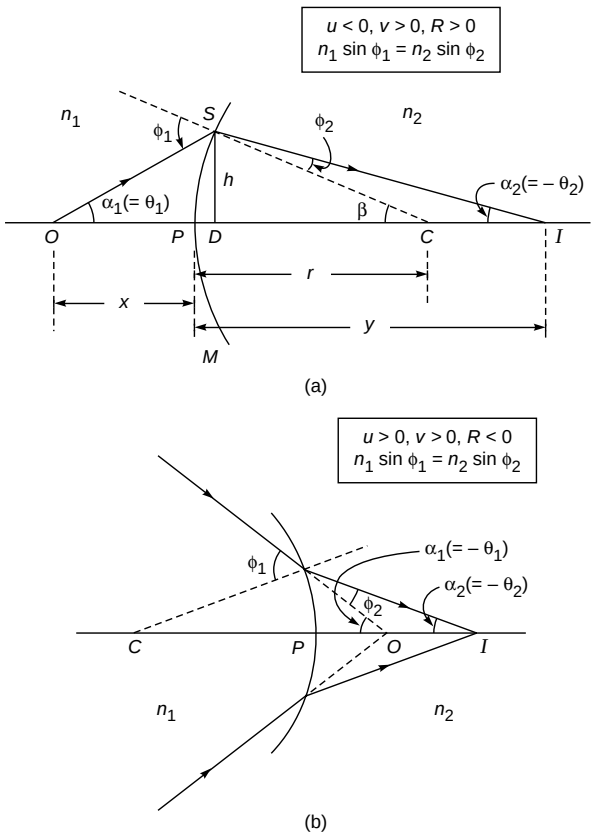


Fig. 4.1 (a) Paraxial image formation by a spherical refracting surface separating media of refractive indices n_1 and n_2. Point *O* represents the object point and *I* the paraxial image point. (b) Corresponds to positive *u*.

We next make use of the paraxial approximation, viz., all angles ϕ_1, ϕ_2, α_1, α_2, and β are small, so we may write

$$\sin \phi_1 \approx \tan \phi_1 \approx \phi_1 \qquad \text{etc.}$$

where the angles are obviously measured in radians. Thus, we have

$$\sin \phi_1 \approx \phi_1 = \beta + \alpha_1 \approx \tan \beta + \tan \alpha_1 \approx \frac{h}{r} + \frac{h}{x} \qquad \textbf{(2)}$$

and

$$\sin \phi_2 \approx \phi_2 = \beta - \alpha_2 \approx \tan \beta - \tan \alpha_2 \approx \frac{h}{r} - \frac{h}{y} \qquad \textbf{(3)}$$

where the distances *h*, *x*, *y*, and *r* are defined in Fig. 4.1(a) and we have assumed that the foot of the perpendicular *D* is very close to the point *P* so that $OD \approx OP = x$, $ID \approx IP = y$, etc. We now use Eqs. (1) – (3) to obtain (in the paraxial approximation)

$$n_1\left(\frac{h}{r} + \frac{h}{x}\right) = n_2\left(\frac{h}{r} - \frac{h}{y}\right)$$

or

$$\frac{n_2}{y} + \frac{n_1}{x} = \frac{n_2 - n_1}{r} \qquad \textbf{(4)}$$

4.2.1 The Sign Convention

Before we proceed further, we should state the sign convention which we will be using throughout in the book. We refer to Fig. 4.1(a) and consider the point *P* as the origin of the coordinate system. The sign convention is as follows:

1. The rays are always incident from the left on the refracting (or reflecting) surface.
2. All distances to the right of the point *P* are positive, and distances to the left of the point *P* are negative. Thus in Fig. 4.1(a), the object distance *u* is a negative quantity, and the image distance *v* and the radius of curvature *R* are positive quantities. For *u* to be positive, we must have a situation like the one shown in Fig. 4.1(b); in the absence of a refracting surface, the rays converge to a point to the *right* of *P*.
3. The angle that the ray makes with the axis is positive if the axis has to be rotated in the counterclockwise direction (through the acute angle) to coincide with the ray. Conversely, if the axis has to be rotated in the clockwise direction (through the acute angle) to coincide with the ray, then the slope angle is negative. Thus in Fig. 4.1(a) if θ_1 and θ_2 are the angles that the rays *OS* and *SI* make with the axis, then $\theta_1 = \alpha_1$ and $\theta_2 = -\alpha_2$; α_1, α_2, and β represent the magnitudes of the

angles. (If the final result does not depend on the angles, then it is more convenient to use the magnitude of the angles, as has indeed been done in Sec. 4.2.) In Fig. 4.1(b) both θ_1 and θ_2 are negative quantities.

4. The angle that a ray makes with the normal to the surface is positive if the normal has to be rotated in the counterclockwise direction (through the acute angle) to coincide with the ray, and conversely. Thus, in Fig. 4.1(a), ϕ_1 and ϕ_2 are positive quantities.
5. All distances measured upward from the axis (along a perpendicular to the axis) are positive, and all distances measured in the downward direction are negative.

4.2.2 The Gaussian Formula for a Single Spherical Surface

If we now use the sign convention discussed above, then for the ray diagram shown in Fig. 4.1(a), $u = -x$, $v = y$ and $R = r$. Thus Eq. (4) becomes

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R} \qquad \textbf{(5)}$$

which gives the image point due to refraction at a spherical surface [see also Eq. (10) of Chapter 3]. Equation (5) is known as the Gaussian formula for a single spherical surface. Note that corresponding to Fig. 4.1(a), u is negative and v positive, whereas for Fig. 4.1(b), u and v are both positive.

Example 4.1 Consider a medium of refractive index 1.5 bounded by two spherical surfaces $S_1P_1M_1$ and $S_2P_2M_2$ as shown in Fig. 4.2. The radii of curvature of the two surfaces are 15 and 25 cm with their centers at C_1 and C_2, respectively. There is an object at a distance of 40 cm (from P_1) on the line joining C_1 and C_2. Determine the position of the paraxial image.

Solution: We first consider refraction by $S_1P_1M_1$. Obviously $u = -40$ cm, $R = +15$ cm, $n_1 = 1.0$, and $n_2 = 1.5$. Thus

$$\frac{1.5}{v} + \frac{1}{40} = \frac{0.5}{15} \quad \Rightarrow \quad v = +180 \text{ cm}$$

In the absence of the second surface, the image is formed at O' at a distance of 150 cm from P_2. Now O' acts as a virtual object, and since it is to the right of $S_2P_2M_2$, we have, while considering refraction by the second surface, $u = +150$ cm, $R = -25$ cm, $n_1 = 1.5$, and $n_2 = 1.0$. Thus

$$\frac{1.0}{v} - \frac{1.5}{150} = +\frac{0.5}{25}$$

giving

$$v = +33\tfrac{1}{3} \text{ cm}$$

and a real image is formed on the right of P_2 at a distance of $33\frac{1}{3}$ cm.

While we consider refraction by a single surface (as in Fig. 4.1), the axis of the system is defined by the line joining the object point O and the center of curvature C. Thus *any* ray from the point O (like OS) will be in the plane containing the axis and the normal at the point S, and consequently, the refracted ray will *always* intersect the axis. On the other hand, if there is second refracting surface (as in Fig. 4.2 or as in a lens), then the line joining the two centers of curvature is defined as the axis. In the latter case, as can be readily seen, not all rays from an off-axis point will intersect the axis and after refraction at the second surface will, in general, not remain confined to a single plane; these rays are known as *skew rays*. Rays which remain confined to a plane (containing the axis) are known as *meridional rays*; obviously, all rays emanating from a point on the axis are meridional rays.

4.3 REFLECTION BY A SINGLE SPHERICAL SURFACE

We next consider the imaging of a point object O by a spherical mirror SPM (see Fig. 4.3) in the paraxial approximation; the point C represents the center of curvature. We proceed in a manner exactly similar to that in Sec. 4.2, and we refer to

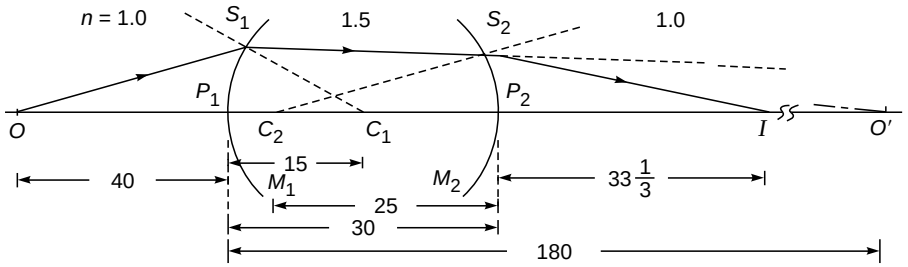


Fig. 4.2 Paraxial image formation by a medium of refractive index 1.5 bounded by two spherical surfaces $S_1P_1M_1$ and $S_2P_2M_2$. All distances are measured in centimeters.

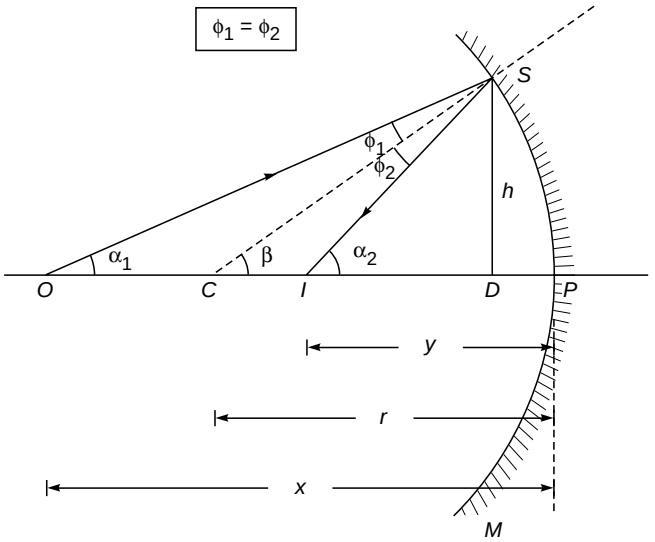


Fig. 4.3 Paraxial image formation by a spherical reflecting surface *SPM*.

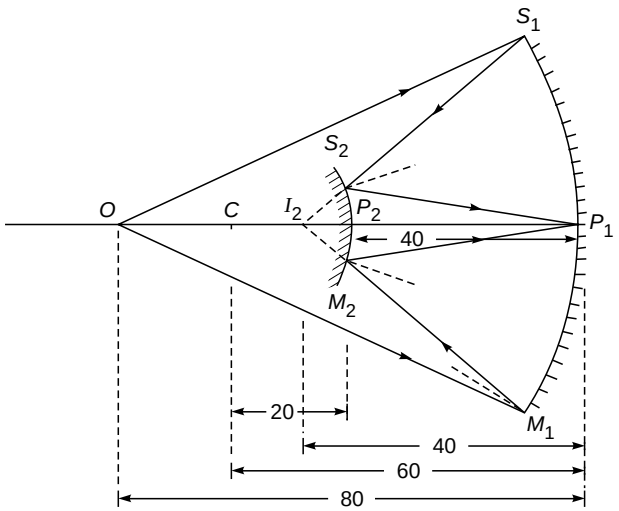


Fig. 4.4 Paraxial image formation by an optical system comprising a concave mirror $S_1P_1M_1$ and a convex mirror $S_2P_2M_2$.

Fig. 4.3 to obtain

$$\phi_1 = \beta - \alpha_1 \approx \frac{h}{r} - \frac{h}{x}$$

and

$$\phi_2 = \alpha_2 - \beta \approx \frac{h}{y} - \frac{h}{r}$$

where the distances x, y, h, and r are defined in Fig. 4.3. Since $\phi_1 = \phi_2$ (the law of reflection), we get

$$\frac{1}{x} + \frac{1}{y} = \frac{2}{r} \qquad \textbf{(6)}$$

If we again use the sign convention that all the distances to the right of P are positive and those to its left negative, then $u = -x$, $v = -y$, and $R = -r$; thus we obtain the mirror equation

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{R} \qquad \textbf{(7)}$$

which is the same as was derived by using Fermat's principle (see Prob. 3.1). If we set $n_2 = -n_1$ in Eq. (5), we get Eq. (7). This follows from the fact that Snell's law of refraction, Eq. (1), becomes the law of reflection if we have $n_2 = -n_1$.

We illustrate the use of Eq. (7) through an example.

Example 4.2 Consider an optical system consisting of a concave mirror $S_1P_1M_1$ and convex mirror $S_2P_2M_2$ of radii of curvatures 60 and 20 cm, respectively (see Fig. 4.4). We would like to determine the final image position of the object point O which is at a distance of 80 cm from the point P_1, the two mirrors being separated by a distance of 40 cm.

We first consider the imaging by $S_1P_1M_1$; since $u = -80$ cm and $R = -60$ cm (because both O and C are on the left of P_1), we have

$$-\frac{1}{80} + \frac{1}{v} = -\frac{2}{60} \quad \Rightarrow \quad v = -48 \text{ cm}$$

In the absence of the mirror $S_2P_2M_2$, a real image will be formed at I_1 which now acts as a virtual object for $S_2P_2M_2$. Since I_1 is to the left of P_2, we have (considering imaging by $S_2P_2M_2$) $u = -8$ cm and $R = -20$ cm, giving

$$\frac{1}{v} - \frac{1}{8} = -\frac{2}{20} \quad \Rightarrow \quad v = +40 \text{ cm}$$

Thus the final image is formed on the right of $S_2P_2M_2$ at a distance of 40 cm, which happens to be the point P_1.

4.4 THE THIN LENS

A medium bounded by two spherical refracting surfaces is referred to as a *spherical lens*. If the thickness of such a lens (shown as t in Fig. 4.5) is very small compared to the object and image distances and to the radii of curvature of the refracting surfaces, then the lens is referred to as a thin spherical lens. In general, a lens may have nonspherical refracting surfaces (e.g., it may have cylindrical surfaces). However, most lenses employed in optical systems have spherical refracting surfaces. Therefore, we will simply use the term *lens* to imply a spherical lens. Different types of lenses are shown in Fig. 4.6. The line joining the centers of curvature of the spherical refracting surfaces is referred to as the *axis* of the lens.

In this section, we will consider the paraxial image formation by a thin lens. The corresponding considerations for a thick lens will be discussed in Prob. 4.6.

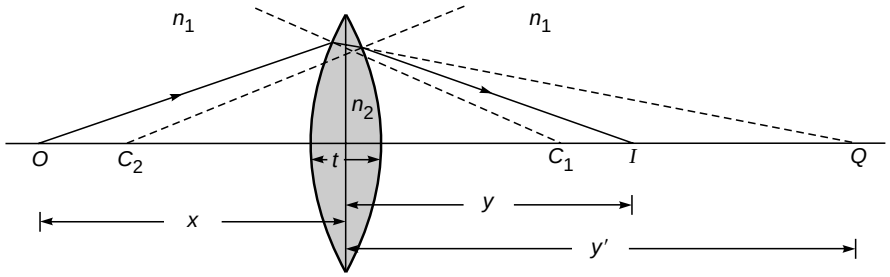


Fig. 4.5 Image formation by a thin lens. The line joining the two centers of curvature is known as the axis of the lens ($u = -x$, $v' = y'$, $v = y$).

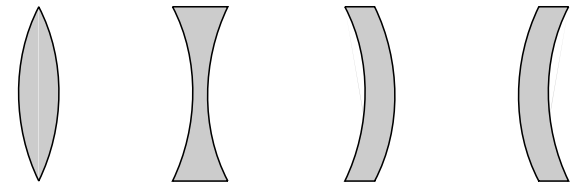


Fig. 4.6 Signs of R_1 and R_2 for different lens types.

We consider a point object O on the axis of a (thin) lens as shown in Fig. 4.5. The lens is placed in a medium of refractive index n_1, and the refractive index of the material of the lens is n_2. Let R_1 and R_2 be the radii of curvature of the left and right surfaces of the lens, respectively; for the lens shown in Fig. 4.5, R_1 is positive and R_2 is negative. To determine the position of the image, we will consider successive refractions at the two surfaces; the image formed by the first surface is considered the object (which may be real or virtual) for the second surface. Thus, if the second refracting surface had not been there, the image of the point O would have been formed at Q whose position (given by v') is determined from the following equation [see Eq. (5)]

$$\frac{n_2}{v'} - \frac{n_1}{u} = \frac{n_2 - n_1}{R_1} \qquad \textbf{(8)}$$

where u is the object distance which is negative for the object point O shown in the figure. Obviously if v' is positive, then the point Q lies to the right of the surface; and if v' is negative, then Q lies to the left of the surface. The point Q now acts as the (virtual) object for the second refracting surface, and the final image is formed at I whose position is determined from the equation

$$\frac{n_1}{v} - \frac{n_2}{v'} = \frac{n_1 - n_2}{R_2} \qquad \textbf{(9)}$$

In Eqs. (8) and (9) the distances are measured from the center of the lens P; this is justified because the lens has been assumed to be thin. Adding Eqs. (8) and (9), we get

$$\frac{1}{v} - \frac{1}{u} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad \textbf{(10)}$$

where

$$n \equiv \frac{n_2}{n_1}$$

Equation (10) is known as the *thin lens formula* and is usually written in the form

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f} \qquad \textbf{(11)}$$

where f, known as the focal length of the lens, is given by

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad \textbf{(12)}$$

For a lens placed in air (which is usually the case), $n > 1$ and if $1/R_1 - 1/R_2$ is a positive quantity, then the focal length is positive and the lens acts as a converging lens [see Fig. 4.7(a)]. Similarly, if $1/R_1 - 1/R_2$ is a negative quantity, then the lens acts as a diverging lens [see Fig. 4.7(b)]. However, if the double convex lens is placed in a medium whose refractive index is greater than that of the material of the lens, then the focal length becomes negative and the lens acts as a diverging lens [see Fig. 4.7(c)]; similarly for the double concave lens [see Fig. 4.7(d)].

4.5 THE PRINCIPAL FOCI AND FOCAL LENGTHS OF A LENS

For a converging lens, the *first principal focus* is defined as the point (on the axis) such that a ray passing through that point will, after refraction through the lens, emerge parallel to the axis—see ray 1 in Fig. 4.8(a); the point F_1 is the first principal focus. For a diverging lens, the ray which (in the absence of

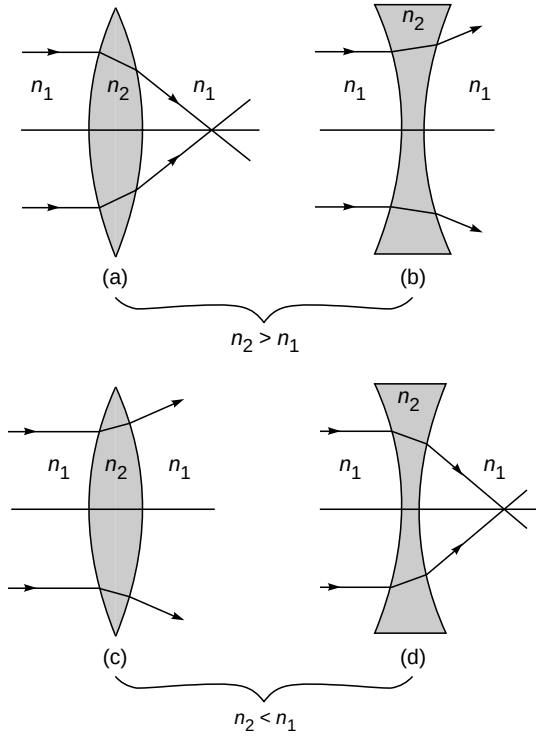


Fig. 4.7 (a) and (b) correspond to the situation when the refractive index of the material of the lens is greater than that of the surroundings, and therefore a biconvex lens acts as a converging lens and a biconcave lens acts as a diverging lens. (c) and (d) correspond to the situation when the refractive index of the material of the lens is smaller than that of the surrounding medium, and therefore a biconvex lens acts as a diverging lens and a biconcave lens as a converging lens.

the lens) would have passed through the first principal focus emerges, after refraction by the lens, as a ray parallel to the axis—see ray 1 in Fig. 4.8(b). Point F_1 is the *first principal focus*, and its distance from the lens (denoted by f_1) is known as the first focal length of the lens. Obviously, f_1 is negative for a converging lens and positive for a diverging lens.

We next consider a ray which travels parallel to the axis [see ray 2 in Fig. 4.8(a) and (b)]. For a converging lens the point at which the ray will intersect the axis [shown as F_2 in Fig. 4.8(a)] is known as the *second principal focus* of the lens. Similarly, for a diverging lens, the point at which the ray would have intersected the axis (if produced backward) is the second principal focus [see the point F_2 in Fig. 4.8(b)]. The distance of the second principal focus from the lens is known as the second focal length and is denoted by f_2. As can be seen from Fig. 4.8, f_2 is positive for a converging lens and negative for a diverging lens.

For a thin lens placed in a medium such that the refractive indices on both sides of the lens are the same ($n_3 = n_1$ in Fig. 4.8), the values of f_1 and f_2 can be readily obtained by considering the thin lens formula [see Eq. (10)] and one gets

$$\frac{1}{f_2} = -\frac{1}{f_1} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = \frac{1}{f} \qquad \textbf{(13)}$$

However, if $n_3 \neq n_1$, then the thin lens formula assumes the following form (see Prob. 4.2):

$$\frac{n_3}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R_1} + \frac{n_3 - n_2}{R_2} \qquad \textbf{(14)}$$

Now, when $v = \infty$, $u = f_1$ (ray 1 in Fig. 4.8) and we have

$$\frac{1}{f_1} = -\frac{1}{n_1}\left(\frac{n_2 - n_1}{R_1} + \frac{n_3 - n_2}{R_2}\right) \qquad \textbf{(15)}$$

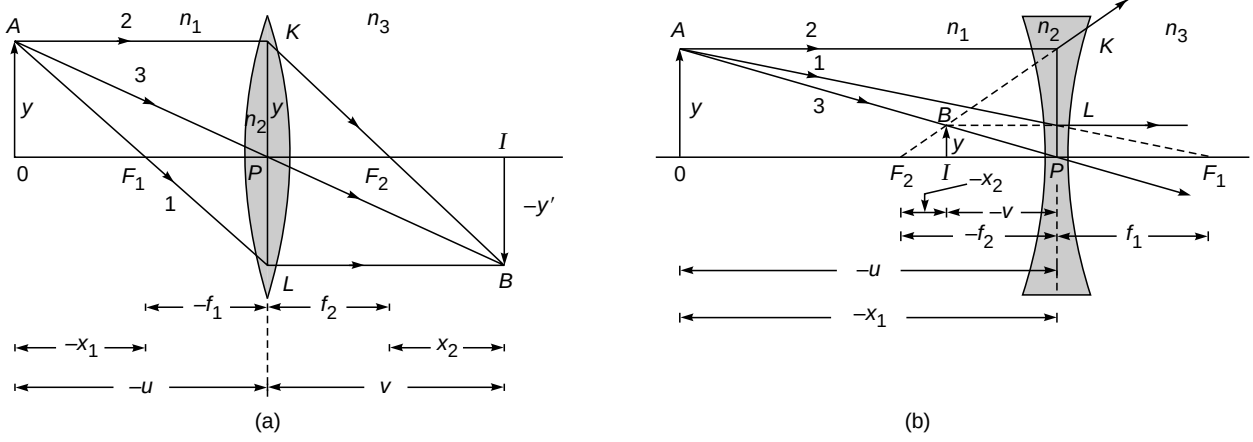


Fig. 4.8 (a) Paraxial imaging by a converging lens; x_1, f_1, and u are negative quantities and x_2, f_2, and v are positive quantities. (b) Paraxial imaging by a diverging lens; here x_1, f_2, u, and v are negative quantities and x_2 and f_1 are positive quantities.

Similarly, when $u = -\infty$, $v = f_2$ (ray 2 in Fig. 4.8), and we have

$$\frac{1}{f_2} = \frac{1}{n_3}\left(\frac{n_2 - n_1}{R_1} + \frac{n_3 - n_2}{R_2}\right) \quad \textbf{(16)}$$

Once we know f_1 and f_2 (and therefore the positions of the first and second principal foci), the (paraxial) image can be graphically constructed from the following rules:

1. A ray passing through the first principal focus will, after refraction, emerge parallel to the axis [see ray 1 in Fig. 4.8(a) and (b)].
2. A ray parallel to the axis will, after refraction, either pass through or appear to come from (depending on the sign of f_2) the second principal focus [see ray 2 in Fig. 4.8(a) and (b)].
3. A ray passing through the center of the lens P will pass through undeviated[2] [see ray 3 in Fig. 4.8(a) and (b)].

4.6 THE NEWTON FORMULA

Let x_1 be the distance of the object from the first principal focus F_1 (x_1 will be positive if the object point is on the right of F_1 and conversely), and let x_2 be the distance of the image from the second principal focus F_2 as shown in Fig. 4.8(a) and (b). Considering similar triangles in Fig. 4.8(a), we have

$$\frac{-y'}{y} = \frac{-f_1}{-x_1} \quad \textbf{(17)}$$

and

$$\frac{-y'}{y} = \frac{x_2}{f_2} \quad \textbf{(18)}$$

where the vertical distances are positive if measured above the line and negative if measured below the line (see Sec. 4.2.1). Equations (17) and (18) give

$$f_1 f_2 = x_1 x_2 \quad \textbf{(19)}$$

which is known as the *Newtonian lens formula*. It may be noted that for a diverging lens [see Fig. 4.8(b)], Eqs. (17) and (18) would be

$$\frac{y'}{y} = \frac{f_1}{-x_1} = \frac{x_2}{-f_2}$$

which are identical to Eqs. (17) and (18).

When the thin lens has the same medium on the two sides, then using Eq. (13), we have

$$x_1 x_2 = -f^2 \quad \textbf{(20)}$$

showing that x_1 and x_2 must be of opposite sign. Thus if the object lies on the left of the first principal focus, then the image will lie on the right of the second principal focus, and vice versa.

4.7 LATERAL MAGNIFICATION

The lateral magnification m is the ratio of the height of the image to that of the object. Considering either Fig. 4.8(a) or Fig. 4.8(b), we readily get

$$m = \frac{y'}{y} = \frac{v}{u} = \frac{f_2 + x_2}{f_1 + x_1} = -\frac{f_1}{x_1} = -\frac{x_2}{f_2} \quad \textbf{(21)}$$

where we have made use of Eqs. (17) and (18). Obviously, if m is positive, the image is erect [as in Fig. 4.8(b)], and conversely if m is negative, the image is inverted [as in Fig. 4.8(a)].

The magnification can also be calculated as the product of the individual magnifications produced by each of the refracting surfaces; referring to Fig. 4.9, the magnification produced by a single refracting surface is given by

$$m = \frac{y'}{y}$$

and considering triangles AOC and ICB, we get

$$\frac{-y'}{y} = \frac{v - R}{-u + R} = \frac{v/R - 1}{-u/R + 1} \quad \textbf{(22)}$$

Now, Eq. (5) gives us

$$\frac{n_2}{n_1} - \frac{v}{u} = \frac{n_2 - n_1}{n_1}\frac{v}{R}$$

and

$$\frac{u}{v} - \frac{n_1}{n_2} = \frac{n_2 - n_1}{n_2}\frac{u}{R}$$

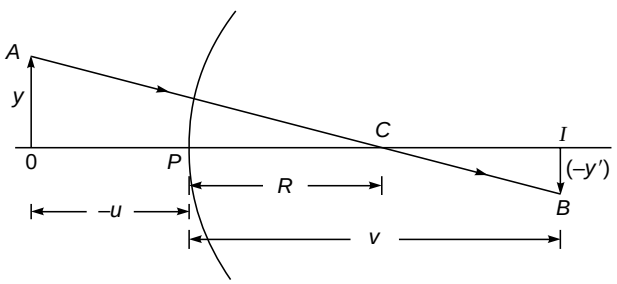


Fig. 4.9 Imaging of an object of height y by a spherical refracting surface.

[2] This follows from the fact that, for a thin lens, when $u = 0$, also v is equal to zero [see Eqs. (10) and (14)].

Substituting for v/R and u/R in Eq. (22), we get

$$m = \frac{y'}{y} = \frac{n_1 v}{n_2 u} \qquad \textbf{(23)}$$

Thus, if m_1 and m_2 represent the magnifications produced by the two refracting surfaces in Fig. 4.8, then

$$m = \frac{n_1}{n_2}\frac{v'}{u}$$

and

$$m_2 = \frac{n_2}{n_1}\frac{v}{v'}$$

where v' represents the distance of the image formed by the first refracting surface. Thus

$$m = m_1 m_2 = \frac{v}{u} \qquad \textbf{(24)}$$

consistent with Eq. (21).

Example 4.3 Consider a system of two thin lenses as shown in Fig. 4.10. The convex lens has a focal length of +20 cm, and the concave lens has a focal length of –10 cm. The two lenses are separated by 8 cm. For an object of height 1 cm (at a distance of 40 cm from the convex lens), calculate the position and size of the image. (The same problem will be solved again in Chap. 5 by using the matrix method.)

Solution: Let us first calculate the position and size of the image formed by the first lens:

$$u = -40 \text{ cm} \qquad f = +20 \text{ cm}$$

Therefore, using Eq. (11), we get

$$\frac{1}{v} = \frac{1}{u} + \frac{1}{f} = -\frac{1}{40} + \frac{1}{20} = +\frac{1}{40}$$

Thus, $v = +40$ cm and $m_1 = -1$; the image is of the same size but inverted. This image acts as a virtual object for the concave lens with $u = +32$ cm and $f = -10$ cm. Thus

$$\frac{1}{v} = \frac{1}{32} - \frac{1}{10} = -\frac{22}{320}$$

giving

$$v \simeq -14.5 \text{ cm}$$

Further,

$$m_2 = -\frac{320/22}{32} = -\frac{1}{2.2}$$

Thus

$$m = m_1 m_2 = +\frac{1}{2.2}$$

The final image is formed at a distance of 14.5 cm on the left of the concave lens. The image is virtual, erect, and smaller by a factor of 2.2.

4.8 APLANATIC POINTS OF A SPHERE

In Sec. 4.2, while discussing image formation by a single refracting surface, we made use of the paraxial approximation; i.e., we considered rays which made small angles with the axis. In this approximation, it was found that the images of point objects are perfect; i.e., *all* rays emanating from a given object point were found to intersect at *one* point which is the image point. If we had considered rays which make large angles with the axis, then we would have observed that in general (after refraction) they do not pass through the same point on the axis (see Fig. 4.11) and a perfect image is not formed. The image is said to be afflicted with aberrations. However, for a given spherical surface, there exist two points for which *all* rays emanating from one point intersect each other at the other point. This point is at a distance equal to $n_2|R|/n_1$ from the center of the spherical surface and a virtual image is formed at a distance of $n_1|R|/n_2$ from the center [see

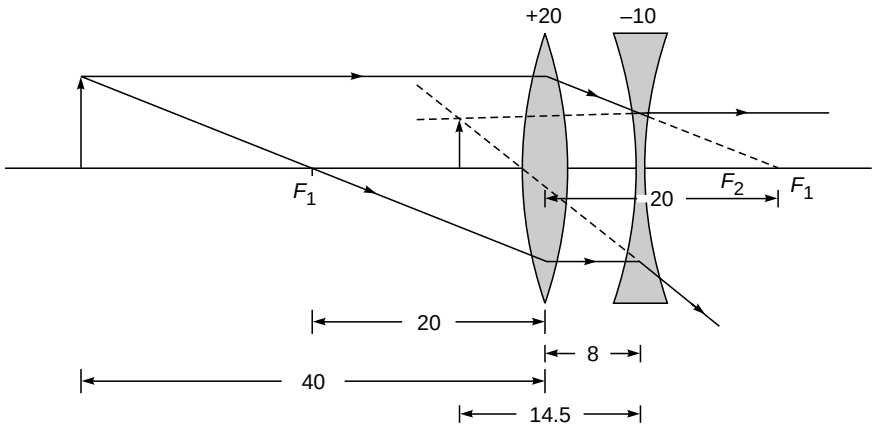


Fig. 4.10 Paraxial imaging by an optical system consisting of a converging lens of focal length 20 cm and a diverging lens of focal length –10 cm separated by 8 cm. All distances in the figure are in centimeters.

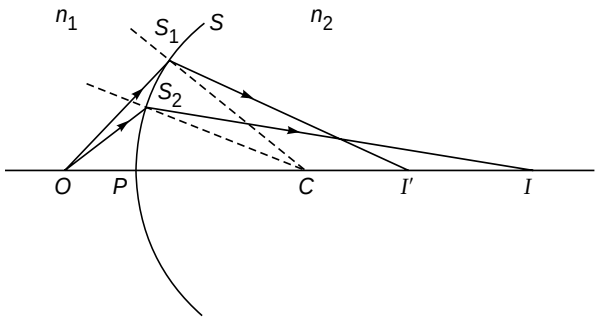


Fig. 4.11 The point I represents the paraxial image point of the object point O formed by a spherical refracting surface SPM. However, if we consider nonparaxial rays such as OS_1 (which make large angles with the axis), then the refracted ray, in general, will not pass through the point I —this leads to aberrations in the image.

Fig. 4.12(a) and (b)]. This can be easily proved by using Fermat's principle (see Prob. 3.7) or by using geometric methods (see Sec. 4.10). The two points are said to be the *aplanatic points* of the sphere and are utilized in the construction of aplanatic lenses (see Fig. 4.13) which are used in wide-aperture oil immersion microscope objectives. The points O and I are the aplanatic points of the spherical surface of radius R_2 (see Fig. 4.13). Thus

$$OP_2 = |R_2| \left(1 + \frac{n_1}{n_2}\right) \tag{25}$$

and

$$IP_2 = |R_2| \left(1 + \frac{n_2}{n_1}\right) \tag{26}$$

Now, the radius of curvature of the first surface (= R_1) is such that the point O coincides with its center of curvature. Hence *all* rays emanating from O hit the first surface normally and move on undeviated. Therefore, for all practical purposes, we may assume O to be embedded in a medium of refractive index n_2. A perfect (virtual) image of O is formed at I.

4.8.1 The Oil Immersion Objective

The principle of aplanatism has a very important application in microscope objectives where one is interested in having as wide a pencil of light as possible without causing any aberrations. We refer to the optical system shown in Fig. 4.14. The hemispherical lens L_1 is placed in contact with a drop of oil whose refractive index is the same as that of the lens. The object O is immersed in the oil, and the distance OC is made equal to $n_3|R_1|/n_2$ so that the point O is the aplanatic point with respect to the hemispherical surface, which is why a perfect (virtual) image is formed at I_1. Now L_2 is an aplanatic lens with respect to the object point at I_1, and therefore a perfect image of I_1 is formed at I. The lateral magnifications caused by the refracting surface R_1 and lens L_2 are

$$m_1 = \frac{n_2(I_1P_1)}{n_3(OP_1)} \tag{27}$$

and

$$m_2 = \frac{n_4(IP_3)}{n_5(I_1P_3)} \tag{28}$$

Thus the oil immersion objective reduces considerably the angular divergence of the rays and results in an increase in lateral magnification without introducing spherical aberration. We should, however, mention that a perfect image is formed

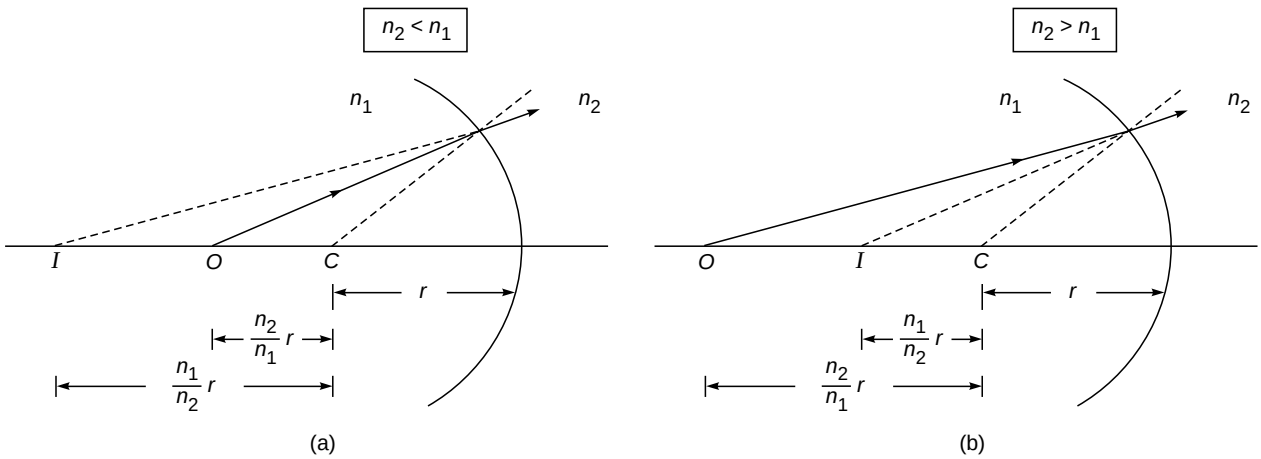


Fig. 4.12 Points O and I represent the aplanatic points of a spherical surface; i.e., *all* rays emanating from O appear to come from I; (a) and (b) correspond to $n_2 < n_1$ and $n_2 > n_1$, respectively.

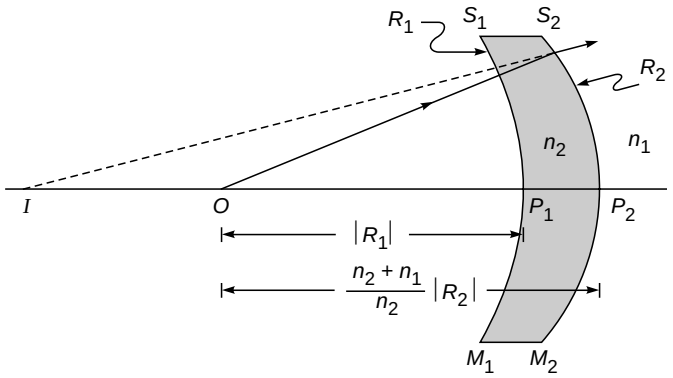


Fig. 4.13 The aplanatic lens. The object point O is at the center of curvature of the first surface $S_1P_1M_1$. The points O and I are the aplanatic points of the spherical surface $S_2P_2M_2$—thus a perfect (virtual) image is formed at I.

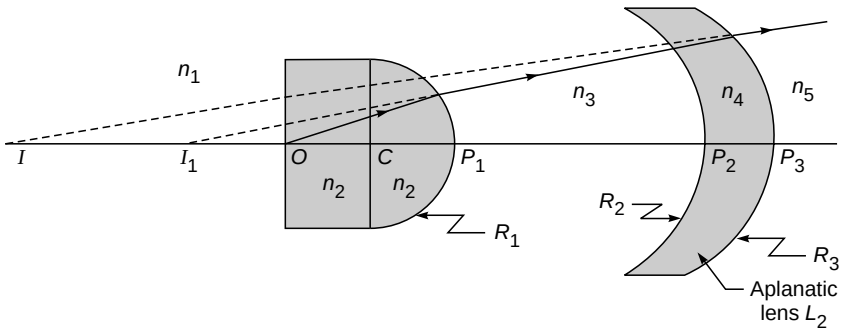


Fig. 4.14 The oil immersion objective. The points O and I_1 are the aplanatic points corresponding to the hemispherical surface of radius R_1; the lens L_2 acts as an aplanatic lens for the (virtual) object at I_1.

only of one point, and therefore nearby points have some aberrations. Moreover, oil immersion objectives have a certain degree of chromatic aberration.

4.9 THE CARTESIAN OVAL

In general, for two points to form perfect images of each other, the refracting surface should not be spherical. Figure 4.15 shows the two points O and I such that all rays emanating from O (and allowed by the system) intersect each other at the other point I. Thus the curve SPM shown in Fig. 4.15 is the locus of the point S such that

$$n_1OS + n_2SI = \text{constant} \qquad \textbf{(29)}$$

The refracting surface is obtained by revolving the curve shown in the figure about the z axis (see also Prob. 2.8). The refracting surface is known as a *Cartesian oval*. When the object point is at infinity, the surface becomes an ellipsoid of revolution (see Prob. 3.5), and under certain circumstances

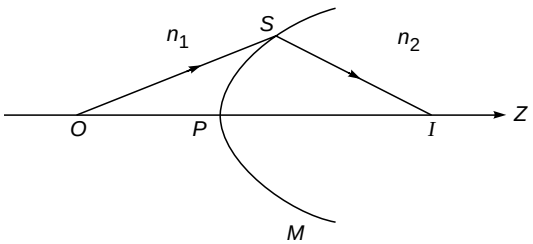


Fig. 4.15 The refracting surface (known as the Cartesian oval) is such that all rays emanating from the point O intersect at I.

the surface is spherical; however, the image is then virtual [see Fig. 4.12(a) and (b)].

4.10 GEOMETRICAL PROOF FOR THE EXISTENCE OF APLANATIC POINTS

In this section we will show the existence of aplanatic points using geometric considerations. We consider a spherical

refracting surface *SPM* of radius *r* separating two media of refractive indices n_1 and n_2 (see Fig. 4.16). We will assume $n_2 < n_1$ and define

$$\mu = \frac{n_1}{n_2} \tag{30}$$

where $\mu > 1$. The point *C* represents the center of the spherical surface *SPM*. With *C* as center, we draw two spheres of radii μr and r/μ, as shown in Fig. 4.16. Let *IOCP* represent *any* common diameter of the three spheres intersecting the outer and inner spheres at *I* and *O*, respectively. From the point *O*, we draw an arbitrary line hitting the refracting surface at the point *S*. We join *I* and *S* and extend the line farther as *SQ*. If we can show that

$$\frac{\sin\alpha}{\sin\beta} = \frac{1}{\mu} \tag{31}$$

for *all* values of θ_1, then *all* rays emanating from the point *O* will appear to come from *I*, and *O* and *I* will be the aplanatic points for the spherical refracting surface *SPM*. Now,

$$\frac{IC}{CS} = \frac{\mu r}{r} = \mu \tag{32}$$

and

$$\frac{CS}{OC} = \frac{r}{r/\mu} = \mu = \frac{IC}{CS} \tag{33}$$

Thus the two triangles *SOC* and *SIC* are similar, and therefore

$$\alpha = \theta_2 \quad \text{and} \quad \beta = \angle ISC = \theta_1 \tag{34}$$

Now, considering the triangle *SOC*, we have

$$\frac{\sin\alpha}{\sin\theta_1} = \frac{r/\mu}{r} = \frac{1}{\mu} \tag{35}$$

and using Eq. (34), we get

$$\frac{\sin\alpha}{\sin\beta} = \frac{1}{\mu} = \frac{n_2}{n_1} \tag{36}$$

proving that *O* and *I* are aplanatic points. We also have

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{\sin\beta}{\sin\alpha} = \frac{n_1}{n_2} \tag{37}$$

It is obvious that the points *O′* and *I′* will also be aplanatic and therefore the image formed by a *small* planar object at *O* will be sharp even for the off-axis points. The system is said to be free not only from spherical aberration but also from coma. Furthermore, the linear magnification is given by

$$m \approx \frac{I'I}{O'O} \approx \frac{\mu r}{r/\mu} = \mu^2 = \left(\frac{n_1}{n_2}\right)^2 \tag{38}$$

4.11 THE SINE CONDITION

We consider a general optical system as shown in Fig. 4.17. We *assume* that the point *O* (on the axis of the system) is

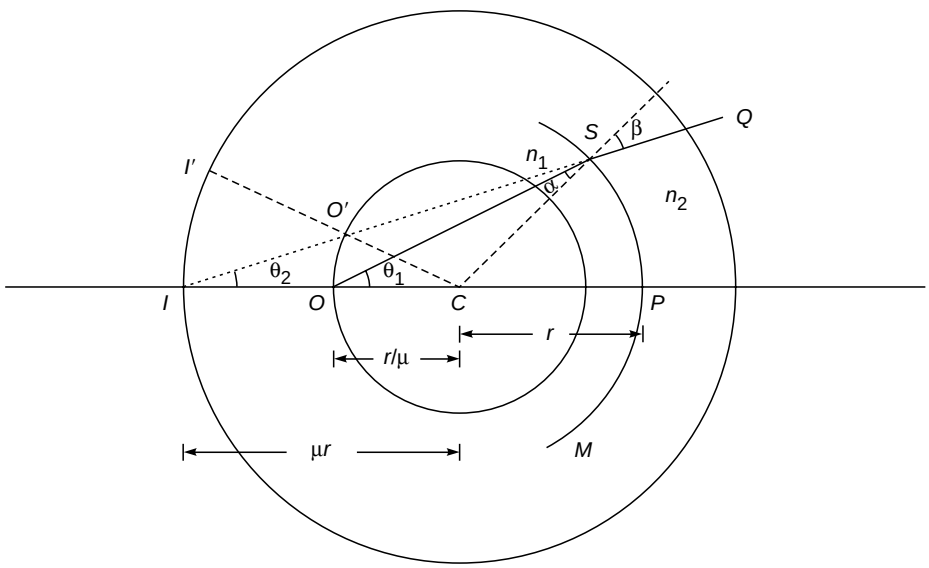


Fig. 4.16 Geometric construction for the derivation of aplanatic points. *SPM* is the refracting the surface of radius *r*. The inner and outer spheres are of radii r/μ and μr, respectively. Points *O* and *I* are the aplanatic points.

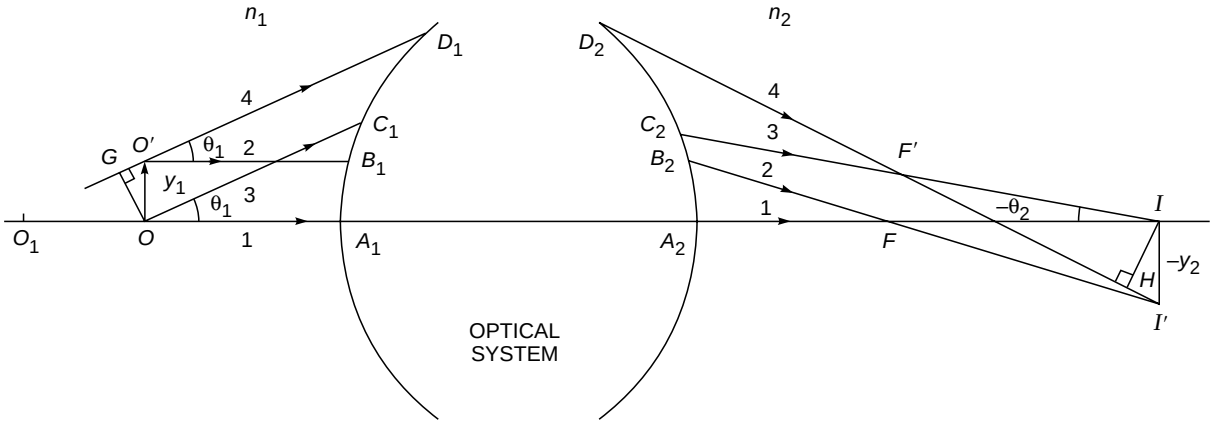


Fig. 4.17 The optical system images perfectly the points O and O' at I and I, respectively.

perfectly imaged at I; i.e. all rays emanating from O intersect each other at I. This implies that the optical system has no spherical aberration corresponding to O. We next consider a slightly off-axis point O' (directly above O), and according to the *sine condition*, for O' to be sharply imaged at I we must have[3]

$$\frac{n_1 \sin\theta_1}{n_2 \sin\theta_2} = \frac{y_2}{y_1} = \text{linear magnification} \qquad \textbf{(39)}$$

where θ_1 and θ_2 are defined in Fig. 4.17. Thus the linear magnification will be constant if the ratio $\sin\theta_1/\sin\theta_2$ is constant for all points on the refracting surface and the image will be free from the aberration known as coma. It is of interest to note that according to Eq. (39) perfect imaging of (nearby) off-axis points requires a condition to be satisfied by rays from an *on-axis point*.

4.11.1 Proof of the Sine Condition[4]

We refer to Fig. 4.17. We will assume that the axial point O is perfectly imaged at I and will use Fermat's principle to determine the condition for perfect imaging of the nearby off-axis point O'. The ray $O'B_1$ is parallel to the ray OA_1 and the ray $O'D_1$ is parallel to OC_1. Now, since I is the image of the point O, we have

$$\text{OPL}(OA_1A_2I) = \text{OPL}(OC_1C_2I) \qquad \textbf{(40)}$$

where OPL stands for the *optical path length*. Further

$$\text{OPL}(O'B_1B_2I') = \text{OPL}(O'D_1D_2I') \qquad \textbf{(41)}$$

Now, the rays $O'B_1$ and OA_1 meet at infinity and therefore

$$\text{OPL}(O'B_1B_2F) = \text{OPL}(OA_1A_2F) \qquad \textbf{(42)}$$

We next consider the triangle FII'.

$$FI' = (FI^2 + |y_2|^2)^{1/2} = FI\left(1 + \frac{1}{2}\frac{|y_2|^2}{FI^2}\right)$$

Thus

$$FI' \approx FI \qquad \textbf{(43)}$$

where we are assuming that $|y_2|$ is small enough that terms proportional to $|y_2|^2$ can be neglected. If we add Eqs. (42) and (43), we get

$$\begin{aligned}\text{OPL}(O'B_1B_2I') &= \text{OPL}(OA_1A_2I) \\ &= \text{OPL}(OC_1C_2I)\end{aligned} \qquad \textbf{(44)}$$

Since the left hand side of the above equation is OPL $(O'D_1D_2I')$, we get

$$\text{OPL}(O'D_1D_2I') = \text{OPL}(OC_1C_2I) \qquad \textbf{(45)}$$

Now the rays 3 and 4 meet at infinity and intersect at F', so that

$$\text{OPL}(GD_1D_2F') = OPL\,(OC_1CF') \qquad \textbf{(46)}$$

where the point G is the foot of the perpendicular drawn from the point O on ray 4. We subtract Eq. (45) from Eq. (46) to obtain

$$\text{OPL}(F'I') - \text{OPL}(GO') = \text{OPL}(F'I) \qquad \textbf{(47)}$$

[3] If we use Eqs. (37) and (38), we get

$$m = \frac{y_2}{y_1} = \left(\frac{n_1}{n_2}\right)^2 = \frac{n_1}{n_2}\frac{\sin\theta_1}{\sin\theta_2}$$

consistent with Eq. (39).

[4] For a rigorous proof of the sine condition, see Ref. 3.

or

$$n_2(F'I') - n_1(GO') = n_2(F'I)$$

or

$$n_1(GO') = n_2(F'I' - F'I) \qquad \textbf{(48)}$$

But

$$GO' = y_1 \sin\theta_1 \qquad \textbf{(49)}$$

and

$$F'I' - F'I \approx HI' \approx -y_2 \sin(-\theta_2) \qquad \textbf{(50)}$$

where H is the foot of the perpendicular from the point I on ray 4. Substituting the above two equations in Eq. (48), we get

$$\frac{n_1 \sin\theta_1}{n_2 \sin\theta_2} = \frac{y_2}{y_1} = \text{linear magnification} \qquad \textbf{(51)}$$

showing that the linear magnification is constant if the ratio $\sin\theta_1/\sin\theta_2$ is constant for all points on the refracting surface. The sine condition is of extensive use in the design of optical systems.

Summary

- Consider refraction at a spherical surface separating two media of refractive indices n_1 and n_2. For a point object at a distance $|u|$ on the left, the paraxial image is formed at a distance v where

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R}$$

  The sign convention is as follows:

  1. The rays are always incident from the left on the refracting surface.
  2. All distances to the right of the refracting surface are positive, and distances to the left of the refracting surface are negative.

- For a thin lens of refractive index n (placed in air), let R_1 and R_2 be the radii of curvature of the left and right surface respectively of the lens; then the image distance is given by

$$\frac{1}{v} - \frac{1}{u} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

  which is usually referred to as the *thin lens formula*; the quantity f is known as the focal length of the lens.

- For a given spherical surface, there are two points for which *all* rays emanating from one point intersect each other at the other point. This point is at a distance equal to $n_2|R|/n_1$ from the center of the spherical surface, and a virtual image is formed at a distance of $n_1|R|/n_2$ from the center. The two points are said to be the *aplanatic points* of the sphere and are utilized in the construction of aplanatic lenses.

- For two points to form perfect images of each other, the refracting surface is a Cartesian oval.

Problems

4.1 (a) Consider a thin biconvex lens (as shown in Fig. 4.18) made of a material whose refractive index is 1.5. The radii of curvature of the first and second surfaces (R_1 and R_2) are +100 and –60 cm, respectively. The lens is placed in air (i.e., $n_1 = n_3 = 1$). For an object at a distance of 100 cm from the lens, determine the position and linear magnification of the (paraxial) image. Also calculate x_1 and x_2 and verify Newton's formula [Eq. (20)].

[**Ans:** $x_1 = -25$ cm and $x_2 = +225$ cm]

(b) Repeat the calculations of part(a) when the object is at a distance of 50 cm.

4.2 Consider a thin lens (made of a material of refractive index n_2) having different media on the two sides; let n_1 and n_3 be the refractive indices of the media on the left and on the right of the lens, respectively. Using Eq. (5) and considering successive refractions at the two surfaces, derive Eq. (14).

4.3 Referring again to Fig. 4.18, assume a biconvex lens with $|R_1| = 100$ cm, $|R_2| = 60$ cm with $n_1 = 1.0$ but $n_3 = 1.6$. For $u = -50$ cm determine the position of the (paraxial) image. Also determine the first and second principal foci, and verify Newton's formula. Draw the ray diagram.

[**Ans:** $x_1 = 250$ cm, $x_2 = 576$ cm]

4.4 (a) In Fig. 4.18, assume the convex lens to be replaced by a (thin) biconcave lens with $|R_1| = 100$ cm and $|R_2| = 60$ cm. Assume $n_1 = n_3 = 1$ and $n_2 = 1.5$. Determine the position of the image and draw an approximate ray diagram for $u = -100$ cm.

(b) In (a), assume $n_1 = n_3 = 1.5$ and $n_2 = 1.3$. Repeat the calculations and draw the ray diagram. What is the qualitative difference between the systems in (a) and (b)?

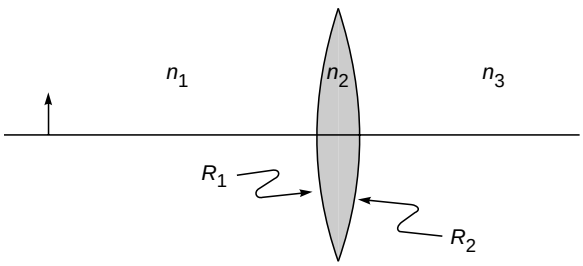


Fig. 4.18

4.5 Consider an object of height 1 cm placed at a distance of 24 cm from a convex lens of focal length 15 cm (see Fig. 4.19). A concave lens of focal length –20 cm is placed beyond the convex lens at a distance of 25 cm. Draw the ray diagram and determine the position and size of the final image.

[**Ans:** Real image at a distance of 60 cm from the concave lens.]

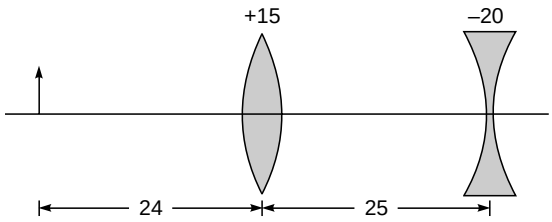


Fig. 4.19 An optical system consisting of a thin convex and a thin concave lens. All distances are measured in centimeters.

4.6 Consider a thick biconvex lens whose magnitudes of the radii of curvature of the first and second surfaces are 45 and 30 cm, respectively. The thickness of the lens is 5 cm, and the refractive index of the material that it is made of is 1.5. For an object of height 1 cm at distance of 90 cm from the first surface, determine the position and size of the image. Draw the ray diagram for the axial point of the object.

[**Ans:** Real image at a distance of 60 cm from the second surface.]

4.7 In Prob. 4.6 assume that the second surface is silvered so that it acts as a concave mirror. For an object of height 1 cm at a distance of 90 cm from the first surface, determine the position and size of the image and draw the ray diagram.

[**Ans:** Real image at a distance of about 6.2 cm from the first surface. (Remember the sign convention.)]

4.8 Consider a sphere of radius 20 cm of refractive index 1.6 (see Fig. 4.20). Show that the paraxial focal point is at a distance of 6.7 cm from the point P_2.

4.9 Consider a hemisphere of radius 20 cm and refractive index 1.5. Show that parallel rays will focus at a point 40 cm from P_2 (see Fig. 4.21).

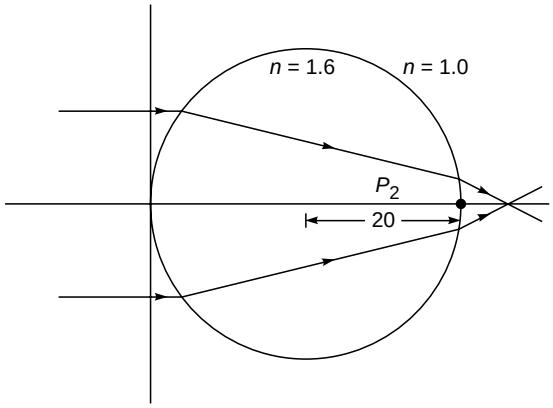


Fig. 4.20

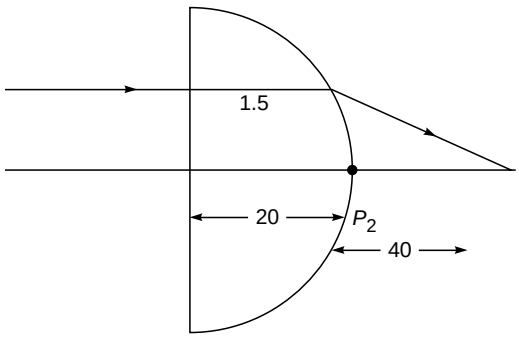


Fig. 4.21

4.10 Consider a lens of thickness 1 cm, made of a material of refractive index 1.5, placed in air. The radii of curvature of the first and second surfaces are +4 and –4 cm, respectively. Determine the point at which parallel rays will focus.

[**Ans:** At a distance of about 4.55 cm from the second surface]

REFERENCES AND SUGGESTED READINGS

1. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, 1975.
2. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley Publishing Co., Reading, Mass., 1965.
3. A. K. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978. [Reprinted by Macmillan India, New Delhi.]
4. R. H. Penfield, 'Consequences of Parameter Invariance in Geometrical Optics," *American Journal of Physics*., Vol. 24**,** p. 19, 1956.

Chapter Five

THE MATRIX METHOD IN PARAXIAL OPTICS

In dealing with a system of lenses we simply chase the ray through the succession of lenses. That is all there is to it.

—Richard Feynman, *Feynman Lectures on Physics*

5.1 INTRODUCTION[1]

Let us consider a ray PQ incident on a refracting surface SQS' separating two media of refractive indices n_1 and n_2 (see Fig. 5.1). Let NQN' denote the normal to the surface. The direction of the refracted ray is completely determined from the following conditions:

1. The incident ray, the refracted ray and the normal lie in the same plane,
2. If θ_1 and θ_2 represent the angles of incidence and refraction, respectively, then

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_2}{n_1} \tag{1}$$

Optical systems, in general, are made up of a large number of refracting surfaces (as in a combination of lenses), and any ray can be traced through the system by using the above conditions. To obtain the position of the final image due to such a system, one has to calculate step by step the position of the image due to each surface, and this image will act as an object for the next surface. Such a step-by-step analysis becomes complicated as the number of elements of an optical system increases. In this chapter, we develop the matrix method which can be applied with ease under such situations. This method indeed lends itself to direct use in computers for tracing rays through complicated optical systems.

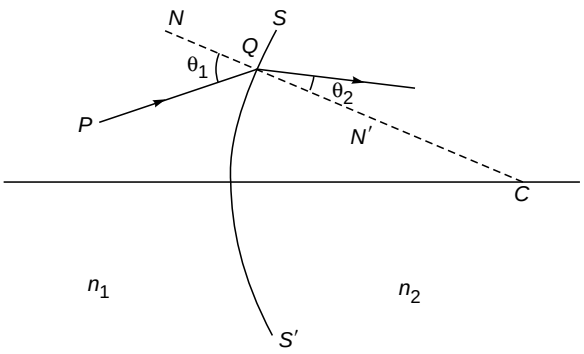


Fig. 5.1 Refraction of a ray by a surface SQS' which separates two media of refractive indices n_1 and n_2; NQN' denotes the normal at the point Q. If the refracting surface is spherical, then the normal NQN' will pass through the center of curvature C.

Before we describe the matrix formulation of geometric optics, it is necessary to mention the rule of matrix multiplication and the use of matrices for solving linear equations. An $m \times n$ matrix has m rows and n columns and has $m \times n$ elements; thus the matrix

$$A = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \tag{2}$$

has 2 rows and 3 columns and has $2 \times 3 = 6$ elements. A $m \times n$ matrix can be multiplied only by an $n \times p$ matrix to obtain an $m \times p$ matrix. Let

$$B = \begin{pmatrix} g \\ h \\ i \end{pmatrix} \tag{3}$$

represent a 2×1 matrix. Then the product

$$AB = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \begin{pmatrix} g \\ h \\ i \end{pmatrix} = \begin{pmatrix} (ag + bh + ci) \\ (dg + eh + fi) \end{pmatrix} \tag{4}$$

will be a 2×1 matrix, and the product BA has no meaning.

[1] The author thanks Prof. K. Thyagarajan for his help in writing this chapter.

If we define a 2×3 matrix

$$A' = \begin{pmatrix} a' & b' & c' \\ d' & e' & f' \end{pmatrix}$$

then $A' = A$

if and only if $a' = a$, $b' = b$, $c' = c$, $d' = d$, $e' = e$, and $f' = f$, i.e., all the elements must be equal. The set of two equations

$$\begin{aligned} x_1 &= ay_1 + by_2 \\ x_2 &= cy_1 + dy_2 \end{aligned} \quad \textbf{(5)}$$

can be written in the following form:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} (ay_1 + by_2) \\ (cy_1 + dy_2) \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \textbf{(6)}$$

the last step follows from the rule of matrix multiplication.

Further, if we have

$$\begin{aligned} y_1 &= ez_1 + fz_2 \\ y_2 &= gz_1 + hz_2 \end{aligned} \quad \textbf{(7)}$$

and

then

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} e & f \\ g & h \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad \textbf{(8)}$$

Consequently,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad \textbf{(9)}$$

or $$X = FZ, \quad \textbf{(10)}$$

where X and Z represent 2×1 matrices:

$$X \equiv \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad Z \equiv \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad \textbf{(11)}$$

and F represents a 2×2 square matrix

$$F = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} (ae + bg) & (af + bh) \\ (ce + dg) & (cf + dh) \end{pmatrix} \quad \textbf{(12)}$$

Equations (9) and (12) tell us that

$$\begin{aligned} x_1 &= (ae + bg)\, z_1 + (af + bh)\, z_2 \\ x_2 &= (ce + dg)\, z_1 + (df + dh)\, z_2 \end{aligned} \quad \textbf{(13)}$$

and

which can be verified by direct substitution. We will now use the matrix method to trace paraxial rays through a cylindrically symmetric optical system.

5.2 THE MATRIX METHOD

We will consider a cylindrically symmetric optical system similar to the one shown in Fig. 5.2. The axis of symmetry is chosen as the z axis. We will be considering only paraxial rays in this chapter; nonparaxial rays lead to what are known as aberrations, which will be discussed in Chap.6.

In the paraxial approximation, rays remain close to the optical axis, thus making small angles. Such a ray can be specified by its distance from the axis of the system and the angle made by the ray with the axis; for example, in Fig. 5.2, point P on the ray is at a distance x_1 from the axis and makes an angle α_1 with the axis. The quantities (x_1, α_1) represent the coordinates of the ray. However, instead of specifying the angle made by the ray with the z axis, we will specify the quantity.

$$\lambda = n \cos \psi \; (= n \sin \alpha)$$

which represents the product of the refractive index and the sine of the angle that the ray makes with the z axis, this quantity is known as the optical direction cosine.

Now, when a ray propagates through an optical system, it undergoes only two operations: translation and refraction. The rays undergo translation when they propagate through a homogeneous medium as in the region PQ (see Fig. 5.2). However, when a ray strikes an interface of two media, it undergoes refraction. We will now study the effect of translation and of refraction on the coordinates of the ray.

(a) Effect of Translation

Consider a ray traveling in a homogeneous medium of refractive index n_1 which is initially at a distance x_1 from the

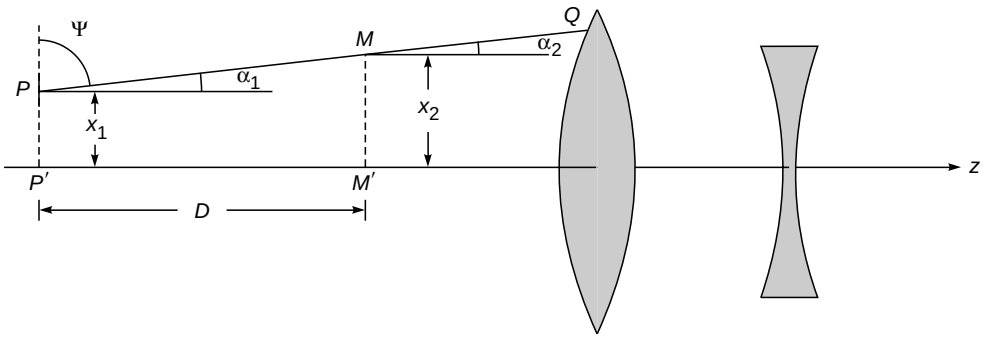


Fig. 5.2 In a homogeneous medium the ray travels in a straight line.

z axis and makes an angle α_1 with the axis (see point P in Fig. 5.2). Let (x_2, α_2) represent the coordinates of the ray at the point M (see Fig. 5.2). Since the medium is homogeneous, the ray travels in a straight line and, therefore,

$$\alpha_2 = \alpha_1 \tag{14}$$

Further, if PP' and MM' are perpendiculars on the axis and if $P'M' = D$, then

$$x_2 = x_1 + D \tan \alpha_1 \tag{15}$$

Since we are interested only in paraxial rays, α_1 is very small and hence we can make use of the approximation $\tan \alpha_1 \simeq \alpha_1$, where α_1 is measured in radians. Thus, Eq. (15) reduces to

$$x_2 \simeq x_1 + \alpha_1 D \tag{16}$$

If $$\lambda_1 = n_1 \alpha_1 \tag{17}$$

and $$\lambda_2 = n_2 \alpha_2 \tag{18}$$

then, using Eqs. (15) and (17), we get

$$\lambda_2 = \lambda_1$$

and

$$x_2 = x_1 + \frac{D}{n_1}\lambda_1 \tag{19}$$

which may be combined into the following matrix equation:

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ D/n_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \tag{20}$$

Thus, if a ray is initially specified by a 2×1 matrix with elements λ_1 and x_1, then the effect of translation through a distance D in a homogeneous medium of refractive index n_1 is completely given by the 2×2 matrix

$$T = \begin{pmatrix} 1 & 0 \\ D/n_1 & 1 \end{pmatrix} \tag{21}$$

and the final ray is given by Eq. (20). The matrix T is known as the translation matrix. Notice that

$$\det T = \begin{vmatrix} 1 & 0 \\ D/n_1 & 1 \end{vmatrix} = 1 \tag{22}$$

(b) Effect of Refraction

We will now determine the matrix which would represent the effect of refraction through a spherical surface of radius of curvature R. Consider the ray AP intersecting a spherical surface (separating two media of refractive indices n_1 and n_2, respectively) at point P and getting refracted along PB (see Fig. 5.3). If θ_1 and θ_2 are the angles made by the incident and the refracted ray with the normal to the surface at P (i.e., with the line joining P to the center of curvature C), then according to Snell's law

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \tag{23}$$

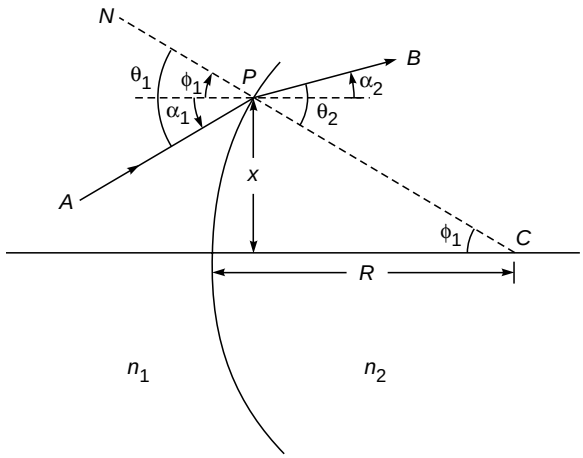


Fig. 5.3 The refraction of a ray at a spherical surface.

Since we are dealing with paraxial rays, we can use the approximation $\sin \theta \simeq \theta$. Thus Eq. (23) reduces to

$$n_1 \theta_1 \simeq n_2 \theta_2 \tag{24}$$

From Fig. 5.3 it follows that

$$\theta_1 = \phi_1 + \alpha_1 \quad \text{and} \quad \theta_2 = \phi_1 + \alpha_2 \tag{25}$$

where α_1, α_2, and ϕ_1 are, respectively, the angles that the incident ray, the refracted ray, and the normal to the surface make with the z axis. Also since ϕ_1 is small, we may write

$$\phi_1 = \frac{x}{R} \tag{26}$$

Now, from Eqs. (24) and (25), we get

$$n_1(\phi_1 + \alpha_1) \simeq n_2(\phi_1 + \alpha_2)$$

or

$$n_2 \alpha_2 \simeq n_1 \alpha_1 - \frac{n_2 - n_1}{R} x \tag{27}$$

where we have used Eq. (26). Thus

$$\lambda_2 = \lambda_1 - Px \tag{28}$$

where

$$P = \frac{n_2 - n_1}{R} \tag{29}$$

is known as the *power* of the refracting surface. Also since the height of the ray at P before and after refraction is the same (i.e., $x_2 = x_1$), we obtain for the refracted ray

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \tag{30}$$

Thus, refraction through a spherical surface can be characterized by a 2×2 matrix:

$$\mathfrak{R} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix} \tag{31}$$

Note that

$$\det \Re = \begin{vmatrix} 1 & -P \\ 0 & 1 \end{vmatrix} = 1 \qquad \textbf{(32)}$$

In general, an optical system made up of a series of lenses can be characterized by the refraction and translation matrices.

If a ray is specified by $\begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$ when it enters an optical system and is specified by $\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix}$ when it leaves the system, then we can, in general, write

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \qquad \textbf{(33)}$$

where the matrix

$$S = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \qquad \textbf{(34)}$$

is called the *system matrix* and is determined solely by the optical system. The negative signs in some of the elements of S have been chosen for convenience. Since the only two operations a ray undergoes in traversing an optical system are refraction and translation, the system matrix is, in general, a product of refraction and translation matrices. Also, using the property that the determinant of the product of matrices is the product of the determinant of the matrices, we obtain

$$\det S = 1 \qquad \textbf{(35)}$$

i.e.,

$$bc - ad = 1 \qquad \textbf{(36)}$$

The quantities b and c are dimensionless. The quantities a and P have the dimension of inverse length, and the quantity d has the dimension of length. In general, the units will not be given; however, it will be implied that a and P are in cm^{-1} and d is in cm.

5.2.1 Imaging by a Spherical Refracting Surface

As a simple illustration of the use of the matrix method, we consider imaging by a spherical surface separating two media of refractive indices n_1 and n_2 (see Fig. 5.4); the same problem was discussed in Chap. 4 using the standard geometrical method. Let (λ_1, x_1), (λ', x'), (λ'', x''), and (λ_2, x_2) represent the coordinates of the ray at O, A' (just before refraction), A'' (just after refraction), and I respectively.

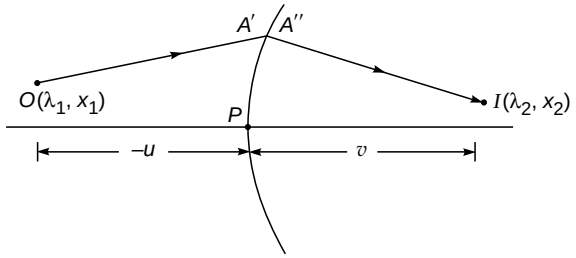


Fig. 5.4 Imaging by a spherical refracting surface separating two media of refractive indices n_1 and n_2.

We will be using the analytical geometry sign convention so that the coordinates on the left of the point P are negative and coordinates on the right of P are positive (see Sec. 4.2.1). Thus

$$\begin{pmatrix} \lambda' \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -u/n_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

$$\begin{pmatrix} \lambda'' \\ x'' \end{pmatrix} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda' \\ x' \end{pmatrix}$$

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ v/n_2 & 1 \end{pmatrix} \begin{pmatrix} \lambda'' \\ x'' \end{pmatrix}$$

or

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ v/n_2 & 1 \end{pmatrix} \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -u/n_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

Simple manipulations give

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 + \dfrac{Pu}{n_1} & -P \\ \dfrac{v}{n_2}\left(1 + \dfrac{Pu}{n_1}\right) - \dfrac{u}{n_1} & \left(1 - \dfrac{vP}{n_2}\right) \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \qquad \textbf{(37)}$$

from which we obtain

$$x_2 = \left[\frac{v}{n_2}\left(1 + \frac{Pu}{n_1}\right) - \frac{u}{n_1}\right]\lambda_1 + \left(1 - \frac{vP}{n_2}\right)x_1 \qquad \textbf{(38)}$$

For a ray emanating from an axial object point (i.e., for $x_1 = 0$) the image plane is determined by the condition $x_2 = 0$. Thus in the above equation, the coefficient of λ_1 should vanish and therefore

$$\frac{u}{n_1} = \frac{v}{n_2}\left(1 + \frac{Pu}{n_1}\right)$$

or

$$\frac{n_2}{v} - \frac{n_1}{u} = P = \frac{n_2 - n_1}{R} \qquad \textbf{(39)}$$

which is the same as derived Chap. 4. Hence, on the image plane

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1+\dfrac{Pu}{n_1} & -P \\ 0 & 1-\dfrac{vP}{n_2} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \qquad \textbf{(40)}$$

giving

$$x_2 = \left(1-\frac{vP}{n_2}\right) x_1$$

Thus the magnification is given by

$$m = \frac{x_2}{x_1} = 1 - \frac{vP}{n_2}$$

which on using Eq. (39) gives

$$m = \frac{n_1 v}{n_2 u}$$

consistent with Eq. (23) of Chap. 4.

5.2.2 Imaging by a Coaxial Optical System

We will next derive the position of the image plane for an object plane, which is at distance $-D_1$ from the first refracting surface of the optical system (see Fig. 5.5). Let the image be formed at a distance D_2 from the last refracting surface. Now, according to our sign convention, for points on the left of a refracting surface, the distances will be negative, and for points on the right of the refracting surface the distances will be positive; thus D_1 is an intrinsically negative quantity. Further, if D_2 is found to be positive, the image is real and is formed on the right of the refracting surface; on the other hand, if D_2 is found to be negative, the image will be virtual and will be formed on the left of the last refracting surface.

Let us consider a ray $O'P$ starting from point O' which lies in the object plane. Let QI' be the ray emerging from the last surface; point I' is assumed to lie on the image plane—see Fig. 5.5 (point I is the paraxial image of point O, and image plane is defined to be the plane which contains point I and is normal to the axis). Let (λ_1, x_1), (λ', x'), (λ'', x''), and (λ_2, x_2) represent the coordinates of the ray at O', P, Q, and I' respectively. Then

$$\begin{pmatrix} \lambda' \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -D_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

$$\begin{pmatrix} \lambda'' \\ x'' \end{pmatrix} = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} \lambda' \\ x' \end{pmatrix}$$

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ D_2 & 1 \end{pmatrix} \begin{pmatrix} \lambda'' \\ x'' \end{pmatrix}$$

Thus

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ D_2 & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -D_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \qquad \textbf{(41)}$$

where the first and the third matrices on the right-hand side correspond to translations by distances D_2 and $-D_1$, respectively (in a medium of refractive index unity); the second matrix corresponds to the system matrix of the optical system. Carrying out the matrix multiplications, we obtain

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} b + aD_1 & -a \\ bD_2 + aD_1D_2 - cD_1 - d & c - aD_2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \qquad \textbf{(42)}$$

Thus

$$x_2 = (bD_2 + aD_1D_2 - cD_1 - d)\lambda_1 + (c - aD_2)x_1$$

For a ray emanating from the axial object point (i.e., for $x_1 = 0$) the image plane is determined by the condition $x_2 = 0$. Thus, for the image plane we must have

$$bD_2 + aD_1D_2 - cD_1 - d = 0 \qquad \textbf{(43)}$$

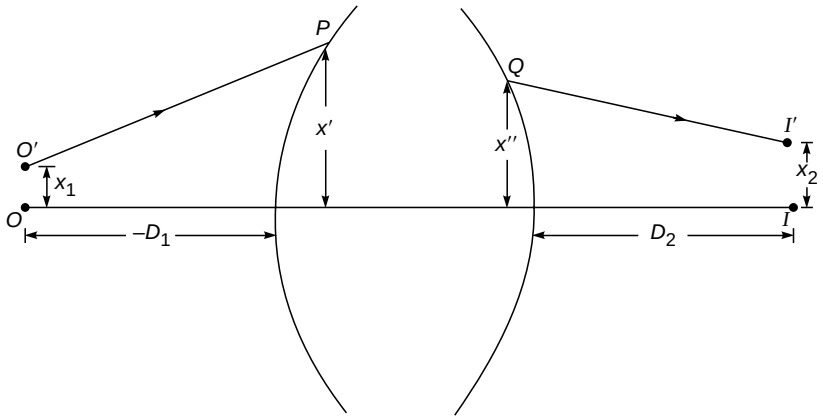


Fig. 5.5 The object point O is at a distance $(-D_1)$ from the first refracting surface. The paraxial image is assumed to be formed at a distance D_2 from the last refracting surface.

which would give us the relationship between the distances D_1 and D_2. Thus, corresponding to the image plane, we have

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} b + aD_1 & -a \\ 0 & c - aD_2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad \textbf{(44)}$$

For $x_2 \neq 0$, we obtain

$$x_2 = (c - aD_2)x_1$$

Consequently, the magnification of the system $M\left(= x_2/x_1\right)$ would be given by

$$M = \frac{x_2}{x_1} = c - aD_2 \quad \textbf{(45)}$$

Further, since

$$\begin{vmatrix} b + aD_1 & -a \\ 0 & c - aD_2 \end{vmatrix} = 1$$

we obtain

$$b + aD_1 = \frac{1}{c - aD_2} = \frac{1}{M} \quad \textbf{(46)}$$

Hence, if x_1 and x_2 correspond to object and image planes, then for a general optical system we may write

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1/M & -a \\ 0 & M \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad \textbf{(47)}$$

Example 5.1 Obtain the system matrix for a thick lens, and derive the thin lens and thick lens formulas.

Solution: Let us consider a lens of thickness t and made of a material of relative refractive index n (see Fig. 5.6). Let R_1 and R_2 be the radii of curvature of the two surfaces. The ray is assumed to strike the first surface of the lens at P and emerge from point Q; let the coordinates of the ray at P and Q be

$$\begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} \quad \textbf{(48)}$$

where λ_1 and λ_2 are the optical direction cosines of the ray at P and Q; x_1 and x_2 are the distances of points P and Q from the axis (see Fig. 5.6). The ray, in propagating from P to Q, undergoes two refractions [one at the first surface (whose radius of curvature is R_1) and the other at the second surface (whose radius of curature is R_2)] and a translation through a distance[2] t in a medium of refractive index n. Thus

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -P_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t/n & 1 \end{pmatrix} \begin{pmatrix} 1 & -P_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad \textbf{(49)}$$

where

$$P_1 = \frac{n-1}{R_1} \quad \text{and} \quad P_2 = \frac{1-n}{R_2} = -\frac{n-1}{R_2} \quad \textbf{(50)}$$

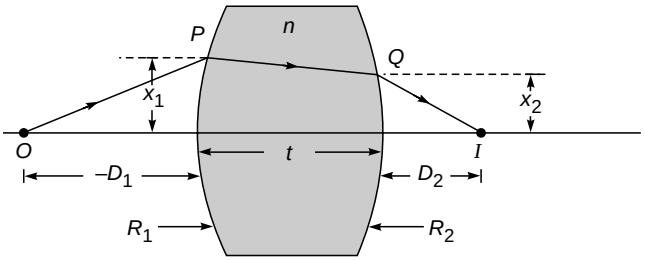


Fig. 5.6 A paraxial ray passing through a thick lens of thickness t.

represent the powers of the two refracting surfaces. Thus our system matrix is given by

$$S = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} = \begin{pmatrix} 1 & -P_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t/n & 1 \end{pmatrix} \begin{pmatrix} 1 & -P_1 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 - \frac{P_2 t}{n} & -P_1 - P_2\left(1 - \frac{t}{n}P_1\right) \\ \frac{t}{n} & 1 - \frac{t}{n}P_1 \end{pmatrix} \quad \textbf{(51)}$$

For a thin lens, $t \to 0$ and the system matrix takes the following form:

$$S = \begin{pmatrix} 1 & -P_1 - P_2 \\ 0 & 1 \end{pmatrix} \quad \textbf{(52)}$$

Thus for a thin lens,

$$a = P_1 + P_2 \qquad b = 1 \qquad c = 1 \qquad d = 0 \quad \textbf{(53)}$$

Substituting the above values of a, b, c, and d in Eq. (43), we obtain

$$D_2 + (P_1 + P_2)\, D_1 D_2 - D_1 = 0$$

or

$$\frac{1}{D_2} - \frac{1}{D_1} = P_1 + P_2$$

$$= (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \quad \textbf{(54)}$$

or

$$\frac{1}{D_2} - \frac{1}{D_1} = \frac{1}{f} \quad \textbf{(55)}$$

where

$$f = \frac{1}{P_1 + P_2} = \left[(n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)\right]^{-1} \quad \textbf{(56)}$$

represents the focal length of the lens. Equation (55) is the well-known thin lens formula. (The signs of R_1 and R_2 for different kinds of lenses are shown in Fig. 5.7). Thus the system matrix for a thin lens is given by

$$\mathrm{S} = \begin{pmatrix} 1 & -\frac{1}{f} \\ 0 & 1 \end{pmatrix} \quad \textbf{(57)}$$

[2] Notice that since we are dealing with paraxial rays, the distance between P and Q is approximately t.

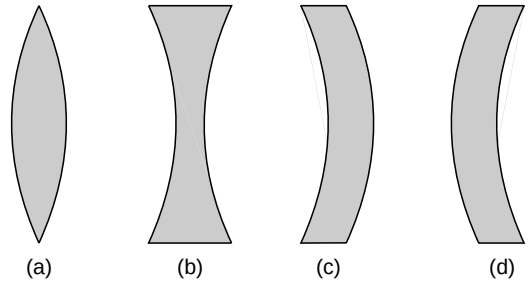


Fig. 5.7 Signs of R_1 and R_2 for different lens types.

For a thick lens, we have from Eq. (51)

$$a = P_1 + P_2\left(1 - \frac{t}{n}P_1\right) \qquad b = 1 - \frac{P_2 t}{n}$$
$$c = 1 - \frac{t}{n}P_1 \qquad d = -\frac{t}{n} \tag{58}$$

If we substitute the above values for a, b, c, and d in Eq. (43), we get the required relation between D_1 and D_2; however, for thick lenses it is more convenient to define the unit and the nodal planes which we shall do in the following sections.

5.3 UNIT PLANES

The unit planes are two planes, one each in the object and the image space, between which the magnification M is unity; i.e., any paraxial ray emanating from the unit plane in the object space will emerge at the same height from the unit plane in the image space. Thus, if d_{u1} and d_{u2} represent the distances of the unit planes from the refracting surfaces (see Fig. 5.8)[3] we obtain from Eq. (46)

$$b + ad_{u1} = \frac{1}{c - ad_{u2}} = 1 \tag{59}$$

or
$$d_{u1} = \frac{1-b}{a} \tag{60}$$

$$d_{u2} = \frac{c-1}{a} \tag{61}$$

Hence the unit planes are determined completely by the elements of the system matrix S.

It will be convenient to measure distances from the unit planes. Thus if u is the distance of the object plane from the first unit plane and v is the distance of the corresponding image plane from the second unit plane (see Fig. 5.8), we obtain

$$D_1 = u + d_{u1} = u + \frac{1-b}{a} \tag{62}$$

and

$$D_2 = v + d_{u2} = v + \frac{c-1}{a} \tag{63}$$

Now, from Eq. (43) we have

$$D_2 = \frac{d + cD_1}{b + aD_1} \tag{64}$$

Substituting for D_1 and D_2 from Eqs. (62) and (63), we get

$$v + \frac{c-1}{a} = \frac{d + cu + c(1-b)/a}{b + au + (1-b)}$$

or
$$v = \frac{ad - bc + c(au+1) - (c-1)(1+au)}{a(1+au)}$$
$$= \frac{au}{a(1+au)} \tag{65}$$

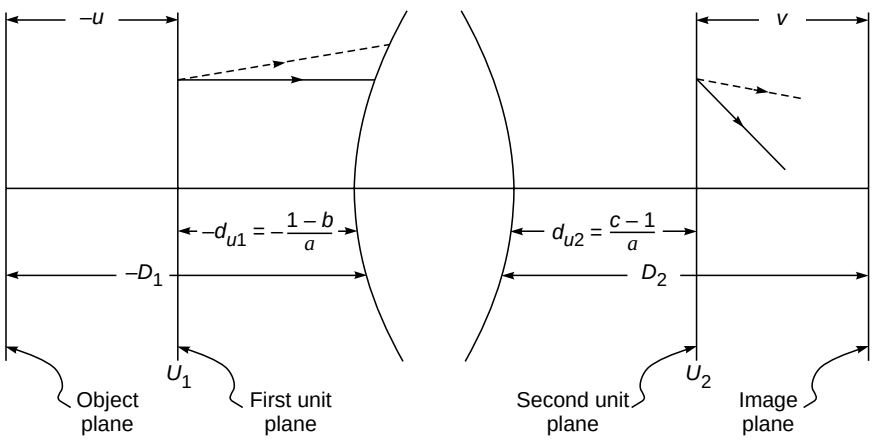


Fig. 5.8 Planes U_1 and U_2 are the two unit planes. A ray emanating at any height from the first unit plane will cross the second unit plane at the same height.

[3] Obviously, if we consider U_1 as an object plane, then U_2 is the corresponding image plane.

where we have used the condition that

$$\det S = bc - ad = 1 \quad \textbf{(66)}$$

On simplification, we obtain

$$\frac{1}{v} - \frac{1}{u} = a \quad \textbf{(67)}$$

Thus $1/a$ represents the focal length of the system if the distances are measured from the two unit planes. For example, for a thick lens one obtains [using Eqs. (58), (60). and (61)]

$$d_{u1} = \frac{P_2 t}{n} \frac{1}{P_1 + P_2\left[1 - (t/n)P_1\right]} \quad \textbf{(68)}$$

and

$$d_{u2} = -\frac{t}{n} \frac{P_1}{P_1 + P_2\left[1 - (t/n)P_1\right]} \quad \textbf{(69)}$$

For a thick double convex lens with $|R_1| = |R_2|$

$$P_1 = P_2 = \frac{n-1}{R} \quad \textbf{(70)}$$

where $R = |R_1| = |R_2|$. Thus

$$d_{u1} = \frac{t}{n} \frac{1}{2 - \frac{t}{n}\frac{n-1}{R}} \simeq \frac{t}{2n} \quad \textbf{(71)}$$

and

$$d_{u2} = -\frac{t}{n} \frac{1}{2 - \frac{t}{n}\frac{n-1}{R}} \simeq -\frac{t}{2n} \quad \textbf{(72)}$$

where we have assumed $t << R$ which is indeed the case for most thick lenses. The positions of the unit planes are shown in Fig. 5.9. To calculate the focal length, we note from Eq. (67) that

$$\frac{1}{f} = a = P_1 + P_2\left(1 - \frac{t}{n}P_1\right) \quad \textbf{(73)}$$

where we have used Eq. (58). Thus

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{(n-1)^2 t}{n R_1 R_2} \quad \textbf{(74)}$$

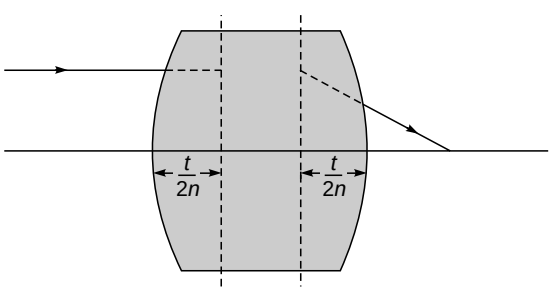


Fig. 5.9 Unit planes of a thick biconvex lens.

5.4 NODAL PLANES

Nodal points are two points on the axis which have a relative angular magnification of unity; i.e., a ray striking the first point at an angle α emerges from the second point at the same angle (see Fig. 5.10). The planes which pass through these points and are normal to the axis are known as nodal planes.

To determine the position of the nodal points, we consider two axial points N_1 and N_2 at distances d_{n1} and d_{n2} from the two refracting surfaces, respectively (see Fig. 5.10). From the definition of nodal points, we require that a ray incident at an angle α_1 on the point N_1 emerge from the optical system at the same angle α_1 from the other point N_2. Since we have assumed the media on either side of the system to have the same refractive index, this condition requires the equality of λ_1 and λ_2. Also, since we are considering an axial object point $x_1 = 0$, we get from Eq. (44)

$$\lambda_2 = (b + ad_{n1})\lambda_1 = \lambda_1 \quad \textbf{(75)}$$

Thus

$$b + ad_{n1} = 1 \quad \textbf{(76)}$$

or

$$d_{n1} = \frac{1-b}{a} \quad \textbf{(77)}$$

Comparing this with Eq. (60), we find that $d_{n1} = d_{u1}$. This has arisen because of the equality of the indices of refraction on either side of the optical system. Similarly we can get

$$d_{n2} = \frac{c-1}{a} \quad \textbf{(78)}$$

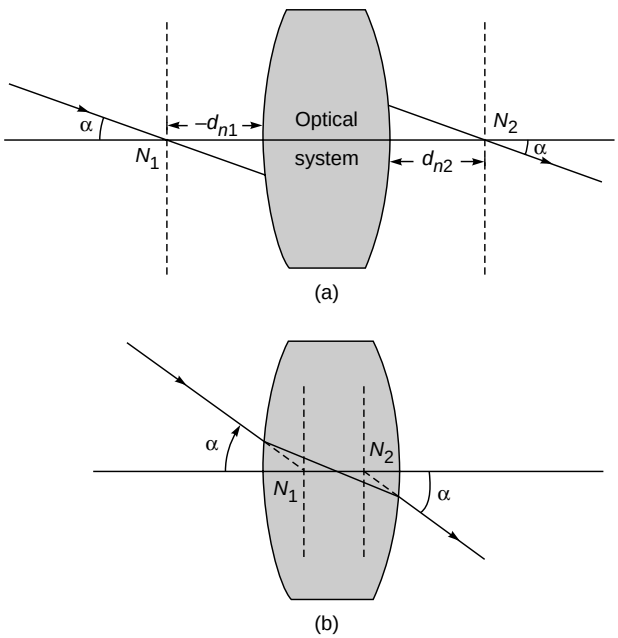


Fig. 5.10 Points N_1 and N_2 denote the two nodal points of an optical system. The nodal points can also lie inside the optical system, as shown in (b).

Thus, when the media on either side of an optical system have the same refractive index (which is indeed the case for most optical systems), the nodal planes coincide with the unit planes. In general, if we know the elements of the system matrix S (i.e., if we know a, b, c, and d which are also called the Gaussian constants of the system), we can obtain all the properties of the system.

Example 5.2 Consider a thick equiconvex lens (made of a material of refractive index 1.5) of the type shown in Fig. 5.9. The magnitudes of the radii of curvature of the two surfaces are 4 cm. The thickness of the lens is 1 cm, and the lens is placed in air. Obtain the system matrix, and determine the focal length and the positions of unit planes.

Solution:

$$R_1 = +4 \text{ cm} \qquad R_2 = -4 \text{ cm} \qquad t = 1 \text{ cm}$$

Both surfaces have equal power

$$P_1 = P_2 = \frac{n-1}{R_1} = \frac{0.5}{4} = 0.125 \text{ cm}^{-1}$$

Thus the system matrix is, from Eq. (51),

$$\begin{pmatrix} 1-\dfrac{0.125\times 1}{1.5} & -0.125-0.125\left(1-\dfrac{1}{1.5}\times 0.125\right) \\ \dfrac{1}{1.5} & 1-\dfrac{0.125}{1.5} \end{pmatrix}$$

$$= \begin{pmatrix} 0.9167 & -0.240 \\ 0.6667 & 0.9167 \end{pmatrix}$$

Thus

$$a = \frac{1}{f} = 0.24 \quad \Rightarrow \quad f \simeq 4.2 \text{ cm}$$

$$b = 0.9167 = c \qquad d = -0.6667$$

Using Eqs. (60) and (61), we get the positions of the unit planes

$$d_{u1} = \frac{1-b}{a} \simeq 0.35 \text{ cm}$$

$$d_{u2} = \frac{c-1}{a} \simeq -0.35 \text{ cm}$$

Thus the unit planes are as shown in Fig. 5.9. The nodal planes coincide with the unit planes because the lens is immersed in air.

Example 5.3 Consider a sphere of radius 20 cm of refractive index 1.6 (see Fig. 5.11). Find the positions of the paraxial focal point and the unit planes.

Solution: The matrices from the first refracting surface to the image plane are given by

Second surface to image	Refraction at second surface	Transmission through glass	Refraction at the first surface

$$\begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix} \begin{pmatrix} 1 & (1-1.6)/20 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 40/1.6 & 1 \end{pmatrix} \begin{pmatrix} 1 & -(1.6-1)/20 \\ 0 & 1 \end{pmatrix}$$

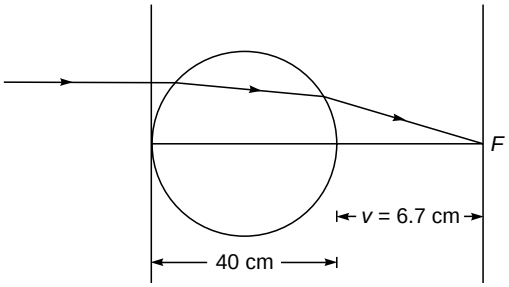


Fig. 5.11 Imaging by a sphere of radius 20 cm and refractive index 1.6.

$$= \begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix} \begin{pmatrix} 0.25 & -0.0375 \\ 25 & 0.25 \end{pmatrix}$$

$$= \begin{pmatrix} 0.25 & -0.0375 \\ 25+0.25v & 0.25-0.0375v \end{pmatrix}$$

Thus at the image plane, the ray coordinates are

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.25 & -0.0375 \\ 25+0.25v & 0.25-0.0375v \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

This gives us

$$x_2 = (25 + 0.25v)\lambda_1 + (0.25 - 0.0375v)\, x_1$$

To determine the focal distance v, consider a ray incident parallel to the axis for which $\lambda_1 = 0$. The focal plane would be that plane for which x_2 is also zero. This gives us

$$0.0375v = 0.25 \quad \text{or} \quad v = 6.7 \text{ cm}$$

The system matrix elements are

$$a = \frac{1}{f} = 0.0375 \text{ cm}^{-1} \quad \Rightarrow \quad f \simeq 26.7 \text{ cm}$$

$$b = 0.25 \qquad c = 0.25 \qquad d = -25 \text{ cm}$$

The unit planes are given by

$$d_{u1} = \frac{1-b}{a} = 20 \text{ cm}$$

and

$$d_{u2} = \frac{c-1}{a} = -20 \text{ cm}$$

Thus both the unit planes pass through the center of the sphere. In this example, we cannot use the approximation $t << R$.

5.5 A SYSTEM OF TWO THIN LENSES

We finally use the matrix formulation for the analysis of a combination of two thin lenses of focal lengths f_1 and f_2 separated by a distance t. The system matrix for the combination of the

two lenses can be obtained by noting that the matrices of the two lenses are [see Eq. (57)]

$$\begin{pmatrix} 1 & -\dfrac{1}{f_1} \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & -\dfrac{1}{f_2} \\ 0 & 1 \end{pmatrix} \tag{79}$$

and the matrix for translation through a distance t (in air is) is

$$\begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} \tag{80}$$

Thus the system matrix S is given by

$$S = \begin{pmatrix} 1 & -\dfrac{1}{f_2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} \begin{pmatrix} 1 & -\dfrac{1}{f_1} \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \left(1 - \dfrac{t}{f_2}\right) & -\left(\dfrac{1}{f_1} + \dfrac{1}{f_2} - \dfrac{t}{f_1 f_2}\right) \\ t & \left(1 - \dfrac{t}{f_1}\right) \end{pmatrix} \tag{81}$$

Thus

$$a = \frac{1}{f_1} + \frac{1}{f_2} - \frac{t}{f_1 f_2} \qquad b = 1 - \frac{t}{f_2}$$
$$c = 1 - \frac{t}{f_1} \qquad d = -t \tag{82}$$

As already noted, the element a in the system matrix represents the inverse of the focal length of the system. Thus, the focal length of the combination is

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{t}{f_1 f_2} = a \tag{83}$$

The positions of the unit planes are given by [see Eqs. (60) and (61)]

$$d_{u1} = \frac{1-b}{a} = \frac{tf}{f_2}$$
$$d_{u2} = \frac{c-1}{a} = -\frac{tf}{f_1} \tag{84}$$

It is easy to see that if we have a system of four thin lenses, we simply have to multiply seven matrices [four of them being of the type given by Eq. (79) and three of them of the type given by Eq. (80)].

Example 5.4 Consider a lens combination consisting of a convex lens (of focal length +15 cm) and a concave lens (of focal length –20 cm) separated by 25 cm (see Fig. 5.12 and Prob. 4.5). Determine the system matrix elements and the positions of the unit planes. For an object (of height 1 cm) placed at a distance of 27.5 cm from the convex lens, determine the size and position of the image.

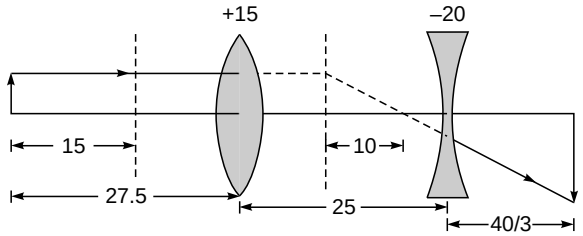


Fig. 5.12

Solution:

$f_1 = +15$ cm $\quad f_2 = -20$ cm $\quad t = 25$ cm

Thus, using Eq. (82), we readily get

$$a = \frac{1}{10} = \frac{1}{f} \qquad b = \frac{45}{20} \qquad c = -\frac{2}{3} \qquad d = -25$$

and

$$d_{u1} = \frac{1-b}{a} = -12.5 \text{ cm} \qquad d_{u2} = \frac{c-1}{a} = -\frac{50}{3} \text{ cm}$$

Thus the distance of the object from the first unit plane is given by

$$u = -27.5 - (-12.5) = -15 \text{ cm}$$

Since $f = +10$ cm, we get [using Eq. (67)]

$$v = 30 \text{ cm}$$

which represents the distance of the image plane from the second unit plane. Thus the image is at a distance of 30 –50/3 = 40/3 cm from the concave lens. The magnification is given by

$$M = \frac{v}{u} = -2$$

Example 5.5 Consider a system of two thin lenses as shown in Fig. 4.10. For a 1 cm tall object at a distance of 40 cm from the convex lens, calculate the position and size of the image.

Solution: Let v be the distance of the image plane from the concave lens. Thus the matrix, which when operated on the object column matrix gives the image column matrix, is given by

Concave lens to image	Concave lens	Convex lens to concave lens	Convex lens	Object to convex lens
$\begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & +1/10 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 8 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & -1/20 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 40 & 1 \end{pmatrix}$

$$= \begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix} \begin{pmatrix} 2.2 & 0.01 \\ +32 & 0.6 \end{pmatrix}$$

$$= \begin{pmatrix} 2.2 & 0.01 \\ 2.2v + 32 & 0.6 + 0.01v \end{pmatrix}$$

The image plane would correspond to

$$32 + 2.2v = 0$$

or

$$v \simeq -14.5 \text{ cm}$$

i.e., it is at a distance of 14.5 cm to the left of the concave lens. If we compare this with Eq. (45), we obtain

$$M = 0.6 + 0.01v = 0.6 - 0.01\left(\frac{32}{2.2}\right) = +\frac{1}{2.2}$$

Example 5.6 In Example 5.5, determine the system matrix and hence the positions of the unit planes. Finally, use Eq. (67) to determine the position of the image.

Solution: The system matrix is given by

$$S = \begin{pmatrix} 1 & 1/10 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 8 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1/20 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 9/5 & 1/100 \\ 8 & 3/5 \end{pmatrix}$$

Thus

$$a = -\frac{1}{100} \quad \Rightarrow \quad f = -100 \text{ cm}$$

$$b = \frac{9}{5} \qquad c = \frac{3}{5} \qquad d = -8$$

If we now use Eqs. (60 and (61), we have

$$d_{u1} = \frac{1-b}{a} = 80 \text{ cm}$$

and

$$d_{u2} = \frac{c-1}{a} = 40 \text{ cm}$$

Thus the first unit plane is at a distance of 80 cm to the right of the convex lens, and the second unit plane is at 40 cm to the right of the concave lens. The object distance from the first unit plane is therefore given by

$$u = -(80 + 40) = -120 \text{ cm}$$

We now use Eq. (67) to obtain

$$\frac{1}{v} = a + \frac{1}{u} = -\frac{1}{100} - \frac{1}{120} = -\frac{22}{1200}$$

$$\Rightarrow \qquad v = -\frac{600}{11} \text{ cm}$$

Thus the image is at 54.5 cm to the left of the second unit plane or at 14.5 cm to the left of the concave lens, as shown in Fig. 4.10. The magnification is

$$M = \frac{v}{u} = +\frac{1}{2.2}$$

Summary

- In the paraxial approximation we may confine ourselves to rays which pass through the axis of the system; these rays remain confined to a single plane. Such a ray can be specified by its distance from the axis of the system x and the quantity $\lambda = n \sin \alpha$, which represents the product of the refractive index with the sine of the angle that the ray makes with z axis.
- If a ray is initially specified by a 2×1 matrix with elements λ_1 and x_1, then the effect of translation through a distance D in a homogeneous medium of refractive index n_1 is given by

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = T\begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

  where the translation matrix T is given by

$$T = \begin{pmatrix} 1 & 0 \\ D/n_1 & 1 \end{pmatrix}$$

- The effect of refraction through a spherical refracting surface (separating media of refractive indices n_1 and n_2) is given by

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \mathcal{R}\begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

  where the refraction matrix is given by

$$\mathcal{R} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix}$$

  with

$$P = \frac{n_2 - n_1}{R}$$

- By successive application of the above matrices one can study paraxial imaging by a coaxial optical system.
- In an optical system, unit planes are two planes, one each in the object and the image space, between which the magnification M is unity; i.e., any paraxial ray emanating from the unit plane in the object space will emerge at the same height from the unit plane in the image space.
- Nodal points are two points on the axis which have a relative angular magnification of unity; i.e., a ray striking the first point at an angle α emerges from the second point at the same angle. The planes which pass through these points and are normal to the axis are known as nodal planes.

Problems

5.1 Consider a system of two thin convex lenses of focal lengths 10 and 30 cm separated by a distance of 20 cm in air.

(a) Determine the system matrix elements and the positions of the unit planes.

(b) Assume a parallel beam of light incident from the left. Use Eq. (67) and the positions of the unit planes to determine the image point. Using the unit planes, draw the ray diagram.

[**Ans:** (a) $a = 1/15$, $b = 1/3$, $c = -1$, $d = -20$; the first convex lens is in the middle of the two unit planes. (b) The final image is virtual and is 15 cm away (on the left) from the second lens.]

5.2 Consider a thick biconvex lens whose magnitudes of the radii of curvature of the first and second surfaces are 45

and 30 cm, respectively. The thickness of the lens is 5 cm, and the refractive index of the material of the lens is 1.5. Determine the elements of the system matrix and positions of the unit planes, and use Eq. (67) to determine the image point of an object at a distance of 90 cm from the first surface.

[**Ans:** a = 0.02716, b = 0.9444, c = 0.9630, d = –3.3333, d_{u1} = 2.0455, d_{u2} = –1.3636. Final image at a distance of 60 cm from the second surface.]

5.3 Consider a hemisphere of radius 20 cm and refractive index 1.5. If H_1 and H_2 denote the positions of the first and second principal points, respectively, then show that AH_1 = 13.3 cm and that H_2 lies on the second surface, as shown in Fig. 5.13. Further, show that the focal length is 40 cm.

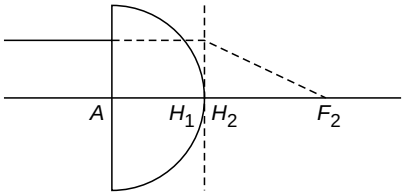


Fig. 5.13

5.4 Consider a thick lens of the form shown in Fig. 5.14; the radii of curvature of the first and second surfaces are

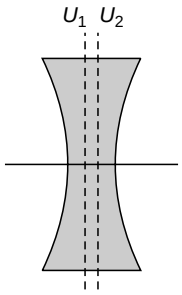


Fig. 5.14

–10 and +20 cm, respectively, and the thickness of the lens is 1.0 cm. The refractive index of the material of the lens is 1.5. Determine the positions of the principal planes.

[**Ans:** d_{u1} = 20/91 cm, d_{u2} = 40/91 cm]

5.5 Consider a combination of two thin lenses of focal lengths f_1 and f_2 separated by a distance $f_1 + f_2$. Show that the angular magnification of the lens combinations (which is just $\lambda_2/\lambda_1 = \alpha_2/\alpha_1$) is given by $-f_1/f_2$. Interpret the negative sign in the expression for magnification.

5.6 Consider a spherical refracting surface as shown in Fig. 4.12. Using the matrix method, show that for an object at a distance of $(1 + n_2/n_1)\, r$ from the surface, the image is virtual and at a distance of $(1 + n_1/n_2)\, r$ from the surface.

REFERENCES AND SUGGESTED READINGS

1. J. N. Blaker, *Geometrical Optics: The Matrix Theory,* Marcel Dekker, New York, 1971.
2. W. Brouwer, *Matrix Methods in Optical Instrumental Design,* Benjamin, New York, 1964.
3. D. M. Eakin and S. P. Davis, "An application of Matrix Optics," *American Journal of Physics,* Vol. 34, p. 758, 1966.
4. A. Gerrard and J. M. Burch, *Introduction to Matrix Methods in Optics,* John Wiley & Sons, New York, 1975.
5. K. Halbach, "Matrix Representation of Gaussian Optics," *American Journal of Physics,* Vol. 32, p. 90, 1964.
6. A. Nussbaum, *Geometric Optics: An Introduction,* Addison-Wesley Publishing Co., Reading, Mass., 1968.
7. J. W. Simmons and M. J. Guttmann, *States, Waves and Photons: A Modern Introduction to Light,* Addison-Wesley Publishing Co., Reading, Mass., 1970.

Chapter Six

ABERRATIONS

Geometrical optics is either very simple or else it is very complicated. . . . If one has an actual, detailed problem in lens design, including analysis of aberrations, then he has to simply trace the rays through the various surfaces using the law of refraction and find out where they come out and see if they form a satisfactory image. People have said that this is too tedious, but today, with computing machines, it is the right way to do it. One can set up the problem and make the calculation one ray after another very easily. So the subject is really ultimately quite simple, and involves no new principles.

—Richard Feynman, *Feynman Lectures on Physics*, Vol. I

6.1 INTRODUCTION

In Chap. 4, while studying the formation of images by refracting surfaces and thin lenses, we made the assumption that the object point does not lie far away from the axis of the optical system and that the rays taking part in image formation are essentially those which make small angles with the axis of the system. In practice, neither of the above assumptions is true; one in fact has to deal with rays making large angles with the axis. The domain of optics dealing with rays lying close to the optical axis and making small angles with it is called *paraxial optics*. We found that in the realm of paraxial optics, the images of objects were perfect; i.e., all rays emanating from a single object point converged to a single image point, and the magnification of the system was a constant of the optical system, independent of the particular ray under consideration. Since in real optical systems, nonparaxial rays also take part in image formation, the actual images depart from the ideal images. This departure leads to what are known as aberrations.

It can be shown that the primary aberrations of any rotationally symmetric system can be specified by five coefficients. The five coefficients represent the spherical aberration, coma, astigmatism, curvature of field, and distortion. These are called the Seidel aberrations. Since these aberrations are present even for light of a single wavelength, they are also called monochromatic aberrations. In this chapter, we will consider the five kinds of aberrations separately and discuss the effect on the image when each one of them is present separately.

Note that if a polychromatic source (such as white light) is used for image formation (which is indeed the case for many optical instruments), then, in general, the images will be colored; this is known as chromatic aberration. Physically, chromatic aberration is due to the dependence of the refractive index of the material of the lens on the wavelength of the radiation under consideration. Since image formation is accompanied by refraction at refractive index discontinuities, the wavelength dependence of the refractive index results in the colored image. For a polychromatic source, different wavelength components (after refraction) proceed along different directions and form images at different points; this leads to colored images. Since chromatic aberration is the easiest to understand, we discuss this first. This is followed by a discussion of monochromatic aberrations.

6.2 CHROMATIC ABERRATION

Let us consider a parallel beam of white light incident on a thin convex lens, as shown in Fig. 6.1. Since blue light gets refracted more than red light, the point at which the blue light will focus is nearer the lens than the point at which the red light will focus. Thus, the image will appear to be colored; note that this aberration is independent of the five Seidel aberrations discussed in later sections.

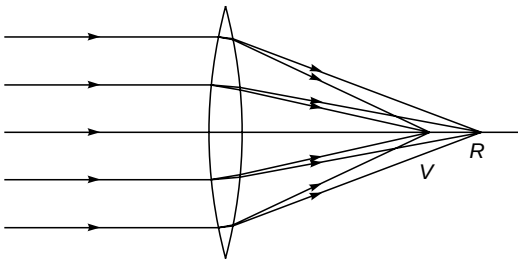


Fig. 6.1 When white light consisting of a continuous range of wavelengths is incident on a lens, each wavelength refracts by a different amount; this leads to chromatic aberration. This aberration is independent of the five Seidel aberrations.

For the case of a thin lens, the expression for chromatic aberration can be easily derived. The focal length of a thin lens is given by

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad \textbf{(1)}$$

If a change of n by δn (the change of n is due to the change in the wavelength of the light) results in a change of f by δf, then we obtain, by differentiating Eq. (1),

$$-\frac{\delta f}{f^2} = \delta n \left(\frac{1}{R_1} - \frac{1}{R_2}\right) = \frac{\delta n}{n-1}\frac{1}{f}$$

i.e., $$\delta f = -f\,\frac{\delta n}{n-1} \qquad \textbf{(2)}$$

which represents the chromatic aberration of a thin lens. If n_b and n_r represent the refractive indices for the blue and red colors respectively, then

$$f_r - f_b = f\left(\frac{n_b - n_r}{n-1}\right) \qquad \textbf{(3)}$$

represents the chromatic aberration.

6.2.1 The Achromatic Doublet

We first consider an optical system of two thin lenses made of different materials placed in contact with each other. For example, one of the lenses may be made of crown glass and the other of flint glass. We will find the condition for this lens combination to have the same focal length for the blue and red colors. Let n_b, n_y, and n_r represent the refractive indices for the material of the first lens corresponding to the blue, yellow, and red colors, respectively. Similarly, n_b', n_y', and n_r' represent the corresponding refractive indices for the second lens. If f_b and f_b' represent the focal lengths for the first and second lenses corresponding to the blue color, and if F_b represents the focal length of the combination of the two lenses (placed in contact), then

$$\frac{1}{F_b} = \frac{1}{f_b} + \frac{1}{f_b'} = (n_b - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + (n_b' - 1)\left(\frac{1}{R_1'} - \frac{1}{R_2'}\right) \qquad \textbf{(4)}$$

where R_1 and R_2 represent the radii of curvature of the first and second surfaces for the first lens and, as before, the primed quantities refer to the second lens. Thus, we may write

$$\frac{1}{F_b} = \frac{n_b - 1}{n-1}\frac{1}{f} + \frac{n_b' - 1}{n' - 1}\frac{1}{f'} \qquad \textbf{(5)}$$

where

$$\frac{1}{f} \equiv (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

$$\frac{1}{f'} \equiv (n'-1)\left(\frac{1}{R_1'} - \frac{1}{R_2'}\right) \qquad \textbf{(6)}$$

$$n \equiv \frac{n_b + n_r}{2} \approx n_y \qquad n' \equiv \frac{n_b' + n_r'}{2} \approx n_y' \qquad \textbf{(7)}$$

and f and f' represent the focal lengths of the first and second lenses corresponding to a mean color which is around the yellow region. Similarly, the focal length of the combination corresponding to the red color is given by

$$\frac{1}{F_r} = \frac{n_r - 1}{n-1}\frac{1}{f} + \frac{n_r' - 1}{n' - 1}\frac{1}{f'} \qquad \textbf{(8)}$$

For the focal length of the combination to be equal for blue and red colors, we must have

$$\frac{n_b - 1}{n-1}\frac{1}{f} + \frac{n_b' - 1}{n' - 1}\frac{1}{f'} = \frac{n_r - 1}{n-1}\frac{1}{f} + \frac{n_r' - 1}{n' - 1}\frac{1}{f'}$$

or $$\frac{\omega}{f} + \frac{\omega'}{f'} = 0 \qquad \textbf{(9)}$$

where $$\omega = \frac{n_b - n_r}{n-1} \quad \text{and} \quad \omega' = \frac{n_b' - n_r'}{n' - 1} \qquad \textbf{(10)}$$

are known as the dispersive powers. Since ω and ω' are both positive, f and f' must have opposite signs for the validity of Eq. (9). A lens combination which satisfies Eq. (9) is known as an *achromatic doublet* (see Fig. 6.2). If the two lenses are made of the same material, then $\omega = \omega'$ and Eq. (9) would imply $f = -f'$; such a combination will have an infinite focal length. Thus, for an achromatic doublet the two lenses must be of different materials.

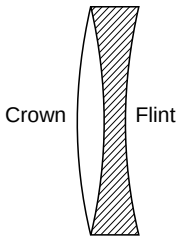


Fig. 6.2 An achromatic doublet.

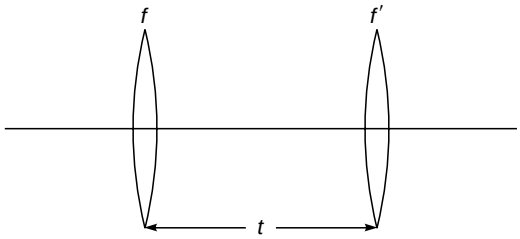


Fig. 6.3 The separated doublet.

Example 6.1 An achromatic doublet of focal length 20 cm is to be made by placing a convex lens of borosilicate crown glass in contact with a diverging lens of dense flint glass. Assuming n_r = 1.51462, n_b = 1.52264, n_r' = 1.61216, and n_b' = 1.62901, calculate the focal length of each lens; here the unprimed and the primed quantities refer to the borosilicate crown glass and dense flint glass, respectively.

Solution:

$$n \approx \frac{n_b + n_r}{2} = \frac{1.52264 + 1.51462}{2} = 1.51863$$

$$n' \approx \frac{n_b' + n_r'}{2} = \frac{1.62901 + 1.61216}{2} = 1.62058$$

Thus,

$$\omega = \frac{1.52264 - 1.51462}{1.51863 - 1} = 0.01546$$

and

$$\omega' = \frac{1.62901 - 1.61216}{1.62058 - 1} = 0.02715$$

Substituting in Eq. (9), we obtain

$$\frac{0.01546}{f} + \frac{0.02715}{f'} = 0$$

or
$$\frac{f}{f'} = -0.56942$$

Now, for the lens combination to be of focal length 20 cm, we must have

$$\frac{1}{f} + \frac{1}{f'} = \frac{1}{20}$$

or
$$\frac{1}{f}(1 - 0.56942) = \frac{1}{20}$$

or
$$f = 20 \times 0.43058 = 8.61 \text{ cm}$$

and
$$f' = -\frac{f}{0.56942} \simeq -15.1 \text{ cm}$$

6.2.2 Removal of Chromatic Aberration of a Separated Doublet

Let us consider two thin lenses of focal lengths f and f' and separated by a distance t (see Fig. 6.3). The focal length of the combination F is

$$\frac{1}{F} = \frac{1}{f} + \frac{1}{f'} - \frac{t}{ff'} \qquad \textbf{(11)}$$

The focal length of the first lens is given by

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad \textbf{(12)}$$

with a similar expression for $1/f'$. If Δf and Δn represent the changes in the focal length and in the refractive index due to a change $\Delta\lambda$ in the wavelength, then by differentiating Eq. (12) we obtain

$$-\frac{\Delta f}{f^2} = \Delta n \left(\frac{1}{R_1} - \frac{1}{R_2}\right) = \frac{\Delta n}{(n-1)f}$$

Thus, differentiating Eq. (11), we obtain

$$-\frac{\Delta F}{F^2} = -\frac{\mathrm{D}f}{f^2} - \frac{\mathrm{D}f'}{f'^2} + \frac{t}{f}\frac{\mathrm{D}f'}{f'^2} + \frac{t}{f'}\frac{\mathrm{D}f}{f^2}$$

$$= \frac{\Delta n}{(n-1)f} + \frac{\Delta n'}{(n'-1)f'} - \frac{t}{f}\frac{\Delta n'}{(n'-1)f'} - \frac{t}{f'}\frac{\Delta n}{(n-1)f}$$

$$= \frac{\omega}{f} + \frac{\omega'}{f'} - \frac{t}{ff'}(\omega + \omega') \qquad \textbf{(13)}$$

where, as before, ω and ω' represent the dispersive powers. Consequently, for the combination to have the same focal length for blue and red colors we should have

$$\frac{t(\omega + \omega')}{ff'} = \frac{\omega}{f} + \frac{\omega'}{f'}$$

or
$$t = \frac{\omega f' + \omega' f}{\omega + \omega'} \qquad \textbf{(14)}$$

If both the lenses are made of the same material, then $\omega = \omega'$ and the above equation simplifies to

$$t = \frac{f + f'}{2} \qquad \textbf{(15)}$$

implying that the chromatic aberration is very small if the distance between the two lenses is equal to the mean of the focal lengths. This is indeed the case for the Huygens eyepiece.

6.3 MONOCHROMATIC ABERRATIONS

6.3.1 Spherical Aberration

Let a beam of light parallel to the axis be incident on a thin lens (see Fig. 6.4). The light rays after passing the lens bend toward the axis and cross the axis at some point. If we restrict ourselves to the paraxial region, then we can see that all rays cross the z axis at the same point, which is at a distance f_p from the lens; f_p represents the paraxial focal length of the lens. If one does not restrict to the paraxial region, then in general, rays which are incident at different heights on the lens hit the axis at different points. For example, for a convex lens, the marginal rays (which are incident near the periphery of the lens) focus at a point closer than the focal point of paraxial rays [see Fig. 6.4(a)]. Similarly, for a concave lens, rays which are incident farther from the axis appear to be emerging from a point which is nearer to the lens [see Fig. 6.4(b)]. The point F_P at which the paraxial rays strike the axis is called the paraxial focus, and the point F_M at which the rays near the periphery strike is called the marginal focus. The distance between the two foci is a measure of spherical aberration in the lens. Thus if O represents an axial object, then different rays emerging from the object converge to different points; consequently, the image of a point object will not be a point. The distance along the axis between the paraxial image point and the image corresponding to marginal rays (i.e., rays striking the edge of the lens) is termed *longitudinal spherical aberration*. Similarly, the distance between the paraxial image point and the point at which the marginal ray strikes the paraxial image plane is called the *lateral spherical aberration* [see Fig. 6.4(a)]. The image on any plane (normal to the z axis) is a circular patch of light; however, as can be seen from Fig. 6.4(a), on a plane AB the circular patch has the least diameter. This is called the **circle of least confusion** (see Fig. 6.5). For an object lying on the axis of a cylindrically symmetric system (such as system of coaxial lenses), the image will suffer only from spherical aberration. All other off-axis aberrations such as coma and astigmatism will be absent.

To see how the rays hitting the refracting surface at different heights could focus to different points on the axis, let us consider the simple case of a plane refracting surface as shown

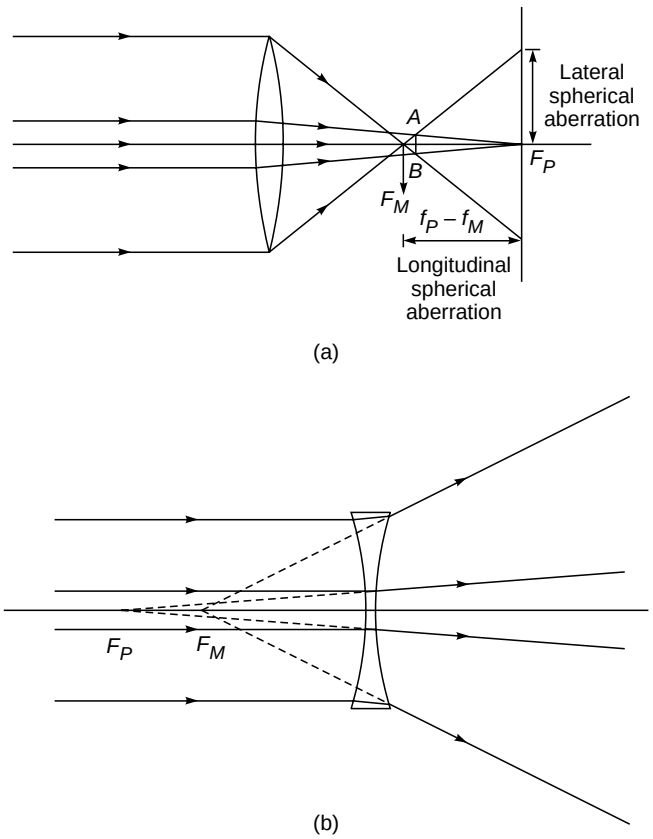


Fig. 6.4 (a) For a converging lens, the focal point for marginal rays lies closer to the lens than the focal point for paraxial rays. The distance between the paraxial focal point and the marginal focal point is known as the longitudinal spherical aberration, and the radius of the image at the paraxial focal plane is known as the lateral spherical aberration. The combined effect of defocusing and spherical aberration leads to the formation of a circle of least confusion, where the image has the minimum diameter. (b) The spherical aberration of a diverging lens.

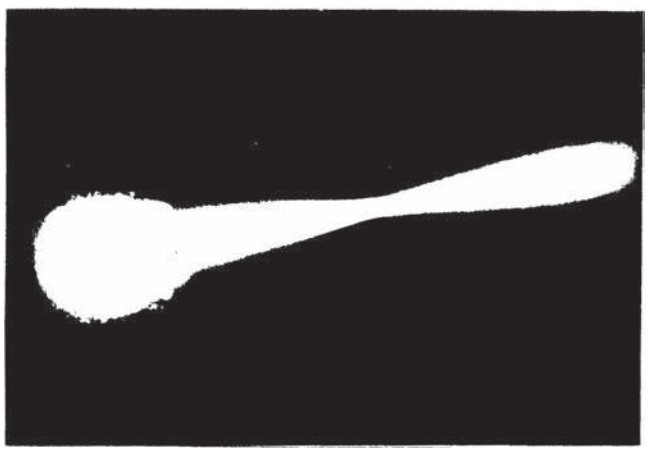


Fig. 6.5 The spherical aberration of a convex lens (photograph courtesy Dr. K. K. Gupta).

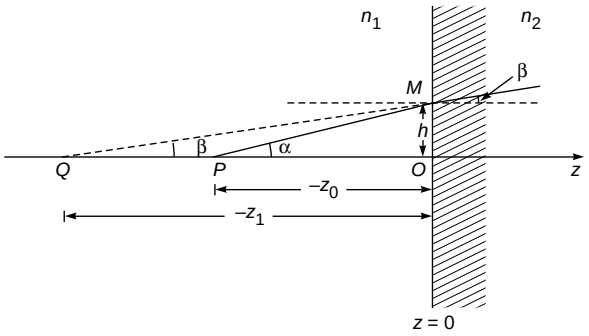


Fig. 6.6 Refraction at a plane surface.

in Fig. 6.6. Let the plane of the refracting surface be chosen as the plane $z = 0$. Let P be the object point. The z axis is chosen to be along the normal (PO) from point P to the surface. The plane $z = 0$ separates two media of refractive indices n_1 and n_2 (see Fig. 6.6); in the figure we have assumed $n_2 > n_1$. Consider a ray PM incident on the refracting surface (from the object) at a height h as shown in Fig. 6.6. The refracted ray appears to emerge from the point Q. We assume the origin to be at the point O. Let the z coordinates of the points P and Q be z_0 and z_1, respectively. Obviously, both z_0 and z_1 will be negative quantities, and the distances OP and OQ will be $-z_0$ and $-z_1$, respectively (see Fig. 6.6). We have to determine z_1 in terms of z_0. From Snell's law we know that

$$\sin\alpha = n \sin\beta \tag{16}$$

where α and β are the angles that the incident and refracted rays make with the z axis and

$$n = \frac{n_2}{n_1} \tag{17}$$

Now, from Fig. 6.6 we have

$$-z_1 = h \cot\beta = \frac{h}{\sin\beta}\sqrt{1-\sin^2\beta}$$

or

$$z_1 = -\frac{nh}{\sin\alpha}\left(1-\frac{1}{n^2}\sin^2\alpha\right)^{1/2} \tag{18}$$

where we have used Eq. (16). Since

$$\sin\alpha = \frac{h}{\sqrt{h^2+z_0^2}} \tag{19}$$

we obtain

$$z_1 = -\frac{nh}{h}(h^2+z_0^2)^{1/2}\left(1-\frac{1}{n^2}\frac{h^2}{h^2+z_0^2}\right)^{1/2} \tag{20}$$

or

$$z_1 = -n|z_0|\left(1+\frac{h^2}{z_0^2}\right)^{1/2}\left[1-\frac{h^2}{n^2 z_0^2}\left(1+\frac{h^2}{z_0^2}\right)^{-1}\right]^{1/2} \tag{21}$$

The value of z_1 given in Eq. (21) is an exact expression in terms of z_0. At once it can be seen that the image distance z_1 is a complicated function of the height h at which the ray strikes the refracting surface. In the limit of $h \to 0$, i.e., for paraxial rays, we get

$$z_1 = -n|z_0| \tag{22}$$

which is the expression for the image distance in the paraxial region. To the next order of approximation, assuming $|h/z_0| << 1$, we get

$$z \simeq -n|z_0|\left(1+\frac{h^2}{2z_0^2}\right)\left(1-\frac{h^2}{2n^2z_0^2}\right)$$

$$\simeq -n|z_0|\left[1+\frac{h^2}{2z_0^2 n^2}(n^2-1)\right] \tag{23}$$

Thus the aberration is given by

$$\Delta z = -\frac{h^2}{2n|z_0|}(n^2-1) \tag{24}$$

Equation (24) gives the longitudinal spherical aberration. The negative sign implies that the nonparaxial rays appear to emanate from a point which is farther away from the paraxial image point.

From the above example, it can be seen that even a single plane refracting surface suffers from spherical aberration. Thus, spherical refracting surfaces and thin lenses must also suffer from spherical aberration.

The calculation of the spherical aberration even for a single spherical refracting surface is quite cumbersome (see, e.g., Ref. 5); we just give the final results:

$$\Delta z = -\frac{n_2-n_1}{2n_2\left(\frac{1}{z_0}+\frac{n_2-n_1}{n_1R}\right)^2}\left(\frac{1}{R}+\frac{1}{z_0}\right)^2 \times\left(-\frac{n_2+n_1}{n_1z_0}+\frac{1}{R}\right)h^2 \tag{25}$$

where R represents the radius of curvature of the surface, and n_1 and n_2 represent the refractive indices of the media on the left and right, respectively, of the spherical surface (see Fig. 6.7). For a plane surface $R = \infty$, Eq. (25) reduces to Eq. (24) with $n = n_2/n_1$.

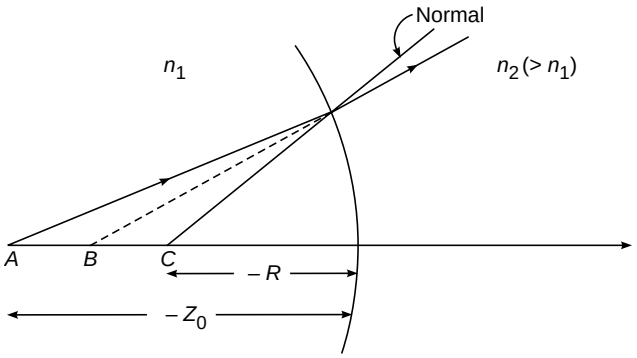


Fig. 6.7 The aplanatic points of a spherical refracting surface.

Example 6.2 Consider a spherical refracting surface of radius R. Show that for a point A [see Fig. 6.7(b)] such that

$$z_0 = \frac{n_1 + n_2}{n_1} R \qquad \textbf{(26)}$$

the spherical aberration is zero. Notice that both R and z_0 are negative quantities. The corresponding image point B is at a distance $[(n_2–n_1)/n_2]\ z_0$. Points A and B are known as the aplanatic points and are utilized in microscope objectives.

Solution: For $z_0 = [(n_1+n_2)/n_1]\ R$, one of the factors in Eq. (25) vanishes and the spherical aberration is zero. Indeed, it can be rigorously shown that all rays emanating from point A appear to diverge from point B (see also Sec. 4.8).

Example 6.3 Consider a refracting surface obtained by revolving an ellipse about its major axis (see Fig. 6.8). Show that all the rays parallel to the major axis will focus at one of the foci if the eccentricity of the ellipse is equal to n_1/n_2.

[*Hint:* The eccentricity of the ellipse is given by

$$\varepsilon = \frac{OF}{a} = \left(1 - \frac{b^2}{a^2}\right)^{1/2}$$

where a and b are the semimajor and semiminor axes, respectively. If we assume $n_1(QP) + n_2(PF) = n_2(BF)$, then one can easily show that the coordinates of the point $P(x, y)$ will satisfy the equation of the ellipse.]

In a similar manner, for a set of rays incident parallel to the axis, one can show that the coefficient of spherical aberration of a thin lens made of a material of refractive index n and placed in air, with the surfaces having radii of curvature R_1 and R_2, is given by

$$A = -\frac{f(n-1)}{2n^2} \times \left\{ -\left(\frac{1}{R_2} - P\right)^2 \left[\frac{1}{R_2} - P(n+1)\right] + \frac{1}{R_1^3} \right\} \qquad \textbf{(27)}$$

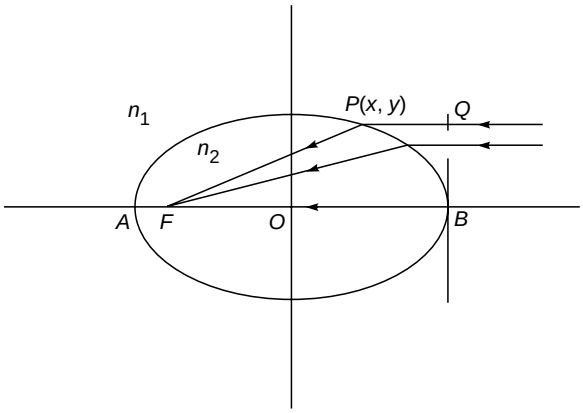


Fig. 6.8 For Example 6.3.

where

$$P = \frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad \textbf{(28)}$$

represents the power of the lens. The coefficient A is such that when it is multiplied by the cube of the height of the ray at the lens, one obtains the lateral spherical aberration. Thus the lateral spherical aberration for rays hitting the lens at a height h is

$$S_{\text{lat}} = Ah^3 = -\frac{f(n-1)h^3}{2n^2} \times \left\{ -\left(\frac{1}{R_2} - P\right)^2 \left[\frac{1}{R_2} - P(n+1)\right] + \frac{1}{R_1^3} \right\} \qquad \textbf{(29)}$$

The longitudinal spherical aberration which corresponds to the difference between the marginal focal length and the paraxial focal length is given by

$$S_{\text{long}} = Ah^2 f = -\frac{(n-1) f^2 h^2}{2n^2} \times \left[\frac{1}{R_1^3} - \left(\frac{1}{R_2} - \frac{n+1}{f}\right)\left(\frac{1}{R_2} - \frac{1}{f}\right)^2 \right] \qquad \textbf{(30)}$$

For a converging lens, S_{long} will always be negative, implying that the marginal rays focus closer to the lens.

For a thin lens of given power (i.e., of a given focal length), one can define a quantity q, called the shape factor, by the relation

$$q = \frac{R_2 + R_1}{R_2 - R_1} \qquad \textbf{(31)}$$

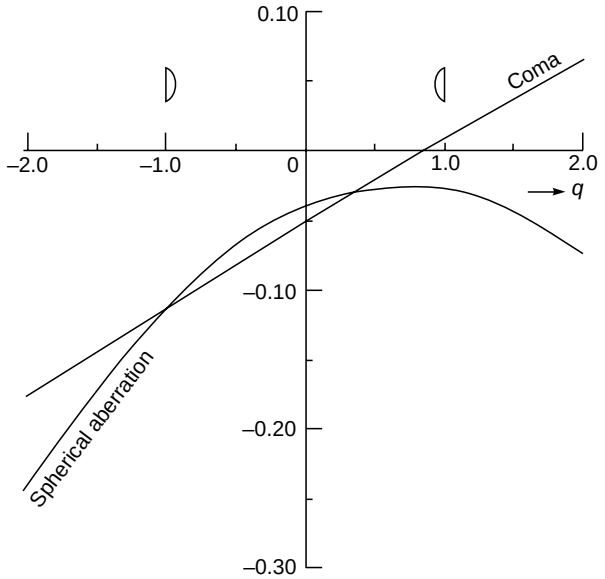


Fig. 6.9 Variation of spherical aberration and coma with the shape factor of a thin lens with $n = 1.5$, $f = 40$ cm, and $h = 1$ cm. For calculating the coma we have assumed $\tan\theta = 1$; i.e., rays make an angle of 45° with the axis.

where R_1 and R_2 are the radii of curvature of the two surfaces. For a given focal length of the lens, one can control the spherical aberration by changing the value of q. This procedure is called bending of the lens. Figure 6.9 shows the variation of spherical aberration with q for $n = 1.5$, $f = 40$ cm (i.e.; $P = 0.025$ cm^{-1}), and $h = 1$ cm. For values of q lying near $q \simeq +0.7$, the (magnitude of the) spherical aberration is minimum (but not zero). Thus, by choosing proper values of the radii, the spherical aberration can be minimized. The value $q = +1$ implies $R_2 = \infty$ and hence it corresponds to a planoconvex lens with the convex side facing the incident light. On the other hand, for a planoconvex lens with the plane side facing the incident light, $R_1 = \infty$ and $q = -1$. Thus the spherical aberration is dependent on how the deviation is divided between the surfaces.

The physical reason for the minimum of $|S_{\text{long}}|$ to occur at $q \simeq 0.7$ is as follows: It has already been mentioned before that (for a converging lens) the marginal rays undergo a large deviation which results in the spherical aberration [see Fig. 6.4(a)]. As such we should expect the spherical aberration to be minimum when the angle of deviation δ [see Fig. 6.10(a)] is minimum. As in the case of the prism [see Fig. 6.10(b)], this would occur when the deviations suffered at each of the refracting surfaces were exactly equal, i.e.,

$$\delta_1 = \delta_2 \qquad (\delta = \delta_1 + \delta_2) \qquad \textbf{(32)}$$

Indeed for $q = 0.7$, the deviations suffered at each of the surfaces are equal, and one obtains minimum spherical aberration.

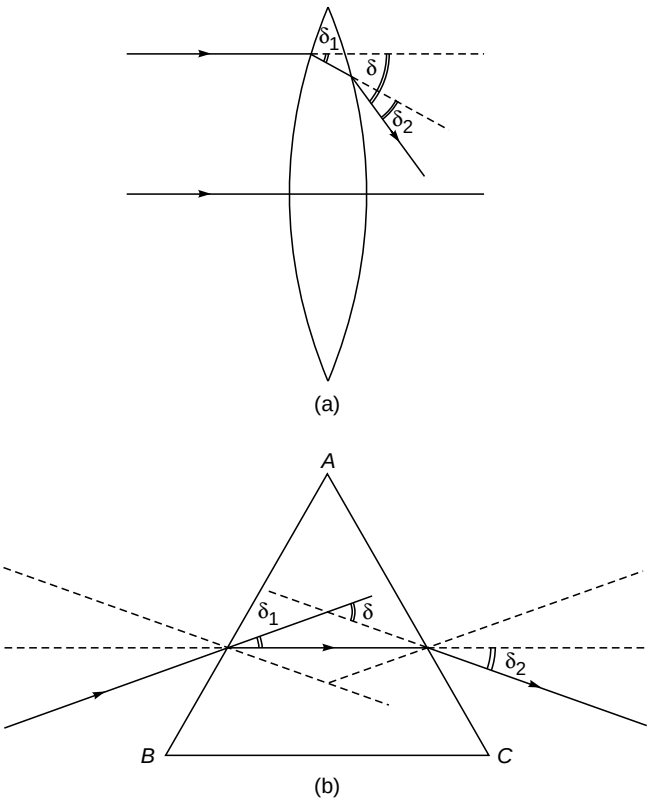


Fig. 6.10 (a) Refraction at the two refracting surfaces of a thin lens; the diagram is exaggerated to show clearly the angles. (b) For a prism, the minimum deviation position corresponds to $\delta_1 = \delta_2$.

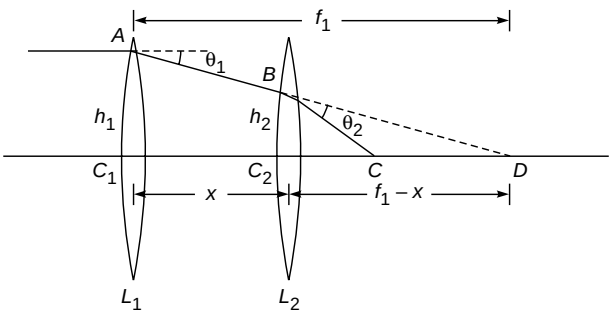


Fig. 6.11 Condition for minimum spherical aberration for a combination of two thin lenses.

Using the criterion of equal deviation discussed above, we will determine the separation between two thin lenses which would lead to minimum spherical aberration. Let L_1 and L_2 be two lenses of focal lengths f_1 and f_2, respectively, separated by a distance x (see Fig. 6.11). If θ_1 and θ_2 represent the deviations of the ray at the two lenses, then for minimum spherical aberration we get

$$\theta_1 = \theta_2 \qquad \textbf{(33)}$$

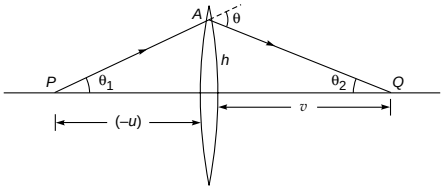


Fig. 6.12 Calculation of the angle of deviation.

To obtain an expression for the deviation suffered by a ray when it encounters a lens, we refer to Fig. 6.12 where a ray *PA* gets refracted along *AQ* after suffering a deviation through an angle θ. From triangle *PAQ*, we can see that

$$\theta = \theta_1 + \theta_2 \simeq \frac{h}{v} + \frac{h}{-u}$$

$$= h\left(\frac{1}{v} - \frac{1}{u}\right) = \frac{h}{f} \qquad \textbf{(34)}$$

where we have used the paraxial relation

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f} \qquad \textbf{(35)}$$

The quantity u is an intrinsically negative quantity. Thus Eq. (33) becomes

$$\frac{h_1}{f_1} = \frac{h_2}{f_2} \qquad \textbf{(36)}$$

From similar triangles AC_1D and BC_2D (see Fig. 6.11), we can write

$$\frac{h_1}{f_1} = \frac{h_2}{f_1 - x} \qquad \textbf{(37)}$$

If we use Eqs. (36) and (37), we obtain

$$x = f_1 - f_2 \qquad \textbf{(38)}$$

Thus the spherical aberration of a combination of two thin lenses is a minimum when their separation is equal to the difference in their focal lengths. Indeed, in the Huygens eyepiece (see Fig. 6.13), the focal length of the field lens is 3f, where f represents the focal length of the eye lens. The distance between the two lenses is 2f. We can immediately see that the conditions for achromatism [see Eq. (15)] and minimum spherical aberration [see Eq. (38)] are simultaneously satisfied. Since the eyepiece as a whole is corrected and the individual lenses are not, the image of the cross wires (which are placed in plane *PQ*) will show aberrations. A discussion of the procedure for reducing the aberrations in various optical instruments requires a very detailed analysis involving the tracing of the rays, which is beyond the scope of this book.

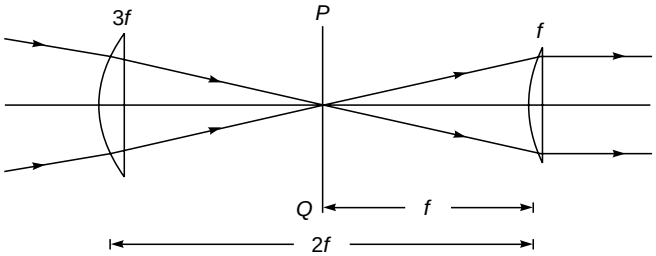


Fig. 6.13 The Huygens eyepiece.

Note that even when the system is free from all aberrations, the image of a point object will still not be a point because of diffraction effects (see Sec. 18.3). For example, if a perfectly spherical wave is emanating from a lens, the ray theory predicts a point image whereas the diffraction theory (which takes into account the finiteness of the wavelength) predicts that the image formed in the image plane will be an Airy pattern [see Fig. 18.8(c)], and the first dark ring will occur at a distance of $1.22\lambda f/D$ from the paraxial image point (see Fig. 6.14) where D is the diameter of the exit pupil. The Airy pattern shown in Fig. 6.14 is highly magnified. For example, for $\lambda = 5000$ Å, $D = 5$ cm, and $f = 10$ cm, the radii of the first and second dark rings in the Airy

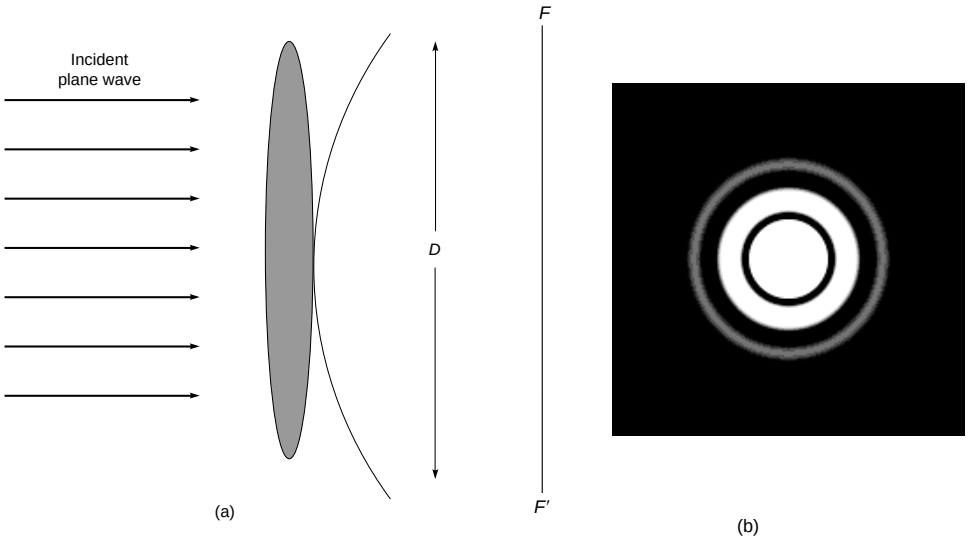


Fig. 6.14 A perfectly spherical wave (converging on the plane FF') will produce an Airy pattern on the image plane.

pattern will be about 0.00012 and 0.00022 mm, respectively (see Sec. 18.3). The spatial extent of the Airy pattern will become larger with a decrease in the value of D. Often one uses a "stop" to restrict to the paraxial region; however, if the diameter of the stop is made very small, then the diffraction effects will dominate. Indeed, a camera gives the best image when $f/D \approx 5.6$; at high apertures aberrations degrade the image, and at low apertures diffraction degrades the image.

6.3.2 Coma

As mentioned earlier, for a point object lying on the axis the image will suffer only from spherical aberration. For off-axis points, the image will also suffer from coma, astigmatism, curvature of field, and distortion. The first off-axis aberration is coma; i.e., for points lying very close to the axis, the image will suffer from spherical aberration and coma only. In this section we will briefly discuss the effect of coma, assuming that all other aberrations are absent.

The effect of coma is schematically shown in Fig. 6.15(a). The rays which proceed near the axis of the lens focus at a point different from that of the marginal rays. Thus, it appears that the magnification is different for different parts of the lens. If we consider the image formation by different zones of a lens, then the spherical aberration arises because different zones have different powers and coma arises because different zones have different magnifications. In Fig. 6.15(a) we have shown only those rays which lie in the meridional plane, i.e., that plane containing the optical axis and the object point. To see the shape of the image, one has to consider the complete set of rays.[1] In Fig. 6.15(b) we have shown a three-dimensional perspective in which we have considered a set of rays that hit the lens at the same distance from the center. Rays which intersect the lens at diametrically opposite points focus to a single point on the paraxial image plane. These different pairs of rays focus to different points in the image plane such that these foci lie on a circle. The radius of the circle and the distance at which the center lies from the ideal image point measure the coma. As the radius of the zone [shown as h in Fig. 6.15(b)] increases, the center of the circle also shifts away from the ideal image. Thus the composite image will have a form shown in Fig. 6.15(c). The image of a point object thus has a cometlike appearance and hence the name *coma* (see Fig. 6.16).

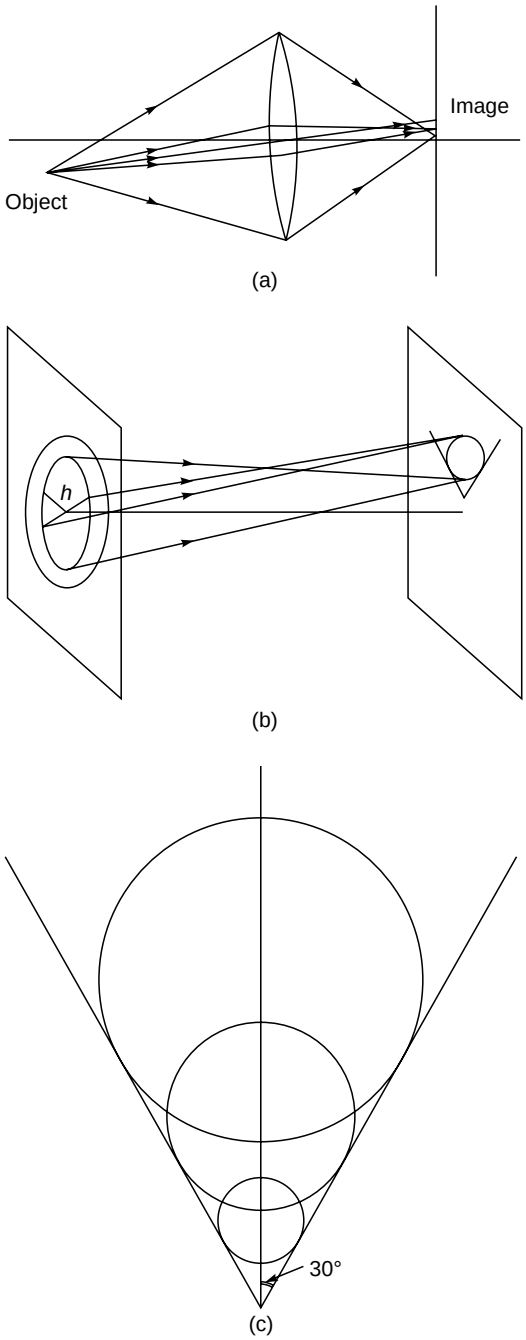


Fig. 6.15 The image formation in the presence of coma. In (a) we have shown only those rays which lie in the meridional plane. (b) A three-dimensional perspective is shown. (c) The composite image.

For a parallel bundle of rays incident on a lens and inclined at an angle θ with the z axis (see Fig. 6.17), one can show that the coma in the image is given by (see, e.g., Ref. 1)

$$\text{Coma} = \frac{3(n-1)}{2} f h^2 \tan^2\theta \times \left[\frac{(n-1)(2n+1)}{nR_1R_2} - \frac{n^2-n-1}{n^2R_1^2} - \frac{n}{R_2^2} \right] \quad \textbf{(39)}$$

In Fig. 6.9 we plotted the variation of coma with the shape factor q. It can immediately be seen that for a lens with $q = +0.8$, coma is zero. It can also be seen that both spherical aberration and coma are close to a minimum for a

[1] It must be mentioned that a proper understanding of the aberrations can only be achieved by a careful and thorough mathematical analysis. This, however, is beyond the scope of this book; interested readers may look up Refs. 1 and 3.

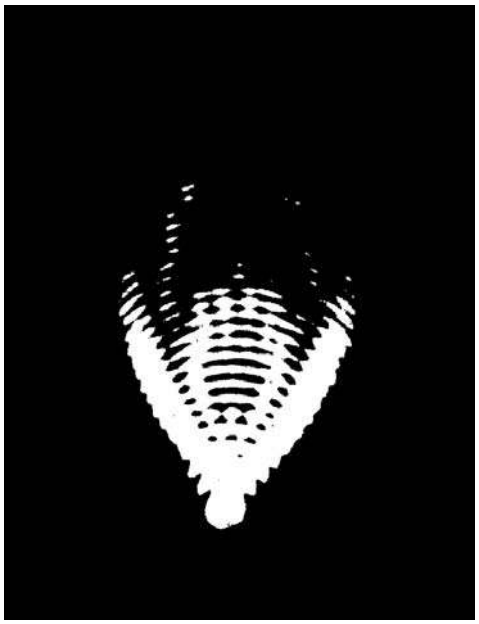


Fig. 6.16 Image of a point source showing coma. [After H. F. Meiners, *Physics Demonstration Experiments,* Vol. ll, The Ronald Press Co., New York, 1970; used with permission.]

planoconvex lens (with the convex side facing the incident light) for which $q = 1.0$, and as such planoconvex lenses are extensively used in eyepieces.

Note that in Sec. 4.11 we derived the Abbe sine condition; when it is satisfied, the optical system is free from spherical aberration and coma.

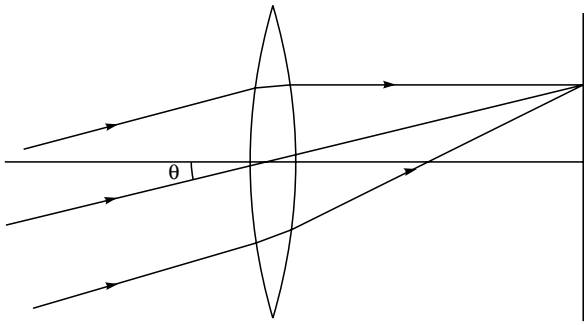


Fig. 6.17 Parallel rays (inclined at an angle θ with the axis) incident on a thin lens.

6.3.3 Astigmatism and Curvature of Field

When an optical system is free from spherical aberration and coma, then the system will image sharply those object points lying on or near the axis. But for points far away from the axis, the image of a point will not be a point and then the optical system is said to be afflicted with astigmatism.

Consider an object point *P* far away from the axis. The plane containing the axis and the object point is called the meridional plane, and the plane perpendicular to the meridional plane (containing the axis) is called the *sagittal plane*. Figure 6.18 shows the image formation when the optical system suffers from astigmatism only. The rays in the meridional plane converge at a different point compared to those in the sagittal plane. For example, rays *PA* and *PB* focus at point *T*, and rays *PC* and *PD* focus at a point *S* which is different

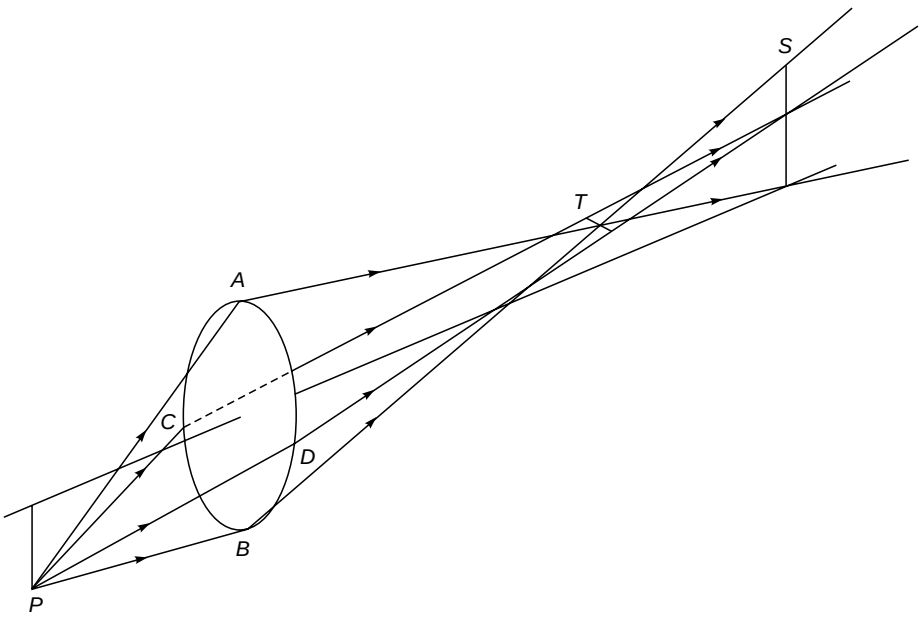


Fig. 6.18 Image formation in the presence of astigmatism.

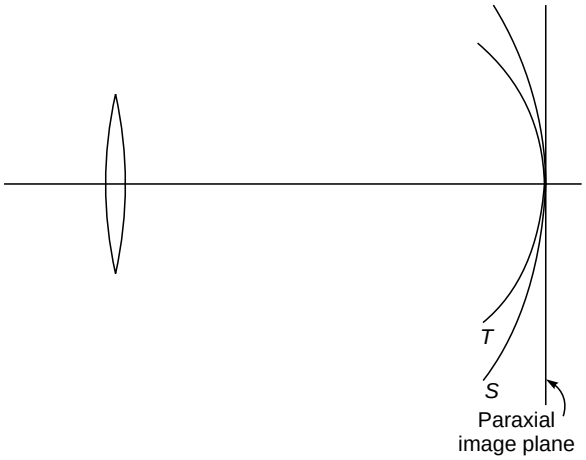


Fig. 6.19 Tangential and sagittal foci.

from *T*. Since at point *T* the rays in the sagittal plane have not still focused, one in fact has a focal line that is normal to the meridional plane. This focal line *T* is called the *tangential focus*. Similarly, since at *S* the rays in the meridional plane have defocused, one obtains a focal line lying in the tangential plane; this is called the sagittal focal line. The distance between *S* and *T* is a measure of astigmatism.

To see the origin of astigmatism, one observes that for a point on the axis (when the lens is free from other aberrations) the wave front emerging from the lens is spherical; and thus as the wave front progresses, it converges to a single point. But when the object point is nonaxial, then the emerging wave front is not spherical; and thus as the wave front converges, it focuses not to a point but to two lines, which are normal to each other and called the tangential and the sagittal focal lines. Somewhere between the two focal lines, the image is circular and is called the circle of least confusion.

The distance between the tangential and sagittal foci increases as the object point moves away from the axis. Thus the tangential foci and the sagittal foci of points at different distances from the axis lie on two surfaces, as shown in Fig. 6.19. The optical system is said to be free from astigmatism when the two surfaces coincide. But even when they coincide, it can be shown that the resultant image surface will be curved. This defect of the image is termed *curvature of the field.*

As an example of image formation in the presence of astigmatism, consider a spoked wheel coaxial with the lens axis, as shown in Fig. 6.20(a). Since on the *T* surface the image of a point source is a line perpendicular to the meridional plane, on the *T* surface, the complete rim of the wheel will be in focus while the spokes will be out of focus, as shown in Fig. 6.20(b). Similarly, since on the *S* surface the image of a point is a line in the meridional plane, the spokes will be in focus, and the rim will not be in focus, as shown in Fig. 6.20(c).

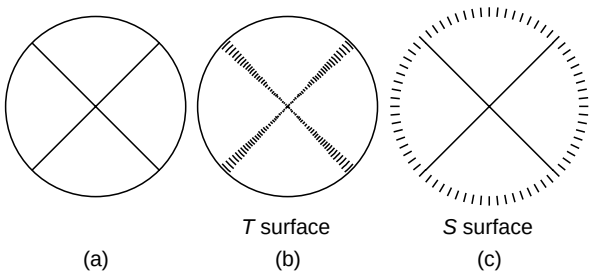


Fig. 6.20 (a) Spoked object coaxial with the axis of the lens; (b), (c) Images on the *T* surface and *S* surface, respectively.

6.3.4 Distortion

The last of the Seidel aberrations is called distortion and is caused by nonuniform magnification of the system. When we discussed spherical aberration, we mentioned that for a point object on the axis of the optical system, the images will suffer only from spherical aberration. Similarly, if we have a pinhole on the axis at any plane of the optical system (see Fig. 6.21), then the image will suffer only from distortion. This is so because corresponding to any point in the object plane, only one of the rays emanating from this point will pass through the pinhole; consequently, all other aberrations will be absent. Obviously, for such a configuration, each point will be imaged as a point; but if the system suffers from nonuniform magnification, the image will be distorted. This can be illustrated if we consider the imaging of four equally spaced points *A*, *B*, *C*, and *D*, which are imaged as *A*′, *B*′, *C*′, and *D*′, respectively. Mathematical analysis shows that (Ref. 3)

$$X_d = Mx_0 + E(x_0^2 + y_0^2)x_0 \qquad \textbf{(40)}$$

and

$$Y_d = My_0 + E(x_0^2 + y_0^2)y_0 \qquad \textbf{(41)}$$

where (x_0, y_0) and (X_d, Y_d) represent the coordinates of the object and the image point, respectively; *M* represents the magnification of the system; and *E* represents the coefficient

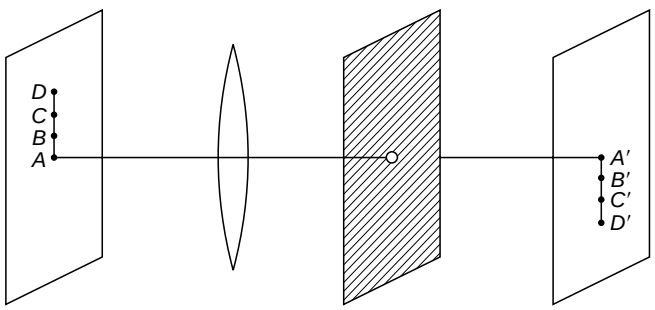


Fig. 6.21 In the presence of a pinhole on the axis, the image suffers only from distortion.

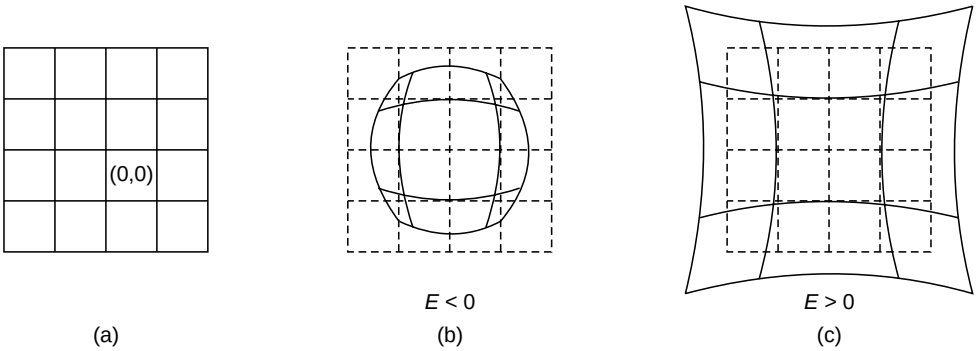


Fig. 6.22 (a) shows the object, (b) represents the image when $E < 0$ and (c) when $E > 0$.

of distortion. Figure 6.22(b) corresponds to a negative value of E and is known as barrel distortion. The distortion of the image can be easily understood if we consider the imaging of a square grid, as shown in Fig. 6.22. Assuming unit magnification (i.e., $M = 1$), the points having coordinates (0, 0), (h, 0), (2h, 0), (3h, 0), (0, h), (0, 2h), (0, 3h), (h, h), (h, 2h), (2h, h), ... are imaged at (0, 0), ($h + Eh^3$, 0), ($2h + 8Eh^3$, 0), ($3h + 27Eh^3$, 0), (0, $h + Eh^3$), (0, $2h + 8Eh^3$), (0, $3h + 27Eh^3$), ($h + Eh^3$, $h + Eh^3$), ($h + Eh^3$, $2h + 8Eh^3$), ($2h + 8Eh^3$, $h + Eh^3$), ..., respectively. If you actually plot these points, then for $E < 0$, you obtain a figure like the one shown in Fig. 6.22(b). Similarly for $E > 0$, you obtain Fig. 6.22(c). Notice that each point is imaged at a point, but the image is distorted because of nonuniform magnification.

Summary

- For a polychromatic source, different wavelength components (after refraction) proceed along different directions and form images at different points; this leads to chromatic aberrations. If we consider two thin lenses made of different materials placed in contact with each other, the focal length of the combination will be the same for blue and red colors if

$$\frac{\omega}{f} + \frac{\omega'}{f'} = 0$$

where

$$\omega = \frac{n_b - n_r}{n - 1} \quad \text{and} \quad \omega' = \frac{n_b' - n_r'}{n' - 1}$$

are known as the dispersive powers. Further,

$$n \equiv \frac{n_b + n_r}{2} \approx n_y \qquad n' \equiv \frac{n_b' + n_r'}{2} \approx n_y'$$

where n_b, n_y, and n_r represent the refractive indices for the material of the first lens corresponding to the blue, yellow, and red colors respectively. Similarly, n_b', n_y', and n_r' represent the refractive indices for the second lens. Since ω and ω' are both positive, f and f' must have opposite signs.

- For a lens, the marginal rays (which are incident near the periphery of the lens) focus at a point which is different from the focal point of paraxial rays. The distance along the axis between the paraxial image point and the image corresponding to marginal rays (i.e., rays striking the edge of the lens) is termed *longitudinal spherical aberration.*
- The spherical aberration of a combination of two thin lenses is a minimum when their separation is equal to the difference in their focal lengths.

Problems

6.1 Consider a plane glass slab of thickness d made of a material of refractive index n, placed in air. By simple application of Snell's law, obtain an expression for the spherical aberration of the slab. What are other kinds of aberrations that the image will suffer from?

[**Ans:** Spherical aberration $= -\dfrac{(n^2-1)dh^2}{2n^3u^2}$, where h is the height at which the ray strikes the slab and u is the distance of the object point from the front surface of the slab.]

6.2 Why can't you obtain an expression for the spherical aberration of a plane glass slab from Eq. (27) by tending R_1, and R_2 to ∞?

6.3 Obtain an expression for the chromatic aberration in the image formed by a plane glass slab.

$$\left[\textbf{Ans}: \approx d\left(\frac{1}{n_r} - \frac{1}{n_b}\right)\right]$$

6.4 Does the image formed by a plane mirror suffer from any aberration?

6.5 Calculate the longitudinal spherical aberration of a thin planoconvex lens made of a material of refractive index 1.5 and whose curved surface has a radius of curvature of 10 cm, for rays incident at a height of 1 cm. Compare the values of

the aberration when (a) the convex side and (b) the plane side face the incident light.

[**Ans:** (a) ≈ -0.058 cm; (b) ≈ -0.225 cm]

6.6 Consider a lens made up of a material of refractive index 1.5 with a focal length 25 cm. Assuming $h = 0.5$ cm and $\theta = 45°$, obtain the spherical aberration and coma for the lens for various values of the shape factor q, and plot the variation in a manner similar to that shown in Fig. 6.9.

6.7 An achromatic cemented doublet of focal length 25 cm is to be made from a combination of an equiconvex flint glass lens ($n_b = 1.50529$, $n_r = 1.49776$) and a crown glass lens ($n_b = 1.66270$, $n_r = 1.64357$). Calculate the radii of curvature of the different surfaces and the focal lengths of each of the two lenses.

[**Ans:** $R_1 = 14.2$ cm $= -R_2 = -R_1'$; $R_2' = -42$ cm]

REFERENCES AND SUGGESTED READINGS

1. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, 1975.
2. M. Cagnet, M. Francon, and J. C. Thierr, *Atlas of Optical Phenomena*, Springer-Verlag, Berlin, 1962.
3. A. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978. [Reprinted by Macmillan India, New Delhi.]
4. H. H. Hopkins, *Wave Theory of Aberrations*, Oxford University Press, London, 1950.
5. C. J. Smith, "A Degree Physics," Part III, *Optics*, Edward Arnold Publishers, London, 1960.
6. W. T. Welford, *Geometrical Optics*, North Holland Publishing Co., Amsterdam, 1962.
7. W. T. Welford, *Aberrations of the Symmetrical Optical System*, Academic Press, New York, 1974.

PART 2
Vibrations and Waves

This part (consisting of six chapters) discusses many interesting experiments such as the physics behind ionospheric reflection, redness of the setting Sun, water waves, and pulse dispersion. Chapter 7 starts with simple harmonic motion (which is the most fundamental vibration associated with wave motion) and is followed by a derivation of the refractive index variation with frequency. Chapters 8 and 9 discuss Fourier series and Fourier transforms which are extensively used in studying the distortion of optical pulses as they propagate through dispersive media (Chap. 10). The derivation and solutions of the wave equation represent the basic physics of wave propagation which are discussed in Chap. 11. Chapter 12 discusses Huygens' principle which is used to derive the laws of reflection and Snell's law of refraction.

Chapter Seven

SIMPLE HARMONIC MOTION, FORCED VIBRATIONS, AND ORIGIN OF REFRACTIVE INDEX

The correct picture of an atom, which is given by the theory of wave mechanics, says that, *so far as problems involving light are concerned*, the electrons behave as though they were held by springs. So we shall suppose that the electrons have a linear restoring force which, together with their mass *m*, makes them behave like little oscillators, with a resonant frequency ω_0 .... The electric field of the light wave polarizes the molecules of the gas, producing oscillating dipole moments. The acceleration of the oscillating charges radiates new waves of the field. This new field, interfering with the old field, produces a changed field which is equivalent to a phase shift of the original wave. Because this phase shift is proportional to the thickness of the material, the effect is equivalent to having a different phase velocity in the material.

—Richard Feynman, *Feynman Lectures on Physics*, Vol. I

7.1 INTRODUCTION

The most fundamental vibration associated with wave motion is the simple harmonic motion; in Sec. 7.2 we will discuss simple harmonic motion, and in Sec. 7.3 we will discuss the effects (on the vibratory motion) due to damping. If a periodic force acts on a vibrating system, the system undergoes what are known as forced vibrations; in Sec. 7.4 we will study such vibrations which will allow us to understand the origin of refractive index (see Sec. 7.5) and even Rayleigh scattering (see Sec. 7.6), which is responsible for the red color of the setting (or rising) Sun and blue color of the sky (see Fig. 3.21 and the color photograph in the insert at the end of the book).

7.2 SIMPLE HARMONIC MOTION

A periodic motion is a motion which repeats itself after regular intervals of time, and the simplest kind of periodic motion is a simple harmonic motion in which the displacement varies sinusoidally with time. To understand simple harmonic motion, we consider a point P rotating on the circumference of a circle of radius a with an angular velocity ω (see Fig. 7.1). We choose the center of the circle as our origin, and we assume that at $t = 0$ the point P lies on the x axis (i.e., at point P_0). At an arbitrary time t the point will be at position P where $\angle POP_0 = \omega t$.

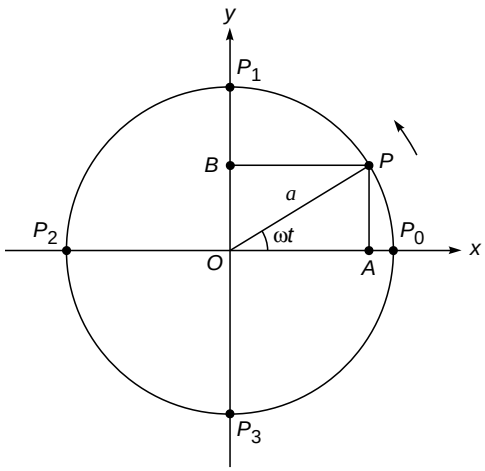


Fig. 7.1 The point P is rotating in the counterclockwise direction on the circumference of a circle of radius a, with uniform angular velocity ω. The foot of the perpendicular on any one of the diameters executes simple harmonic motion. Point P_0 is the position of the point at $t = 0$.

Let A be the foot of the perpendicular from the point P on the x axis. Clearly, the distance

$$OA = a \cos \omega t \quad \textbf{(1)}$$

and as point P rotates on the circumference of the circle, point A moves to and fro about the origin on the diameter. When point P is at P_1, then the foot of the perpendicular is at O. This can also be seen from Eq. (1) because when P coincides with P_1, $\omega t = \pi/2$ and hence $a \cos \omega t = a \cos \pi/2 = 0$. As the point still moves farther, the foot of the perpendicular will lie on the other side of the origin and thus OA will be negative, as is also evident from Eq. (1) because ωt then greater than $\pi/2$. When P coincides with P_2, then $OA = OP_2 = -a$. When point P moves from P_2 to P_3, OA starts decreasing and it finally goes to zero when P coincides with P_2. After P crosses P_3, OA starts increasing again and finally acquires the value a when P coincides with P_0. After crossing the point P_0, the motion repeats itself.

A motion in which the displacement varies sinusoidally with time [as in Eq. (1)] is known as a *simple harmonic motion. Thus, when a point rotates on the circumference of a circle with a uniform angular velocity, the foot of the perpendicular on any one of its diameters will execute simple harmonic motion.* The quantity a is called the amplitude of the motion, and the period of the motion T will be the time required to complete one revolution. Since the angular velocity is ω, the time taken for one complete revolution will be $2\pi/\omega$. Thus,

$$T = \frac{2\pi}{\omega} \quad \textbf{(2)}$$

The inverse of the time period is known as the frequency:

$$\nu = \frac{1}{T} = \frac{\omega}{2\pi}$$

or
$$\omega = 2\pi\nu \quad \textbf{(3)}$$

We could as well have studied the motion of the point B, which is the foot of the perpendicular from point P on the y axis. The distance OB is given by (see Fig. 7.1)

$$OB = y = a \sin \omega t \quad \textbf{(4)}$$

We had conveniently chosen $t = 0$ as the time when P was on the x axis. The choice of the time $t = 0$ is arbitrary, and we could have chosen time $t = 0$ to be the instant when P was at P' (see Fig. 7.2). If the angle $\angle P'OX = \theta$, then the projection on the x axis at any time t is given by

$$OA = x = a \cos (\omega t + \theta) \quad \textbf{(5)}$$

The quantity $\omega t + \theta$ is known as the *phase* of the motion, and θ represents the initial phase. It is obvious from the above discussion that the value of θ is quite arbitrary and depends on the instant from which we start measuring time.

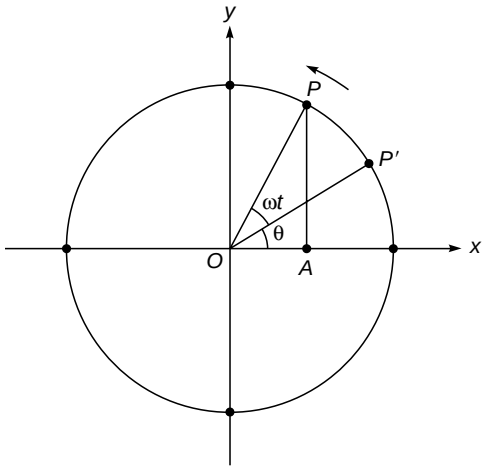


Fig. 7.2 At $t = 0$, point P is at P', and therefore the initial phase is θ.

We next consider two points P and Q rotating on the circle with the same angular velocity, and we let P' and Q' be their respective positions at $t = 0$. Let the angles $\angle P'OX$ and $\angle Q'OX$ be θ and ϕ, respectively (see Fig. 7.3). Clearly at an arbitrary time t, the distance of the foot of the perpendiculars from the origin would be

$$x_P = a \cos (\omega t + \theta) \quad \textbf{(6a)}$$
$$x_Q = a \cos (\omega t + \phi) \quad \textbf{(6b)}$$

The quantity

$$(\omega t + \theta) - (\omega t + \phi) = \theta - \phi \quad \textbf{(7)}$$

represents the phase difference between the two simple harmonic motions; and if $\theta - \phi = 0$ (or an even multiple of π), the

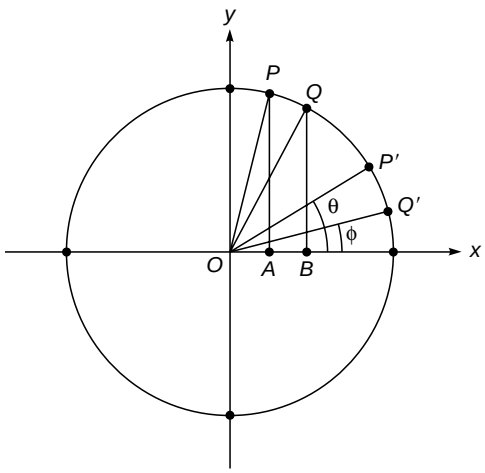


Fig. 7.3 Points A and B execute simple harmonic motions with the same frequency ω. The initial phases of A and B are θ and ϕ, respectively.

motions are said to be in phase, and if $\theta - \phi = \pi$ (or an odd multiple of π), the motions are said to be out of phase. If we choose a different origin of time, the quantities θ and ϕ will change by the same additive constant; consequently, the phase difference $\theta - \phi$ is independent of the choice of the instant $t = 0$.

Thus the displacement of a particle, which executes simple harmonic motion, can be written as

$$x = a \sin(\omega t + \theta) \quad \textbf{(8)}$$

Therefore, the velocity and the acceleration of the particle are given by the following equations:

$$v = \frac{dx}{dt} = a\omega \cos(\omega t + \theta) \quad \textbf{(9)}$$

and

$$f = \frac{d^2x}{dt^2} = -a\omega^2 \sin(\omega t + \theta)$$

or,

$$f = \frac{d^2x}{dt^2} = -\omega^2 x \quad \textbf{(10)}$$

Equation (10) shows that the acceleration of the particle is proportional to the displacement, and the negative sign indicates that the acceleration is always directed toward the origin. Equation (10) can be used to define the simple harmonic motion as the motion of a particle in a straight line in which the acceleration is proportional to the displacement from a fixed point (on the straight line) and always directed toward the fixed point. (Here the point $x = 0$ is the fixed point and is usually referred to as the equilibrium position.) If we multiply Eq. (10) by the mass of the particle, then we obtain the following expression for the force acting on the particle:

$$F = mf = -m\omega^2 x$$

or

$$F = -kx \quad \textbf{(11)}$$

where k ($= m\omega^2$) is known as the *force constant*. We could have equally well started from Eq. (11) and obtained simple harmonic motion. This can be easily seen by noting that since the force is acting in the x direction, the equation of motion will be

$$m\frac{d^2x}{dt^2} = F = -kx$$

or

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0$$

or

$$\frac{d^2x}{dt^2} + \omega^2 x = 0 \quad \textbf{(12)}$$

where $\omega^2 = k/m$. The general solution of Eq. (12) can be written in the form

$$x = A \sin \omega t + B \cos \omega t \quad \textbf{(13)}$$

which can be rewritten in either of the following forms:

$$x = a \sin(\omega t + \theta) \quad \textbf{(14)}$$

or

$$x = a \cos(\omega t + \theta) \quad \textbf{(15)}$$

which describes a simple harmonic motion.

7.2.1 Examples of Simple Harmonic Motion

In this section we discuss three simple examples of simple harmonic motion.

(a) The simple pendulum The simplest example of simple harmonic motion is the motion of the bob of a simple pendulum in the gravitational field. If the bob of the pendulum is displaced slightly from the equilibrium position (see Fig. 7.4), then the forces acting on the bob are the gravitational force *mg* acting vertically downward and the tension *T*, in the direction *B′A*. In the equilibrium position (*AB*) the tension is equal and opposite to the gravitational force. However, in the displaced position the tension *T* is not in the direction of the gravitational force; and if we resolve the gravitational force along the direction of the string and perpendicular to it, we see that the component *mg* cos θ balances the tension in the string and the component *mg* sin θ is the restoring

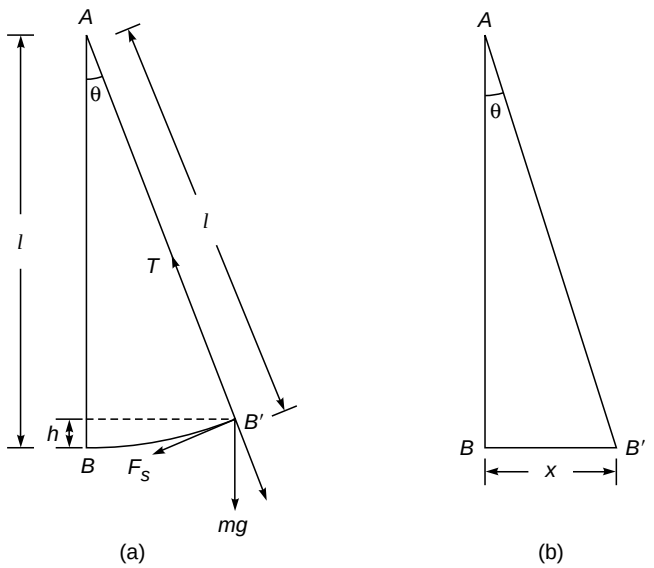


Fig. 7.4 (a) The forces on the bob of the pendulum when it is displaced from its equilibrium position. The restoring force is F_s which is equal to *mg* sin θ. (b) If the angle θ is small, the motion of the bob can be approximately assumed to be in a straight line.

force. The motion of the bob is along the arc of a circle but if the length of the pendulum is large and the angle θ is small, the motion can be assumed to be approximately in a straight line [see Fig. 7.4(b)]. Under such an approximation we may assume that this force is always directed toward point *B* and the magnitude of this force will be[1]

$$mg \sin \theta \approx mg \frac{x}{l} \tag{16}$$

Thus the equation of motion will be

$$F = m \frac{d^2x}{dt^2} = -mg \frac{x}{l} \tag{17}$$

or

$$\frac{d^2x}{dt^2} + \omega^2 x = 0 \tag{18}$$

where $\omega^2 = g/l$. Equation (18) is of the same form as Eq. (12); thus the motion of the bob is simple harmonic with its time period given by

$$T = \frac{2\pi}{\omega} = 2\pi \sqrt{\frac{l}{g}} \tag{19}$$

The expression for the time period is fairly accurate (i.e., the motion is approximately simple harmonic) as long as $\theta \lesssim 4°$.

We next consider the motion of two identical simple pendulums vibrating with the same amplitude *a* (see Fig. 7.5). Let, at $t = 0$, the bob of one of the pendulums be at its extreme right position, moving toward the right [Fig. 7.5(b)]. If we measure the displacement from the equilibrium positions of the pendulums, then the displacements are given by

$$x_1 = a \cos \omega t$$

$$x_2 = a \sin \omega t = a \cos \left(\omega t - \frac{\pi}{2} \right) \tag{20}$$

Thus the two bobs execute simple harmonic motion with a phase difference of π/2, and in fact the first pendulum is ahead in phase by π/2. In Fig. 7.5(b), if the bob were moving toward the left, then the equation of motion would have been

$$x_2 = -a \sin \omega t = a \cos \left(\omega t + \frac{\pi}{2} \right)$$

and then the second pendulum would have been ahead of phase by π/2. Since, in general, the displacement of the bob of the pendulum can be written as

$$x = a \cos (\omega t + \phi) \tag{21}$$

the velocity of the particle is given by

$$\frac{dx}{dt} = -a\omega \sin (\omega t + \phi) \tag{22}$$

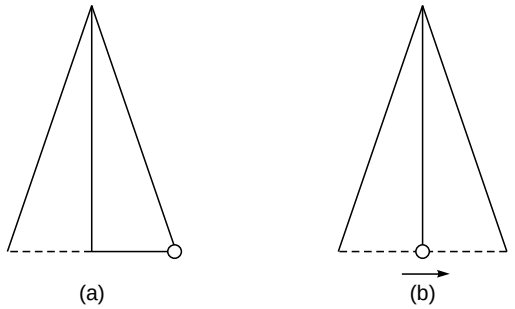


Fig. 7.5 (a) and (b) show the motion of two identical pendulums which are vibrating with the same amplitude but having a phase difference of π/2. The small circles denote the position of the bobs at $t = 0$.

Thus the kinetic energy of the mass is

$$T = \frac{1}{2} m \left(\frac{dx}{dt} \right)^2$$

$$= \frac{1}{2} ma^2\omega^2 \sin^2 (\omega t + \phi) \tag{23}$$

Comparing Eqs. (21) and (23), we see that when the particle is at its extreme positions, the kinetic energy is zero; and when the particle passes through the equilibrium position, the kinetic energy is maximum. At the extreme positions, the kinetic energy gets transformed to potential energy. From Fig. 7.4(a) it can immediately be seen that

Potential energy $V = mgh = mgl (1 - \cos \theta)$

$$= mgl \left(2 \sin^2 \frac{\theta}{2} \right)$$

$$\approx 2\, mgl \left(\frac{\theta}{2} \right)^2$$

(θ measured in radians)

$$\approx \frac{1}{2} mgl \left(\frac{x}{l} \right)^2 = \frac{1}{2} m \left(\frac{g}{l} \right) x^2$$

$$= \frac{1}{2} m\omega^2 x^2 \tag{24}$$

or

$$V = \frac{1}{2} m \omega^2 a^2 \cos^2(\omega t + \phi) \tag{25}$$

where we have used the fact that $\omega^2 = g/l$. The expression for potential energy could have been directly written down by noting that if the potential energies at *x* and at $x + dx$ are *V* and $V + dV$, then

$$dV = -F\, dx = +kx\, dx \tag{26}$$

[1] We will be assuming that θ is small so that $\sin \theta \approx \theta$, where θ is in radians. The above approximation is valid for $\theta \lesssim 0.07$ rad (≈ 4°).

Thus

$$V = \int_0^x kx\,dx = \frac{1}{2}kx^2 \qquad \textbf{(27)}$$

where we have assumed the zero of the potential energy to be at $x = 0$. Thus the total energy E is given by

$$E = T + V = \frac{1}{2}m\omega^2 a^2 \qquad \textbf{(28)}$$

which, as expected, is independent of time. We can also see from Eq. (26) that the energy associated with the simple harmonic motion is proportional to the square of the amplitude and the square of the frequency.

(b) Vibrations of a mass held by two stretched springs Another simple example is the motion of a mass m, held by two stretched springs on a smooth table, as shown in Fig. 7.6. The two springs are of natural length l_0 [Fig. 7.6(a)], and corresponding to the equilibrium position of the mass, the lengths of the stretched springs are l. If the mass is displaced slightly from the equilibrium position, then the resultant force acting on the mass will be

$$F = k[(l - x) - l_0] - k[(l + x) - l_0]$$
$$= -2kx \qquad \textbf{(29)}$$

where k represents the force constant of the spring. Once again we get a force which is proportional to the displacement and directed toward the equilibrium position, and consequently, the motion of the mass on the frictionless table will be simple harmonic.

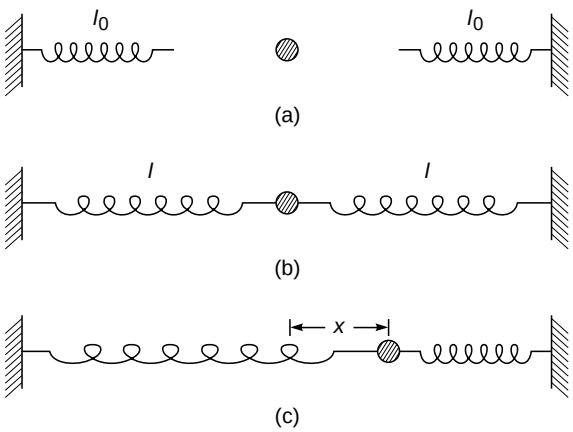


Fig. 7.6 Two springs of natural length l_0 [see (a)] are stretched to a length l [see (b)] to hold the mass. If the mass is displaced by a small distance x from its equilibrium position [see (c)], the mass will execute simple harmonic motion.

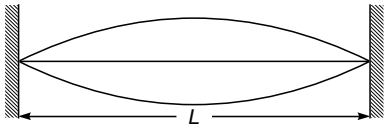


Fig. 7.7 When a string clamped at both the ends is made to vibrate in its fundamental mode, all particles execute simple harmonic motions with the same frequency and same initial phase but having different amplitudes.

(c) Vibrations of a stretched string When a stretched string (as in a sonometer) is made to vibrate in its fundamental mode (see Fig. 7.7), then each point on the string executes simple harmonic motion with different amplitudes but having the same initial phase. The displacement can be written in the form

$$y = a\sin\left(\frac{\pi}{L}x\right)\cos\omega t \qquad \textbf{(30)}$$

The amplitude is therefore zero at $x = 0$ and at $x = L$ and is maximum at $x = L/2$. On the other hand, if the string is vibrating in its first harmonic, then each point on the first half of the string vibrates out of phase with each point on the other half.

7.3 DAMPED SIMPLE HARMONIC MOTION

In Sec. 7.2, we showed that for a particle executing simple harmonic motion (SHM), the equation of motion will be of the form

$$\frac{d^2x}{dt^2} + \omega_0^2 x(t) = 0 \qquad \textbf{(31)}$$

the solution of which is given by

$$x(t) = A\cos(\omega_0 t + \theta) \qquad \textbf{(32)}$$

where A represents amplitude and ω_0 the angular frequency of motion. Equation (32) tells us that the motion will continue forever. However, we know that in actual practice the amplitude of any vibrating system (like that of a tuning fork) keeps on decreasing, and eventually the system stops vibrating. Similarly, the bob of a pendulum comes to rest after a certain time. This phenomenon is due to the presence of damping forces which come into play when the particle is in motion. For a vibrating pendulum, the damping forces are primarily due to the viscosity of the surrounding medium. Consequently, the damping forces will be much larger in liquids than in gases. In general, the exact dependence of

the damping force on the velocity of the particle is quite complicated; however, as a first approximation we may assume it to be proportional to the velocity of the particle. This is also consistent with the fact that there are no damping forces acting on the particle when it is at rest. In this model, the equation of motion will be given by

$$m\,\frac{d^2x}{dt^2} = -\Gamma\,\frac{dx}{dt} - k_0 x \qquad \textbf{(33)}$$

where the constant Γ determines the strength of the damping force; the force constant is now denoted by k_0 to avoid confusion with the wave vector k. Equation (33) can be rewritten in the form

$$\frac{d^2x}{dt^2} + 2K\,\frac{dx}{dt} + \omega_0^2 x(t) = 0 \qquad \textbf{(34)}$$

where

$$2K = \frac{\Gamma}{m} \quad \text{and} \quad \omega_0 = \sqrt{\frac{k_0}{m}} \qquad \textbf{(35)}$$

To solve Eq. (34), we introduce a new variable $\xi(t)$ which is defined by the

$$x(t) = \xi(t)e^{-Kt} \qquad \textbf{(36)}$$

Thus,

$$\frac{dx}{dt} = \left[\frac{d\xi}{dt} - K\,\xi(t)\right]e^{-Kt}$$

and

$$\frac{d^2x}{dt^2} = \left[\frac{d^2\xi}{dt^2} - 2K\,\frac{d\xi}{dt} + K^2\xi(t)\right]e^{-Kt}$$

On substitution in Eq. (34) we get

$$\frac{d^2\xi}{dt^2} + (\omega_0^2 - K^2)\xi(t) = 0 \qquad \textbf{(37)}$$

Equation (37) is similar to Eq. (31); however, depending on the strength of the damping force, the quantity $\omega_0^2 - K^2$ can be positive, negative, or zero. Consequently, we must consider three cases.

Case 1: $\omega_0^2 > K^2$

If the damping is small, ω_0^2 is greater than K^2, and the solution of Eq. (37) is of the form

$$\xi(t) = A\cos\left(\sqrt{\omega_0^2 - K^2}\;t + \theta\right) \qquad \textbf{(38)}$$

or

$$x(t) = Ae^{-Kt}\cos\left(\sqrt{\omega_0^2 - K^2}\;t + \theta\right) \qquad \textbf{(39)}$$

where A and θ are constants which are determined from the amplitude and phase of the motion at $t = 0$. Equation (39) represents a damped simple harmonic motion (see Fig. 7.8). Notice that the amplitude decreases exponentially with time

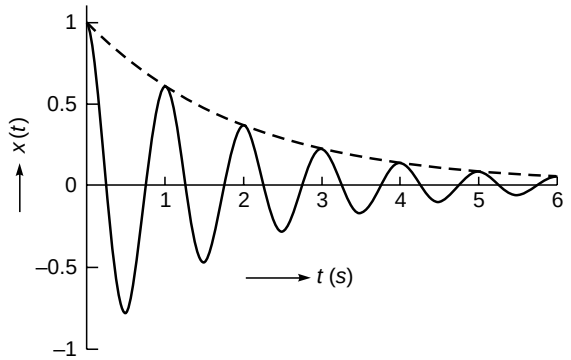


Fig. 7.8 The exponential decrease of amplitude in a damped simple harmonic motion. The figure corresponds to $\frac{2\pi}{\sqrt{\omega_0^2 - K^2}} = 1$ s and $K = 0.5\ \text{s}^{-1}$.

and the time period of vibration $\left(= 2\pi\Big/\sqrt{\omega_0^2 - K^2}\right)$ is greater than in the absence of damping.

Case 2: $K^2 > \omega_0^2$

If the damping is too large, K^2 is greater than ω_0^2, and Eq. (37) should be written in the form

$$\frac{d^2\xi}{dt^2} - (K^2 - \omega_0^2)\xi(t) = 0 \qquad \textbf{(40)}$$

the solution of which is given by

$$\xi(t) = A\exp\left(\sqrt{K^2 - \omega_0^2}\;t\right) + B\exp\left(-\sqrt{K^2 - \omega_0^2}\;t\right) \qquad \textbf{(41)}$$

Thus,

$$x(t) = A\exp\left[\left(-K + \sqrt{K^2 - \omega_0^2}\right)t\right] + B\exp\left[\left(-K - \sqrt{K^2 - \omega_0^2}\right)t\right] \qquad \textbf{(42)}$$

and we can have two kinds of motion; one in which the displacement decreases uniformly to zero or the other in which the displacement first increases, reaches a maximum, and then decreases to zero (see Fig. 7.9). In either case there are no oscillations, and the motion is said to be *overdamped* or *dead beat*. A typical example is the motion of a simple pendulum in a highly viscous liquid (such as glycerine) where the pendulum can hardly complete a fraction of the vibration before coming to rest.

Case 3: $K^2 = \omega_0^2$

When $K^2 = \omega_0^2$, Eq. (37) becomes

$$\frac{d^2\xi}{dt^2} = 0 \qquad \textbf{(43)}$$

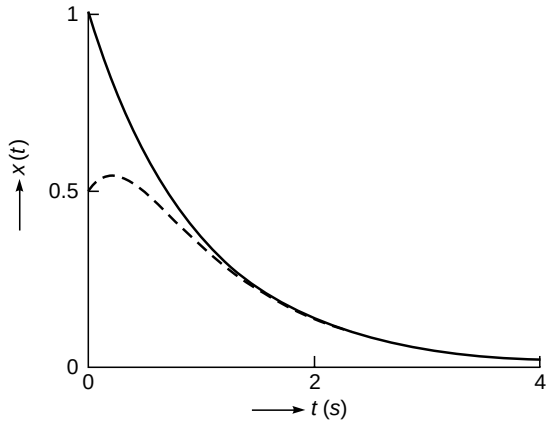


Fig. 7.9 The variation of displacement with time in an overdamped motion. The solid and the dashed curves correspond to $B = 0$ and $B = -A/2$, respectively [see Eq. (42)]. In carrying out the calculations we have assumed $K = 2\ s^{-1}$ and $\sqrt{K^2 - \omega_0^2} = 1\ s^{-1}$.

the solution of which is given by

$$\xi = At + B \tag{44}$$

Thus

$$x(t) = (At + B)e^{-Kt} \tag{45}$$

The motion is again nonoscillatory and is said to correspond to *critical damping*.

7.4 FORCED VIBRATIONS

We consider the effect of a periodic sinusoidal force (see also Sec. 8.3) on the motion of a vibrating system. If the frequency of the external force is ω, then the equation of motion is [cf. Eq. (33)]:

$$m\frac{d^2x}{dt^2} = F\cos\omega t - \Gamma\frac{dx}{dt} - k_0 x \tag{46}$$

where the first term on the RHS represents the external force; the other terms are the same as in Eq. (33). Equation (46) is rewritten in the form[2]

$$\frac{d^2x}{dt^2} + 2K\frac{dx}{dt} + \omega_0^2 x(t) = G\cos\omega t \tag{47}$$

where $G = F/m$ and the other symbols were defined in Sec. 7.3. For the particular solution of Eq. (47) we try

$$x(t) = a\cos(\omega t - \phi) \tag{48}$$

Thus,

$$\frac{dx}{dt} = -a\omega\sin(\omega t - \phi)$$

and

$$\frac{d^2x}{dt^2} = -a\omega^2\cos(\omega t - \phi)$$

Substituting the above forms for $x(t)$, dx/dt, and d^2x/dt^2 in Eq. (47), we obtain

$$-a\omega^2\cos(\omega t - \phi) - 2Ka\omega\sin(\omega t - \phi) + a\omega_0^2\cos(\omega t - \phi) = G\cos[(\omega t - \phi) + \phi] \tag{49}$$

where we have written $G\cos\omega t$ as $G\cos[(\omega t - \phi) + \phi]$. Thus,

$$a(\omega_0^2 - \omega^2)\cos(\omega t - \phi) - 2Ka\omega\sin(\omega t - \phi) = G\cos(\omega t - \phi)\cos\phi - G\sin(\omega t - \phi)\sin\phi \tag{50}$$

For Eq. (50) to be valid for all values of time, we must have

$$a(\omega_0^2 - \omega^2) = G\cos\phi \tag{51}$$

$$2Ka\omega = G\sin\phi \tag{52}$$

If we square and add, we get

$$a = \frac{G}{\left[(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2\right]^{1/2}} \tag{53}$$

Further

$$\tan\phi = \frac{2K\omega}{\omega_0^2 - \omega^2} \tag{54}$$

Since K, ω, and a are positive, φ is uniquely determined by noting that $\sin\phi$ should be positive; i.e., φ must be in either the first or the second quadrant.

To the solution given by Eq. (48), we must add the solution of the homogeneous equation, Eq. (34). Thus, assuming ω_0^2 to be greater than K^2 (i.e., weak damping), the general solution of Eq. (47) will be of the form

$$x(t) = Ae^{-Kt}\cos\left(\sqrt{\omega_0^2 - K^2}\ t - \theta\right) + a\cos(\omega t - \phi) \tag{55}$$

The first term on the RHS represents the transient solution corresponding to the natural vibrations of the system which eventually die out. The second term represents the steady-state solution which corresponds to the forced vibrations imposed by the external force. Notice that the frequency of the forced vibrations is the same as that of the external force.

[2] Notice that the RHS of Eq. (47) is independent of x; such an equation is said to be an inhomogeneous equation. An equation of the type given by Eq. (34) is said to be homogeneous.

7.4.1 Resonance

The amplitude of the forced vibration

$$a = \frac{G}{\left[(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2\right]^{1/2}} \quad \textbf{(56)}$$

depends on the frequency of the driving force and is a maximum when $(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2$ is a minimum, i.e., when

$$\frac{d}{d\omega}\left[(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2\right] = 0$$

or $$2(\omega_0^2 - \omega^2)(-2\omega) + 8K^2\omega = 0$$

or $$\omega = \omega_0\left(1 - \frac{2K^2}{\omega_0^2}\right)^{1/2} \quad \textbf{(57)}$$

Thus the amplitude is maximum[3] when ω is given by Eq. (57). This is known as *amplitude resonance*. When damping is extremely small, the resonance occurs at a frequency very close to the natural frequency of the system. The variation of the amplitude with ω is shown in Fig. 7.10. Notice that as the damping decreases, the maximum becomes very sharp and the amplitude falls off rapidly as we go away from the resonance. The maximum value of *a* is given by

$$a_{\max} = \frac{G}{\left[(2K^2)^2 + 4K^2\omega_0^2\left(1 - 2K^2/\omega_0^2\right)\right]^{1/2}}$$

$$= \frac{G}{2K\left(\omega_0^2 - K^2\right)^{1/2}} = \frac{G}{2K\left(\omega^2 + K^2\right)^{1/2}} \quad \textbf{(58)}$$

Thus, with increase in damping, the maximum occurs at lower values of ω and the resonance becomes less sharp.

To discuss the phase of the forced vibrations, we refer to Eq. (54) from where we find that for small damping the phase angle is small unless it is near resonance. For $\omega = \omega_0$, $\tan\phi = \infty$ and ϕ is $\pi/2$; i.e., the phase of forced vibrations is $\pi/2$ *ahead* of the phase of the driving force. As the frequency of the driving force is increased beyond ω_0, the phase also increases and approaches π (see Fig. 7.11).

All the salient features of forced vibrations can be easily demonstrated by means of an arrangement shown in Fig. 7.12. In the figure, *AC* is a metal rod with a movable bob *B*, and *LM* is a simple pendulum with a bob at *M*. The metal rod and the simple pendulum are suspended from a string *PQ*, as shown in Fig. 7.12. With *B* at the bottom, when the rod *AC* is set in motion, the pendulum *LM* also vibrates. As the bob *B* is moved upward, the time period decreases, the frequency of the rod becomes closer to the natural frequency

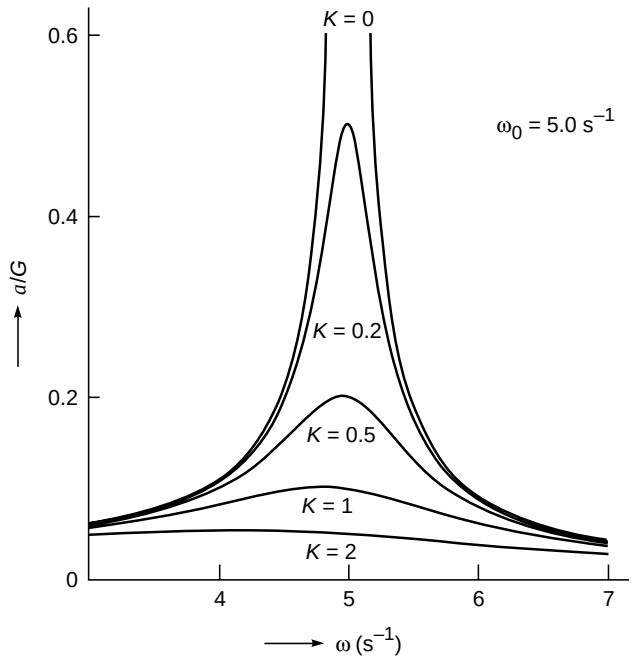


Fig. 7.10 The variation of amplitude with the frequency of the external driving force for various values of *K*. The calculations correspond to $\omega_0 = 5\ s^{-1}$, and the values of *K* are in s^{-1}. Notice that with an increase in damping the resonance occurs at a smaller value of ω.

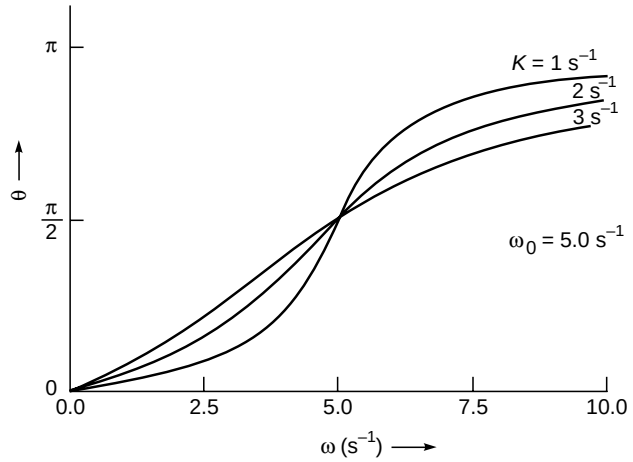


Fig. 7.11 The dependence of the phase of the forced vibration on the frequency of the driving force.

[3] There is no resonance condition when $K^2 \geq \frac{1}{2}\omega_0^2$.

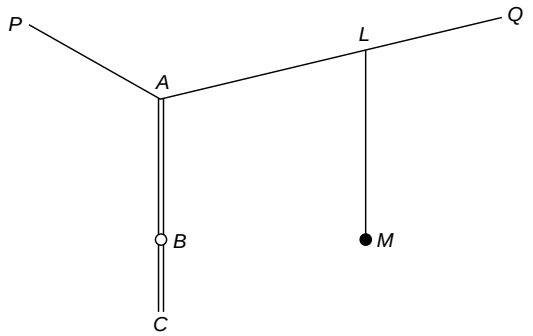


Fig. 7.12 An arrangement for demonstration of forced vibrations.

of the simple pendulum, and eventually the resonance condition is satisfied. At resonance, the amplitude of vibration of the simple pendulum is maximum, and the phase difference between the vibrations is nearly $\pi/2$; i.e., when the metal rod is at its lowest position and moving toward the right, the simple pendulum is at the extreme left position. If the bob *B* is further moved upward, the frequency increases and the amplitude of the forced vibrations decreases.

7.5 ORIGIN OF REFRACTIVE INDEX

In this section we study the origin of the refractive index. We know that an atom consists of a heavy positively charged nucleus surrounded by electrons. In the simplest model of the atom, the electrons are assumed to be bound elastically to their rest positions; thus, when these electrons are displaced by an electric field, a restoring force (proportional to the displacement) will act on the electrons that will tend to return the electrons to their rest positions. In this model, the equation of motion for the electron, in the presence of an external electric field **E**, would be

$$m\frac{d^2\mathbf{x}}{dt^2} + k_0\mathbf{x} = -q\mathbf{E} \qquad \textbf{(59)}$$

or

$$\frac{d^2\mathbf{x}}{dt^2} + \omega_0^2\mathbf{x} = -\frac{q}{m}\mathbf{E} \qquad \textbf{(60)}$$

where **x** represents the position of the electron, m and $-q$ represent the mass and charge, respectively, of the electron ($q \approx +1.6 \times 10^{-19}$ C), k_0 is the force constant, and $\omega_0 \left(=\sqrt{k_0/m}\right)$ represents the frequency of the oscillator. We assume

$$\mathbf{E} = \hat{\mathbf{x}}\, E_0 \cos(kz - \omega t) \qquad \textbf{(61)}$$

i.e., the field is in the x direction having an amplitude E_0 and propagating in the $+z$ direction; $\hat{\mathbf{x}}$ represents the unit vector in the x direction; and $k = 2\pi/\lambda$, with λ representing the wavelength. Thus

$$\frac{d^2x}{dt^2} + \omega_0^2 x = -\frac{qE_0}{m}\cos(kz - \omega t) \qquad \textbf{(62)}$$

where we have replaced the vectors by the corresponding scalar quantities because the displacement and the electric field are in the same direction. Except for the damping term, Eq. (62) is similar to Eq. (46), and therefore the solution corresponding to the forced vibrations will be given by[4]

$$x = -\frac{qE_0}{m(\omega_0^2 - \omega^2)}\cos(kz - \omega t) \qquad \textbf{(63)}$$

In the simplest model of the atom, the center of the negative charge (due to the electrons) is assumed to be at the center of the nucleus. In the presence of an electric field, the center of the negative charge gets displaced from the nucleus which results in a finite value of the dipole moment of the atom. In particular, if we have a positive charge $+q$ at the origin and a negative charge $-q$ at a distance x, then the dipole moment is $-qx$; thus, if there are N dispersion electrons[5] per unit volume, then the polarization (i.e., dipole moment per unit volume) is given by

$$\mathbf{P} = -Nq\mathbf{x} = \frac{Nq^2}{m(\omega_0^2 - \omega^2)}\mathbf{E}$$

$$= \chi\mathbf{E} \qquad \textbf{(64)}$$

where

$$\chi = \frac{Nq^2}{m(\omega_0^2 - \omega^2)} \qquad \textbf{(65)}$$

is known as the electric susceptibility of the material. The dielectric permittivity is therefore given by (see Sec. 23.9)

$$\varepsilon = \varepsilon_0 + \chi \qquad \textbf{(66)}$$

or

$$\frac{\varepsilon}{\varepsilon_0} = 1 + \frac{Nq^2}{m\varepsilon_0(\omega_0^2 - \omega^2)} \qquad \textbf{(67)}$$

Now, $\varepsilon/\varepsilon_0$ is the dielectric constant, which is equal to the square of the refractive index (see Chap. 23). Thus

$$n^2 = 1 + \frac{Nq^2}{m\varepsilon_0\omega_0^2}\left(1 - \frac{\omega^2}{\omega_0^2}\right)^{-1} \qquad \textbf{(68)}$$

[4] Notice that in the absence of damping (i.e., when $\Gamma = 0$), $\phi = 0$; see Eq. (54).

[5] The number of dispersion electrons in a molecule of an ideal gas is the valence number of the molecules. This number is 2 for H_2, 6 for N_2, etc.

showing that the refractive index depends on the frequency; this is known as dispersion. Assuming that the characteristic frequency ω_0 lies in the far ultraviolet [see Eq. (74)],[6] the quantity $\left(1-\omega^2/\omega_0^2\right)^{-1}$ is positive in the entire visible region. Further, as ω increases, n^2 also increases, i.e., the refractive index increases with frequency; this is known as normal dispersion. If we further assume $\omega/\omega_0 << 1$, then

$$\left(1-\frac{\omega^2}{\omega_0^2}\right)^{-1} \approx 1 + \frac{\omega^2}{\omega_0^2}$$

and

$$n^2 \approx 1 + \frac{Nq^2}{m\varepsilon_0\omega_0^2}\left(1+\frac{\omega^2}{\omega_0^2}\right)$$

$$\approx 1 + \frac{Nq^2}{m\varepsilon_0\omega_0^2} + \frac{4\pi^2c^2Nq^2}{m\varepsilon_0\omega_0^4}\,\frac{1}{\lambda_0^2} \qquad \textbf{(69)}$$

where $\lambda_0 = 2\pi c/\omega$ is the free space wavelength. Equation (69) can be written in the form

$$n^2 = A + \frac{B}{\lambda_0^2} \qquad \textbf{(70)}$$

which is the well-known *Cauchy relation*. For hydrogen, the experimental variation of n^2 with λ_0 is approximately given by

$$n^2 = 1 + 2.721\times10^{-4} + \frac{2.11\times10^{-18}}{\lambda_0^2} \qquad \textbf{(71)}$$

where the wavelength is measured in meters; the above numbers correspond to 0°C and 76 cm Hg (see Ref. 9). Thus,

$$\frac{Nq^2}{m\varepsilon_0\omega_0^2} = 2.721\times10^{-4} \qquad \textbf{(72)}$$

and

$$\frac{4\pi^2c^2Nq^2}{m\varepsilon_0\omega_0^4} = 2.11\times10^{-18}\ \text{m}^2 \qquad \textbf{(73)}$$

If we divide Eq. (73) by Eq. (72), we would get

$$\frac{4\pi^2c^2}{\omega_0^2} = \frac{2.11\times10^{-18}}{2.721\times10^{-4}}$$

or

$$\nu_0 = \frac{\omega_0}{2\pi} \approx 3\times10^{15}\ \text{s}^{-1} \qquad \textbf{(74)}$$

which is indeed in the ultraviolet region. We can eliminate ω_0 from Eqs. (72) and (73) to obtain

$$\frac{Nq^2}{4\pi^2c^2\varepsilon_0 m} \approx 3\times10^{10}\ \text{m}^{-2} \qquad \textbf{(75)}$$

Now at NTP, 22,400 cm^3 of H_2 contains 6×10^{23} molecules; thus,

$$N = 2\times\frac{6\times10^{23}}{22,400\times10^{-5}}\ \text{m}^{-3} \approx 5\times10^{25}\ \text{m}^{-3}$$

where the factor 2 arises from the fact that a hydrogen molecule consists of two electrons. Hence,

$$\frac{Nq^2}{4\pi^2c^2\varepsilon_0 m}$$

$$\approx \frac{5\times10^{25}\times(1.6\times10^{-19})^2}{4\times\pi^2\times9\times10^{16}\times8.85\times10^{-12}\times9.1\times10^{-31}}$$

$$\approx 4\times10^{10}\ \text{m}^{-2}$$

which qualitatively agrees with Eq. (75).

We note that for a gas of free electrons (as we have in the upper atmosphere), there is no restoring force and we must set $\omega_0 = 0$. Thus the expression for the refractive index becomes [see Eq. (67)]

$$n^2 = 1 - \frac{Nq^2}{m\varepsilon_0\omega^2} \qquad \textbf{(76)}$$

where N represents the density of free electrons. Equation (75) shows that the refractive index is less than unity; however, this does not imply that one can send signals faster than the speed of light in free space (see Chap. 10). To quote Feynman:

> For free electrons, $\omega_0 = 0$ (there is no elastic restoring force). Setting $\omega_0 = 0$ in our dispersion equation yields the correct formula for the index of refraction for radiowaves in the stratosphere, where N is now to represent the density of free electrons (number per unit volume) in the stratosphere. But let us look again at the equation, if we beam X-rays on the matter, or radiowaves (or any electric waves) on free electrons, the term $(\omega_0^2 - \omega^2)$ become negative, and we obtain the result that n is less than one. That means that the effective speed of the waves in the substance is faster than c! Can that be correct? It is correct. In spite of the fact that it is

[6] This also follows from the fact that according to classical electrodynamics, an oscillating dipole vibrating with frequency ω_0 will radiate electromagnetic waves with frequency ω_0; and as an example, if we consider hydrogen, then $\hbar\omega_0 \approx 13.6$ eV from which we obtain $\omega_0 \approx 2\times10^{16}$ s^{-1}. This frequency corresponds to the far ultraviolet.

said that you cannot send signals any faster than the speed of light, it is nevertheless true that the index of refraction of materials at a particular frequency can be either greater or less than 1.

Equation (76) is usually written in the form

$$n^2 = 1 - \left(\frac{\omega_p}{\omega}\right)^2 \tag{77}$$

where

$$\omega_p = \left(\frac{Nq^2}{m\varepsilon_0}\right)^{1/2} \tag{78}$$

is known as the plasma frequency. For $\omega < \omega_p$, the refractive index is purely imaginary which gives rise to attenuation, and for $\omega > \omega_p$, the refractive index is real. Indeed in 1933, Wood discovered that alkali metals are transparent to ultraviolet light. For example, for sodium if we assume that the refractive index is primarily due to the free electrons and that there is one free electron per atom, then

$$N = \frac{6 \times 10^{23} \times 0.9712}{22.99} \approx 2.535 \times 10^{22}\ \text{cm}^{-3}$$

where we have assumed that the atomic weight of Na is 22.99 and its density is 0.9712 g/cm^3. Substituting the values of $m \approx 9.109 \times 10^{-31}$ kg, $q \approx 1.602 \times 10^{-19}$ C, and $\varepsilon_0 \approx 8.854 \times 10^{-12}$ C/N^{-1}/m^{-2}, we get

$$\lambda_p \left(= \frac{2\pi c}{\omega_p}\right) \approx 2098\ \text{Å}$$

Thus for $\lambda < 2098$ Å, the refractive index of Na becomes real and the metal would become transparent; the corresponding experimental value is 2100 Å. The theoretical and experimental values of λ_p for Li, K, and Rb are discussed in Prob. 7.7.

As mentioned above, Eq. (76) gives the correct dependence of the refractive index of the stratosphere for radio waves; in Sec. 3.4.3 we used Eq. (76) to study the reflection of electromagnetic waves by the ionosphere.

Returning to Eq. (68), we note that as $\omega \to \omega_0$, the refractive index tends to ∞. This is so because we have neglected the presence of damping forces in our treatment. If we do take into account the damping forces, Eq. (62) becomes [see Eq. (46)]

$$m\frac{d^2x}{dt^2} + \Gamma\frac{dx}{dt} + k_0 x = qE_0 \cos(kz - \omega t) \tag{79}$$

To derive an expression for the refractive index, it is more convenient to rewrite the above equation in the form

$$\frac{d^2x}{dt^2} + 2K\frac{dx}{dt} + \omega_0^2 x = \frac{qE_0}{m} e^{i(kz - \omega t)} \tag{80}$$

where the solution of Eq. (79) will be the real part of the solution of Eq. (80). The solution of the homogeneous equation will give the transient behavior which will die out as $t \to \infty$ (see Sec. 7.4); the steady-state solution will correspond to frequency ω. Thus, if we substitute a solution of the type

$$x(t) = Ae^{i(kz - \omega t)} \tag{81}$$

in Eq. (80), we obtain

$$(-\omega^2 - 2iK\omega + \omega_0^2)A = \frac{qE_0}{m}$$

or[7]

$$A = \frac{qE_0}{m(\omega_0^2 - \omega^2 - 2iK\omega)} \tag{82}$$

Thus we get

$$\mathbf{P} = \frac{Nq^2}{m[\omega_0^2 - \omega^2 - 2iK\omega]}\,\mathbf{E} \tag{83}$$

The electric susceptibility is therefore given by

$$\chi = \frac{Nq^2}{m(\omega_0^2 - \omega^2 - 2iK\omega)}$$

Thus

$$n^2 = \frac{\varepsilon}{\varepsilon_0} = 1 + \frac{\chi}{\varepsilon_0}$$

$$= 1 + \frac{Nq^2}{m\varepsilon_0(\omega_0^2 - \omega^2 - 2iK\omega)} \tag{84}$$

Notice that the refractive index is complex, which implies absorption of the propagating electromagnetic wave. Indeed, if we write

$$n = \eta + i\kappa \tag{85}$$

where η and κ are real numbers, then the wave number k, which equals $n\omega/c$, is given by

$$k = (\eta + i\kappa)\frac{\omega}{c} \tag{86}$$

[7] Notice that A is complex; however, if we substitute the expression for A from Eq. (82) in Eq. (81) and take the real part, we get the same expression for $x(t)$ as we obtained in Sec. 7.4.

If we consider a plane electromagnetic wave propagating in the $+z$ direction, then its z and t dependence will be of the form $\exp[i(kz - \omega t)]$; consequently

$$\mathbf{E} = \mathbf{E_0} e^{i(kz-\omega t)}$$
$$= \mathbf{E_0} \exp\left\{ i\left[(\eta + i\kappa)\frac{\omega}{c} z - \omega t \right] \right\}$$
$$= \mathbf{E_0} \exp\left[-i\omega\left(t - \frac{\eta z}{c} \right) - \frac{\kappa\omega}{c} z \right] \qquad \textbf{(87)}$$

which shows an exponential attenuation of the amplitude. This should not be unexpected because damping causes a loss of energy.

To obtain expressions for η and κ, we substitute the expression for n from Eq. (85) into Eq. (84) to obtain

$$(\eta + i\kappa)^2 = 1 + \frac{Nq^2(\omega_0^2 - \omega^2 + 2iK\omega)}{m\varepsilon_0(\omega_0^2 - \omega^2 - 2iK\omega)(\omega_0^2 - \omega^2 + 2iK\omega)}$$

or

$$\eta^2 - \kappa^2 = 1 + \frac{Nq^2(\omega_0^2 - \omega^2)}{m\varepsilon_0[(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2]} \qquad \textbf{(88)}$$

and $$2\eta\kappa = \frac{Nq^2}{m\varepsilon_0} \frac{2K\omega}{(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2} \qquad \textbf{(89)}$$

The above equations can be rewritten in the form

$$\eta^2 - \kappa^2 = 1 - \frac{\alpha\Omega}{\Omega^2 + \beta^2(\Omega + 1)} \qquad \textbf{(90)}$$

and $$2\eta\kappa = \frac{\alpha\beta\sqrt{1+\Omega}}{\Omega^2 + \beta^2(\Omega + 1)} \qquad \textbf{(91)}$$

where we have introduced the following dimensionless parameters:

$$\alpha = \frac{Nq^2}{m\varepsilon_0\omega_0^2}; \quad \Omega = \frac{\omega^2 - \omega_0^2}{\omega_0^2} \quad \text{and} \quad \beta = \frac{2K}{\omega_0}$$

The qualitative variations of $\eta^2 - \kappa^2$ and $2\eta\kappa$ with Ω are shown in Fig. 7.13. It can be easily shown that at $\Omega = -\beta$ and at $\Omega = +\beta$, the function $\eta^2 - \kappa^2$ attains its maximum and minimum values, respectively.

In general, an atom can execute oscillations corresponding to different resonant frequencies, and we have to take into account the various contributions. If $\omega_0, \omega_1, \ldots$ represent the resonant frequencies and if f_j represents the fractional number of electrons per unit volume whose resonant frequency is ω_j, Eq. (84) is modified to:[8]

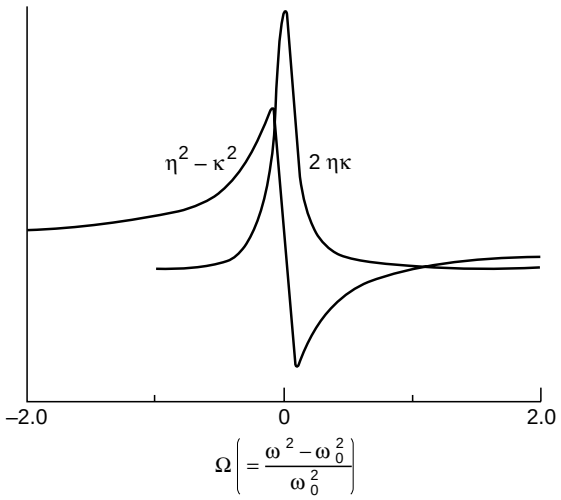


Fig. 7.13 Qualitative variation of $(\eta^2 - \kappa^2)$ and $2\eta\kappa$ with Ω.

$$n^2 = 1 + \frac{Nq^2}{m\varepsilon_0} \sum_j \frac{f_j}{\omega_j^2 - \omega^2 - 2iK_j\omega} \qquad \textbf{(92)}$$

where K_j represents the damping constant corresponding to the resonant frequency ω_j. Indeed, Eq. (92) describes correctly the variation of refractive index for most gases. Figure 7.14 shows the dependence of the refractive index of sodium vapor around $\lambda_0 = 5800$ Å. Since D_1 and D_2 lines occur at 5890 and 5896 Å, respectively one should expect

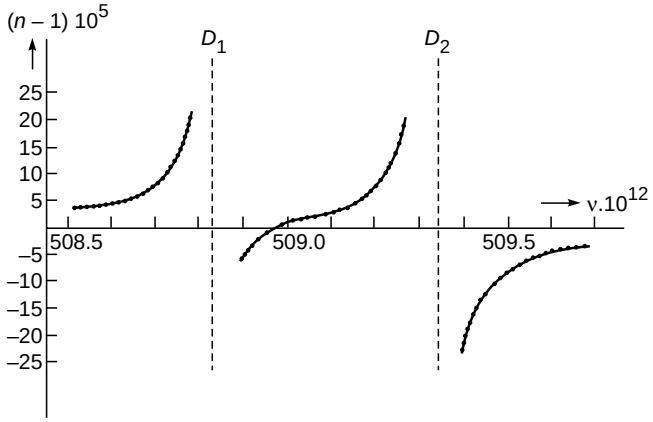


Fig. 7.14 The measured variation of refractive index of sodium with frequency around the D_1 and D_2 lines. The measurements are of Roschdestwensky; the figure has been adapted from Ref. 1.

[8] Quantum mechanics also gives a similar result (see, for example, Ref. 6).

resonant oscillations around these frequencies. This is indeed borne out by the data shown in Fig. 7.14. The variation of the refractive index can be accurately fitted with the formula

$$n^2 = 1 + \frac{A}{\nu^2 - \nu_1^2} + \frac{B}{\nu^2 - \nu_2^2} \tag{93}$$

where we have neglected the presence of damping forces, which is justified except when one is very close to the resonance.

In a liquid, the molecules are very close to one another, and the dipoles interact between themselves. If we take this interaction into account, we get[9]

$$\frac{n^2 - 1}{n^2 + 2} = \frac{Nq^2}{3m\varepsilon_0} \sum_j \frac{f_j}{\omega_j^2 - \omega^2} \tag{94}$$

where we have neglected the presence of damping. For liquids, whose molecules do not have a permanent dipole moment (e.g., H_2, O_2, etc.), Eq. (94) gives a fairly accurate description. However, for liquids whose moleculess possess permanent dipole moments (e.g., H_2O) one has to carry out a different analysis.

7.6 RAYLEIGH SCATTERING

We end this chapter by giving a brief account of Rayleigh scattering. Throughout our analysis we assume that each scattering center behaves independently—an assumption which will be valid for a gas where the average interatomic spacing is greater than the wavelength.

As discussed in Sec. 7.5, the incident electric field **E** produces a dipole moment given by [see Eqs. (64) and (65)]

$$\mathbf{p} = \frac{q^2}{m(\omega_0^2 - \omega^2)}\,\mathbf{E} \tag{95}$$

where ω_0 represents the natural frequency of the atom. To keep the analysis simple, we are neglecting the effect of damping although it can be taken into account without much difficulty. Now, an oscillating dipole given by

$$\mathbf{p} = \mathbf{p}_0\, e^{-i\omega t} \tag{96}$$

radiates energy at a rate (see Sec. 23.4.1)

$$\overline{P} = \frac{\omega^4 p_0^2}{12\pi\varepsilon_0 c^3} \tag{97}$$

or

$$\overline{P} = \frac{\omega^4}{12\pi\varepsilon_0 c^3}\,\frac{q^4}{m^2(\omega_0^2 - \omega^2)^2}\,E_0^2 \tag{98}$$

Thus if N represents the number of atoms per unit volume, then the total energy radiated away (per unit volume) is be $N\overline{P}$.

We assume the electromagnetic wave to be propagating along the x direction. The intensity of the wave is given by [see Eq. (78) of Chap. 23]

$$I = \frac{1}{2}\varepsilon_0 c E_0^{\,2} \tag{99}$$

Thus the change in the intensity of the electromagnetic wave as it propagates through a distance dx is given by

$$dI = -N\overline{P}\,dx$$

or

$$\frac{dI}{I} = -\gamma dx \tag{100}$$

where

$$\gamma = \frac{N\omega^4}{6\pi\varepsilon_0^2 c^4}\,\frac{q^4}{m^2(\omega_0^2 - \omega^2)^2} \tag{101}$$

The integration of Eq. (100) is simply

$$I = I_0 e^{-\gamma x} \tag{102}$$

implying that γ represents the attenuation coefficient. For most atoms ω_0 lies in the ultraviolet region; for example, for the hydrogen atom $\hbar\omega_0 \approx$ few electronvolts. Thus if we assume $\omega << \omega_0$, then γ becomes proportional to ω^4, or

$$\gamma \propto \frac{1}{\lambda^4} \tag{103}$$

which represents the famous $1/\lambda^4$ Rayleigh scattering law and is responsible for the blue color of the sky (because it is the blue component which is predominantly scattered). Similarly, the blue component of the light coming from the setting Sun is predominantly scattered out, resulting in the red color of the setting Sun. Indeed, if the color of the setting (or rising) Sun is deep red, one can infer that the pollution level is high. Now, for a gas,

$$n^2 - 1 = \frac{Nq^2}{m\varepsilon_0(\omega_0^2 - \omega^2)} \tag{104}$$

[see Eq. (68)]. For air, since the refractive index is very close to unity, we may write

$$n - 1 \approx \frac{Nq^2}{2m\varepsilon_0(\omega_0^2 - \omega^2)} \tag{105}$$

By using Eq. (105), Eq. (101) can be written in the convenient form

$$\gamma = \frac{2}{3\pi N}\left(\frac{\omega}{c}\right)^4 (n-1)^2$$

$$= \frac{2k^4}{3\pi N}(n-1)^2 \qquad k = \frac{\omega}{c} \tag{106}$$

[9] See, for example, Ref. 1. Notice that when n is very close to unity (i.e., for a dilute fluid), Eq. (94) reduces to Eq. (92).

For air at NTP, the quantity $n - 1 \approx 2.78 \times 10^{-4}$ in the entire region of the visible spectrum. With $N \approx 2.7 \times 10^{19}$ molecules/cm^3 we obtain

$$L = \frac{1}{\gamma} = 27, 128, \text{and } 188 \text{ km}$$

for λ = 4000 Å (violet), 5900 Å (yellow), and 6500 Å (red), respectively. The quantity L represents the distance in which the intensity decreases by a factor of e.

Rayleigh scattering discussed above is caused by particles whose dimensions are small compared to wavelength (like oxygen and nitrogen molecules in the atmosphere); this is why the sky is blue. Scattering of light by larger size particles (whose size are large compared to the wavelength like in a colloidal suspension) is known as Tyndall scattering and all wavelengths are scattered equally. This is the reason why the clouds (which contain water droplets) appear white against the background of the blue sky.

We conclude this chapter by mentioning that in the 1929 edition of *Encyclopaedia Britannica*, Lord Rayleigh wrote in an article on SKY:

> **SKY:** The apparent covering of the atmosphere, the overarching heaven. . . It is a matter of common observation that the blue of the sky is highly variable, even on days that are free from clouds. The color usually deepens toward the zenith and also with the elevation of the observer. . . Closely associated with the color is the polarization of light from the sky. This takes place in a plane passing through the Sun, and attains a maximum about 90° therefrom.

Summary

- The most fundamental vibration associated with wave motion is the simple harmonic motion.
- When a point rotates on the circumference of a circle with a uniform angular velocity, the foot of the perpendicular on any one of its diameters will execute simple harmonic motion.
- When an external sinusoidal force is applied to a vibrating system, we have forced vibrations. In steady state, the frequency of the forced vibrations is the same as that of the external force.
- When a light wave interacts with an atom, we may assume the electrons to behave as oscillators with resonant frequency ω_0. The electric field of the light wave polarizes the molecules of the gas, producing oscillating dipole moments from which we can make a first principle calculation of the refractive index to obtain

$$n^2(\omega) \approx 1 + \frac{Nq^2}{m\varepsilon_0(\omega_0^2 - \omega^2 - 2iK\omega)}$$

where m is the mass of the electron, q the magnitude of the charge of the electron, N is the number of electrons per unit volume, and K is the damping constant. Because of the fact that an oscillating dipole radiates energy, the light wave gets attenuated; this leads to the famous $1/\lambda^4$ Rayleigh scattering law which is responsible for the red color of the rising Sun and blue color of the sky.

Problems

7.1 The displacement in a string is given by

$$y(x, t) = a \cos\left(\frac{2\pi}{\lambda}x - 2\pi\nu t\right)$$

where a, λ, and ν represent the amplitude, wavelength, and the frequency, respectively, of the wave. Assume a = 0.1 cm, λ = 4 cm, and ν = 1 s^{-1}. Plot the time dependence of the displacement at x = 0, 0.5, 1.0, 1.5, 2, 3, and 4 cm. Interpret the plots physically.

[**Ans:** $y(x = 3.0, t) = -y(x = 1.0, t)$ because the two points are $\frac{\lambda}{2}$ apart; etc.]

7.2 The displacement associated with a standing wave on a sonometer is given by

$$y(x, t) = 2a \sin\left(\frac{2\pi}{\lambda}x\right)\cos 2\pi\nu t$$

If the length of the string is L, then the allowed values of λ are $2L$, $2L/2$, $2L/3$, . . . (see Sec. 13.2). Consider the case when $\lambda = 2L/5$; study the time variation of displacement in each loop, and show that alternate loops vibrate in phase (with different points in a loop having different amplitudes) and adjacent loops vibrate out of phase.

7.3 A tunnel is dug through the Earth as shown in Fig. 7.15. A mass is dropped at point A along the tunnel. Show that it will execute simple harmonic motion. What will the time period be?

[**Ans.:** The time period will be $T = 2\pi\sqrt{\frac{R}{g}}$.]

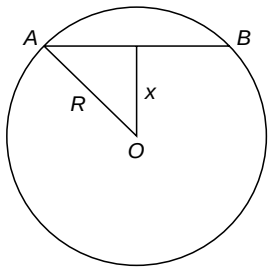


Fig. 7.15 For Prob. 7.3.

7.4 A 1 g mass is suspended from a vertical spring. It executes simple harmonic motion with period 0.1 s. By how much distance had the spring stretched when the mass was attached?

[**Ans:** $\Delta x \approx 0.25$ cm]

7.5 A stretched string is given simultaneous displacement in the x and y directions such that

$$x(z, t) = a \cos\left(\frac{2\pi}{\lambda} z - 2\pi\nu t\right)$$

and $$y(z, t) = a \cos\left(\frac{2\pi}{\lambda} z - 2\pi\nu t\right)$$

Show that the string will vibrate along a direction making an angle $\pi/4$ with the x and y axes.

7.6 In Prob. 7.5, if

$$x(z, t) = a \cos\left(\frac{2\pi}{\lambda} z - 2\pi\nu t\right)$$

and $$y(z, t) = a \sin\left(\frac{2\pi}{\lambda} z - 2\pi\nu t\right)$$

what will be the resultant displacement?

7.7 As mentioned in Sec. 7.5, alkali metals are transparent to ultraviolet light. Assuming that the refractive index is primarily due to the free electrons and that there is one free electron per atom, calculate $\lambda_p \left(= 2\pi c/\omega_p\right)$ for Li, K, and Rb. You may assume that the atomic weights of Li, K, and Rb are 6.94, 39.10, and 85.48, respectively, and that the corresponding densities are 0.534, 0.870, and 1.532 g/cm^3. Also, the values of various physical constants are $m = 9.109 \times 10^{-31}$ kg, $q = 1.602 \times 10^{-19}$ C, and $\varepsilon_0 = 8.854 \times 10^{-12}$/N^{-1}/m^{-2}.

[**Ans:** 1550, 2890, and 3220 Å; the corresponding experimental values are 1551, 3150, and 3400 Å, respectively].

7.8 **(a)** In a metal, the electrons can be assumed to be essentially free. The drift velocity of the electron satisfies the equation

$$m\frac{d\mathbf{v}}{dt} + m\mathbf{v}\nu = \mathbf{F} = -q\mathbf{E_0}e^{-i\omega t}$$

where ν represents the collision frequency. Calculate the steady-state current density ($\mathbf{J} = -Nq\mathbf{v}$) and show that the conductivity is given by

$$\sigma(\omega) = \frac{Nq^2}{m}\frac{1}{\nu - i\omega}$$

(b) If $\mathbf{r}$ represents the displacement of the electron, show that

$$\mathbf{P} = -Nq\mathbf{r} = -\frac{Nq^2}{m(\omega^2 + i\omega\nu)}\mathbf{E}$$

which represents the polarization. Using the above equation, show that

$$\kappa(\omega) = 1 - \frac{Nq^2}{m\varepsilon_0(\omega^2 + i\omega\nu)}$$

which represents the dielectric constant variation for a free-electron gas.

7.9 Assuming that each atom of copper contributes one free electron and that the low-frequency conductivity σ is about 6×10^7 mho/m, show that $\nu \approx 4 \times 10^{13}$ s^{-1}. Using this value of ν, show that the conductivity is almost real for $\omega < 10^{11}$ s^{-1}. For $\omega = 10^8$ s^{-1} calculate the complex dielectric constant, and compare its value with the one obtained for infrared frequencies.

Note that for small frequencies, only one of the electrons of a copper atom can be considered to be free. On the other hand, for X-ray frequencies all the electrons may be assumed to be free (see Probs. 7.10, 7.11, and 7.12). Discuss the validity of the above argument.

7.10 Show that for high frequencies ($\omega >> \nu$) the dielectric constant (as derived in Prob. 7.8) is essentially real with frequency dependence of the form

$$\kappa = 1 - \frac{\omega_p^2}{\omega^2}$$

where $\omega_p = \left(Nq^2/m\varepsilon_0\right)^{1/2}$ is known as the plasma frequency. The above dielectric constant variation is indeed valid for X-ray wavelengths in many metals. Given that at such frequencies all the electrons can be assumed to be free, calculate ω_p for copper for which the atomic number is 29, mass number is 63, and density is 9 g cm^{-3}.

[**Ans:** $\sim 9 \times 10^{16}$ s^{-1}]

7.11 For sodium, at $\lambda = 1$ Å, all the electrons can be assumed to be free; under this assumption show that $\omega_p \approx 3 \times 10^{16}$ s^{-1} and $n^2 \approx 1$ and the metal will be completely transparent.

7.12 In an ionic crystal (such as NaCl, and CaF_2), one has to take into account infrared resonance oscillations of the ions, and Eq. (68) modifies to

$$n^2 = 1 + \frac{Nq^2}{m\varepsilon_0(\omega_1^2 - \omega^2)} + \frac{pNq^2}{M\varepsilon_0(\omega_2^2 - \omega^2)}$$

where M represents the reduced mass of the two ions and p represents the valency of the ion ($p = 1$ for Na^+, Cl^-; $p = 2$ for Ca^{2+}, F_2^{2-}). Show that the above equation can be written in the form[10].

$$n^2 = n_\infty^2 + \frac{A_1}{\lambda^2 - \lambda_1^2} + \frac{A_2}{\lambda^2 - \lambda_2^2}$$

[10] Quoted from Ref. 9; measurements are of Paschen.

where

$$\lambda_1 = \frac{2\pi c}{\omega_1} \qquad \lambda_2 = \frac{2\pi c}{\omega_2}$$

$$A_1 = \frac{Nq^2}{4\pi^2 c^2 \varepsilon_0 m}\lambda_1^4 \qquad A_2 = \frac{pNq^2}{4\pi^2 c^2 \varepsilon_0 M}\lambda_2^4$$

and $n_\infty^2 = 1 + \frac{A_1}{\lambda_1^2} + \frac{A_2}{\lambda_2^2}$

7.13 The refractive index variation for CaF_2 (in the visible region of the spectrum) can be written in the form[10]

$$n^2 = 6.09 + \frac{6.12\times10^{-15}}{\lambda^2 - 8.88\times10^{-15}} + \frac{5.10\times10^{-9}}{\lambda^2 - 1.26\times10^{-9}}$$

where λ is in meters.

(a) Plot the variation of n^2 with λ in the visible region.

(b) From the values of A_1 and A_2 show that $m/M \approx 2.07 \times 10^{-5}$ and compare this with the exact value.

(c) Show that using the constants A_1, A_2, λ_1, and λ_2, we obtain $n_\infty^2 \approx 5.73$, which agrees reasonably well with the experimental value given above.

7.14 **(a)** The refractive index of a plasma (neglecting collisions) is approximately given by (see Sec. 7.6)

$$n^2 = 1 - \frac{\omega_p^2}{\omega^2}$$

where

$$\omega_p = \left(\frac{Nq^2}{m\varepsilon_0}\right)^{1/2} \approx 56.414\ \mathrm{N}^{1/2}\ \mathrm{s}^{-1}$$

is known as the plasma frequency. In the ionosphere, the maximum value of N_0 is $\approx 10^{10}$ to 10^{12} electrons/m^3. Calculate the plasma frequency. Notice that at high frequencies $n^2 \approx 1$; thus high-frequency waves (such as the ones used in TV) are not reflected by the ionosphere. On the other hand, for low frequencies, the refractive index is imaginary (as in a conductor—see Sec. 24.3) and the beam gets reflected. This fact is used in long-distance radio communications (see Fig. 3.20).

(b) Assume that for $x \approx 200$ km, $N = 10^{12}$ electrons/m^3 and that the electron density increases to 2×10^{12} electrons/m^3 at $x \approx 300$ km. For $x < 300$ km, the electron density decreases. Assuming a parabolic variation of N, plot the corresponding refractive index variation.

[**Ans:** For 2×10^5 m $< x < 4 \times 10^5$ m,

$$n^2(x) \approx 1 - \frac{6.4\times10^{15}}{\omega^2}[1 - 5\times10^{-11}(x - 3\times10^5)^2]$$

where ω is measured in s^{-1} and x in m.]

REFERENCES AND SUGGESTED READINGS

1. C. J. F. Bottcher, *Theory of Electric Polarization*, Elsevier Publishing Co., Amsterdam, 1952.
2. H. J. J. Braddick, *Vibrations, Waves and Diffraction*, McGraw-Hill Publishing Co., London, 1965.
3. F. S. Crawford, *Waves and Oscillations: Berkeley Physics Course*, Vol. III, McGraw-Hill Book Co., New York, 1968.
4. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley Publishing Co., Reading, Mass., 1965.
5. A. P. French, *Vibrations and Waves*, Arnold-Heineman India, New Delhi, 1973.
6. R. Loudon, *The Quantum Theory of Light*, Clarendon Press, Oxford, 1973.
7. H. J. Pain, *The Physics of Vibrations and Waves*, John Wiley & Sons, London, 1968.
8. R. Resnick and D. Halliday, *Physics*, Part I, John Wiley & Sons, New York, 1966.
9. A. Sommerfeld, *Optics*, Academic Press, New York, 1964.
10. J. M. Stone, *Radiation and Optics*, McGraw-Hill Book Co., New York, 1963.

Chapter Eight

FOURIER SERIES AND APPLICATIONS

. . . Reimann (in one of his publications in 1867) asserts that when Fourier, in his first paper to the Paris Academy in 1807, stated that a completely arbitrary function could be expressed in such a series, his statement so surprised Lagrange that he denied possibility in the most definite terms. It should also be noted that he (Fourier) was the first to allow that the arbitrary function might be given by different analytical expressions in different parts of the interval.

—H. A. Carslaw (1930)

8.1 INTRODUCTION

Fourier series and Fourier integrals are extensively used in the theory of vibrations and waves. As such, we devote this chapter to the study of Fourier series and Fourier integrals. The results obtained will be used in subsequent chapters. Now, according to Fourier's theorem, any periodic vibration can be expressed as a sum of the sine and cosine functions whose frequencies increase in the ratio of natural numbers. Thus, a periodic function with period T, i.e.,

$$f(t + nT) = f(t) \qquad n = 0, \pm 1, \pm 2, \ldots \tag{1}$$

can be expanded in the form

$$f(t) = \frac{1}{2} a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2n\pi}{T} t\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2n\pi}{T} t\right)$$

$$= \frac{1}{2} a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t \tag{2}$$

where

$$\omega = \frac{2\pi}{T} \tag{3}$$

represents the fundamental frequency. Actually, for the expansion to be possible, the function $f(t)$ must satisfy certain conditions. The conditions are that the function $f(t)$ in one period (i.e., in the interval $t_0 < t < t_0 + T$) must be (1) single-valued, (2) must be piecewise continuous (i.e., it can have at most a finite number of finite discontinuities), and (3) can have only a finite number of maxima and minima. These conditions are known as *Dirichlet's conditions* and are almost always satisfied in all problems that one encounters in physics.

The coefficients a_n and b_n can be easily determined by using the following properties of the trigonometric functions:

$$\int_{t_0}^{t_0+T} \cos n\omega t \cos m\omega t \, dt = \begin{cases} 0 & \text{if } m \neq n \\ T/2 & \text{if } m = n \end{cases} \tag{4}$$

$$\int_{t_0}^{t_0+T} \sin n\omega t \sin m\omega t \, dt = \begin{cases} 0 & \text{if } m \neq n \\ T/2 & \text{if } m = n \end{cases} \tag{5}$$

$$\int_{t_0}^{t_0+T} \sin n\omega t \cos m\omega t \, dt = 0 \tag{6}$$

The above equations can easily be derived. For example, for $m = n$,

$$\int_{t_0}^{t_0+T} \cos n\omega t \cos m\omega t \, dt = \int_{t_0}^{t_0+T} \cos^2 n\omega t \, dt$$

$$= \frac{1}{2} \int_{t_0}^{t_0+T} (1 + \cos 2n\omega t) \, dt = \frac{T}{2}$$

Similarly, for $m \neq n$

$$\int_{t_0}^{t_0+T} \cos n\omega t \cos m\omega t \, dt$$

$$= \frac{1}{2} \int_{t_0}^{t_0+T} [\cos (n - m)\omega t + \cos (n + m)\omega t] \, dt$$

$$= \frac{1}{2}\left[\frac{1}{(n-m)\omega}\sin(n-m)\omega t + \frac{1}{(n+m)\omega}\sin(n+m)\omega t\right]_{t_0}^{t_0+T} = 0$$

To determine the coefficients a_n and b_n, we first multiply Eq. (2) by dt and integrate from t_0 to $t_0 + T$:

$$\int_{t_0}^{t_0+T} f(t)dt = \frac{1}{2}a_0\int_{t_0}^{t_0+T} dt + \sum_{n=1}^{\infty} a_n \int_{t_0}^{t_0+T}\cos n\omega t\, dt + \sum_{n=1}^{\infty} b_n \int_{t_0}^{t_0+T}\sin n\omega t\, dt = \frac{T}{2}a_0$$

where we have used Eqs. (4) and (6) for $m = 0$. Thus

$$a_0 = \frac{2}{T}\int_{t_0}^{t_0+T} f(t)\, dt \qquad \textbf{(7)}$$

Next, if we multiply Eq. (2) by $\cos(m\omega t)\, dt$ and integrate from t_0 to $t_0 + T$, we obtain

$$\int_{t_0}^{t_0+T} f(t)\cos m\omega t\, dt = \frac{1}{2}a_0\int_{t_0}^{t_0+T}\cos m\omega t\, dt + \sum_{n=1}^{\infty} a_n\int_{t_0}^{t_0+T}\cos m\omega t\cos n\omega t\, dt + \sum_{n=1}^{\infty} b_n\int_{t_0}^{t_0+T}\cos m\omega t\sin n\omega t\, dt = \frac{T}{2}a_m$$

where we have used Eqs. (4) and (6). We may combine the above equation with Eq. (7) to write

$$a_n = \frac{2}{T}\int_{t_0}^{t_0+T} f(t)\cos n\omega t\, dt \qquad n = 0, 1, 2, 3, \ldots \qquad \textbf{(8)}$$

Similarly,

$$b_n = \frac{2}{T}\int_{t_0}^{t_0+T} f(t)\sin n\omega t\, dt \qquad n = 1, 2, 3, \ldots \qquad \textbf{(9)}$$

Note that the value of t_0 is quite arbitrary. In some problems it is convenient to choose

$$t_0 = -T/2$$

then

$$a_n = \frac{2}{T}\int_{-T/2}^{+T/2} f(t)\cos n\omega t\, dt \qquad n = 0, 1, 2, \ldots$$

and

$$b_n = \frac{2}{T}\int_{-T/2}^{+T/2} f(t)\sin n\omega t\, dt \qquad n = 0, 1, 2, \ldots$$

Such a choice is particularly convenient when the function is even [i.e., $f(t) = f(-t)$] or odd [i.e., $f(t) = -f(-t)$]. In the former case $b_n = 0$ whereas in the latter case $a_n = 0$. In some problems, it is convenient to choose $t_0 = 0$.

Example 8.1 Consider a periodic function of the form

$$f(t) = t \quad \text{for} \quad -\tau < t < +\tau$$
$$f(t + 2n\tau) = f(t) \qquad \textbf{(10)}$$

(see Fig. 8.1). Such a function is referred to as a *sawtooth function*. In this example, we expand the above function in a Fourier series. Now, since $f(t)$ is an odd function of t, $a_n = 0$ and

$$b_n = \frac{2}{T}\int_{-\tau}^{+\tau} f(t)\sin n\omega t\, dt = \frac{1}{\tau}2\int_0^{\tau} t\sin n\omega t\, dt$$

Notice that the periodicity is $2t$ and, therefore, $\omega = \pi/\tau$. Carrying out the integration, we obtain

$$b_n = \frac{2}{\tau}\left[-\frac{t}{n\omega}\cos n\omega t + \frac{1}{n\omega}\left(\frac{1}{n\omega}\sin n\omega t\right)\right]_0^{\tau} = -\frac{2\tau}{n\pi}\cos n\pi = (-1)^{n+1}\frac{2\tau}{n\pi} \qquad \textbf{(11)}$$

Thus

$$f(t) = \frac{2\tau}{\pi}\sum_{n=1,2,\ldots}\frac{(-1)^{n+1}}{n}\sin n\omega t = \frac{2\tau}{\pi}\left(\sin\omega t - \frac{1}{2}\sin 2\omega t + \frac{1}{3}\sin 3\omega t - \cdots\right) \qquad \textbf{(12)}$$

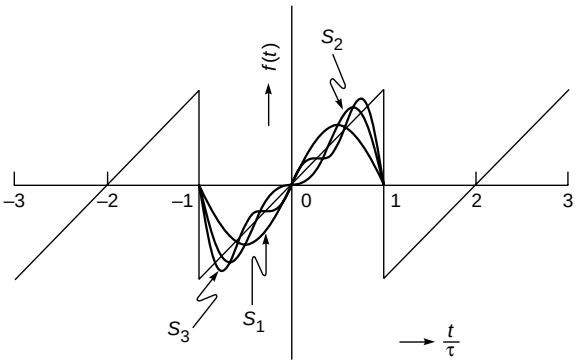


Fig. 8.1 The sawtooth function; S_1, S_2, and S_3 represent the partial sums corresponding to the sawtooth function.

In Fig. 8.1 we have also plotted the partial sums which are given by

$$S_1 = \frac{2\tau}{\pi}\sin \omega t \qquad S_2 = \frac{2\tau}{\pi}\left(\sin \omega t - \frac{1}{2}\sin 2\omega t\right)$$

$$S_3 = \frac{2\tau}{\pi}\left(\sin \omega t - \frac{1}{2}\sin 2\omega t + \frac{1}{3}\sin 3\omega t\right)$$

It can be seen from the figure that as n increases, the sum S_n approaches the function $f(t)$.

Example 8.2 In this example, we will use the Fourier series to expand the function defined by the following equations:

$$f(t) = \begin{cases} -A & \text{for } -\dfrac{T}{2} < t < 0 \\ +A & \text{for } 0 < t < \dfrac{T}{2} \end{cases} \qquad \textbf{(13)}$$

and

$$f(t + T) = f(t)$$

The function is plotted in Fig. 8.2. Once again the function is an odd function; consequently $a_n = 0$ and

$$b_n = \frac{2}{T}2\int_0^{T/2} A\sin n\omega t\, dt = \frac{4A}{T}\frac{1}{n\omega}\left(-\cos n\omega t\right)_0^{T/2}$$

$$= \frac{2A}{n\pi}\left(1 - \cos n\pi\right) = \frac{2A}{n\pi}\left[1 - (-1)^n\right]$$

Thus

$$f(t) = \frac{2A}{\pi}\sum_{n=1,2,3,\ldots}\frac{1}{n}\left[1 - (-1)^n\right]\sin n\omega t$$

$$= \frac{4A}{\pi}\left(\sin \omega t + \frac{1}{3}\sin 3\omega t + \frac{1}{5}\sin 5\omega t + \ldots\right)$$

The partial sums

$$S_1 = \frac{4A}{\pi}\sin \omega t \qquad S_2 = \frac{4A}{\pi}\left(\sin \omega t + \frac{1}{3}\sin 3\omega t\right)$$

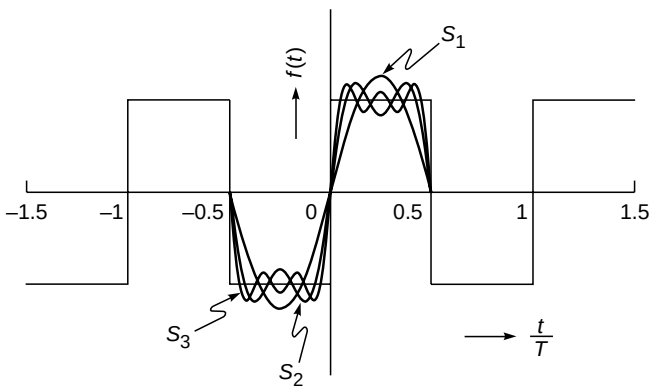


Fig. 8.2 A plot of the periodic step function defined by Eq. (13). S_1, S_2, and S_3 represent the corresponding partial sums.

$$S_3 = \frac{4A}{\pi}\left(\sin \omega t + \frac{1}{3}\sin 3\omega t + \frac{1}{5}\sin 5\omega t\right)$$

are also plotted in Fig. 8.2.

8.2 TRANSVERSE VIBRATIONS OF A PLUCKED STRING

An interesting application of the Fourier series lies in studying the transverse vibrations of a plucked string.

Let us consider a stretched string, fixed at the two ends A and B. One of the ends (A) is chosen as the origin. In the equilibrium position of the string, it is assumed to lie along the x axis (see Fig. 8.3). A point of the string is moved upward by a distance d; the corresponding shape of the string is shown as dashed line in Fig. 8.3. If the displacement occurs at a distance a from the origin, the equation of the string (in its displaced position) is given by

$$y = \begin{cases} \dfrac{d}{a}x & \text{for } 0 < x < a \\ \dfrac{d}{L-a}(L - x) & \text{for } a < x < L \end{cases} \qquad \textbf{(14)}$$

where L represents the length of the string. Now, if the string is released from this position at $t = 0$, we want to determine the shape of the string at any subsequent time.

We will show in Sec. 11.6 that the displacement $y(x, t)$ satisfies the wave equation

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y}{\partial t^2} \qquad \textbf{(15)}$$

where $v\left(=\sqrt{T/\rho}\right)$ represents the speed of the transverse waves, T being the tension in the string and ρ the mass per unit length. We want to solve Eq. (15) subject to the following boundary conditions:

1. $y = 0$ at $x = 0$ and $x = L$ for all values of t **(16)**

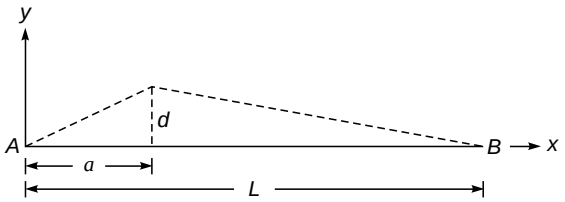


Fig. 8.3 The plucked string; AB represents the equilibrium position. The dashed lines show the displaced position at $t = 0$.

2. At $t = 0$

$$\frac{\partial y}{\partial t} = 0 \qquad \text{for all values of } x \tag{17}$$

$$y(x, t = 0) = \begin{cases} \dfrac{d}{a}x & \text{for } 0 < x < a \\ \dfrac{d}{L-a}(L-x) & \text{for } a < x < L \end{cases} \tag{18}$$

Assuming a time dependence of the form $\cos \omega t$ (or $\sin \omega t$):

$$y(X, t) = X(x) \cos \omega t$$

we obtain

$$\frac{d^2X}{dx^2} = -\frac{\omega^2}{v^2} X(x)$$

or

$$\frac{d^2X}{dx^2} + k^2X(x) = 0 \tag{19}$$

where

$$k = \frac{\omega}{v} \tag{20}$$

The solution of Eq. (19) is simple[1].

$$X(x) = A \sin kx + B \cos kx \tag{21}$$

Thus

$$y(x, t) = (A \sin kx + B \cos kx)(C \cos \omega t + D \sin \omega t)$$

Now

$$y(x, t)\big|_{x=0} = 0 \qquad \text{for all values of } t$$

Thus $B = 0$ and we obtain

$$y(x, t) = \sin kx \, (C \cos \omega t + D \sin \omega t)$$

where we have absorbed A in C and D. Since

$$y(x, t)\big|_{x=L} = 0 \qquad \text{for all values of } t$$

we must have

$$\sin kL = 0$$

or

$$kL = n\pi \qquad n = 1, 2, 3, \ldots \tag{22}$$

Thus, only discrete values of k (and hence of ω) are permissible; these are given by

$$k_n = \frac{n\pi}{L} \qquad n = 1, 2, \ldots \tag{23}$$

giving

$$\omega_n = \frac{n\pi v}{L} \qquad n = 1, 2, \ldots \tag{24}$$

Equation (24) gives the frequencies of the various modes of the string. The mode corresponding to the lowest frequency ($n = 1$) is known as the fundamental mode.

Thus the solution of Eq. (15) satisfying the boundary condition given by Eq. (16) is given by

$$y(x, t) = \sum_{n=1,2,3,\ldots} (\sin k_n x)(C_n \cos \omega_n t + D_n \sin \omega_n t) \tag{25}$$

Differentiating partially with respect to t, we get

$$\left.\frac{\partial y}{\partial t}\right|_{t=0} = \sum_n \sin k_n x \, (-\omega_n C_n \sin \omega_n t + \omega_n D_n \cos \omega_n t)\Big|_{t=0}$$

$$= \sum_n \omega_n D_n \sin k_n x \tag{26}$$

---

[1] Rigorously we should proceed by using the method of separation of variables; thus we assume

$$y(x, t) = X(x)T(t)$$

where $X(x)$ is a function of x alone and $T(t)$ is a function of t alone. Substituting in Eq. (15), we get

$$\frac{1}{X(x)}\frac{d^2X}{dx^2} = \frac{1}{v^2}\frac{1}{T(t)}\frac{d^2T}{dt^2} = -k^2$$

Since the term $\frac{1}{X}\frac{d^2X}{dx^2}$ is a function of x alone and the term $\frac{1}{v^2}\frac{1}{T}\frac{d^2T}{dt^2}$ is a function of t alone, each term must be equal to a constant which we have set equal to $-k^2$. Thus

$$\frac{d^2T}{dt^2} + \omega^2 T(t) = 0$$

and

$$\frac{d^2X}{dx^2} + k^2X(x) = 0$$

where

$$\omega = kv$$

Since

$$\left.\frac{\partial y}{\partial t}\right|_{t=0} = 0 \qquad \text{for all values of } x$$

we must have

$$D_n = 0 \qquad \text{for all } n$$

Thus

$$y(x, t) = \sum_{n=1,2,3,\ldots} C_n \sin k_n x \cos \omega_n t \qquad \textbf{(27)}$$

or

$$y(x, 0) = \sum_n C_n \sin\left(\frac{n\pi}{L} x\right) \qquad \textbf{(28)}$$

The above equation is essentially a Fourier series, and to determine C_n, we multiply both sides of Eq. (28) by $\sin\left(\frac{m\pi}{L} x\right)dx$ and integrate from 0 to L to obtain

$$C_m = \frac{2}{L}\int_0^L y(x, 0) \sin\left(\frac{m\pi}{L} x\right) dx \qquad \textbf{(29)}$$

where we have used the relation

$$\int_0^L \sin\frac{n\pi x}{L}\sin\frac{m\pi x}{L}\, dx = \begin{cases} 0 & \text{if } m \neq n \\ L/2 & \text{if } m = n \end{cases} \qquad \textbf{(30)}$$

[cf. Eq. (5)]. Substituting the expression for $y(x, 0)$ from Eq. (18), we obtain

$$C_n = \frac{2}{L}\left[\frac{d}{a}\int_0^a x \sin\left(\frac{n\pi}{L} x\right) dx + \frac{d}{L-a}\int_a^L (L-x) \sin\left(\frac{n\pi}{L} x\right) dx\right]$$

$$= \frac{2dL^2}{a(L-a)\pi^2 n^2}\sin\left(\frac{n\pi}{L} a\right)$$

On substituting in Eq. (27), we finally obtain

$$y(x,t) = \frac{2dL^2}{a(L-a)\pi^2} \sum_{n=1,2,3,\ldots} \frac{1}{n^2}\sin\left(\frac{n\pi}{L} a\right) \times \sin\left(\frac{n\pi}{L} x\right)\cos\left(\frac{n\pi v}{L} t\right) \qquad \textbf{(31)}$$

Equation (31) can be used to determine the shape of the string at an arbitrary time t. If the string is plucked at the center (i.e., $a = L/2$), terms corresponding to $n = 2, 4, 6, \ldots$ are absent (i.e., the even harmonics are absent) and Eq. (31) simplifies to

$$y(x, t) = \frac{8d}{\pi^2}\sum_m (-1)^{m+1}\frac{1}{(2m-1)^2} \sin\frac{(2m-1)\pi x}{L} \times \cos\frac{(2m-1)(\pi v t)}{L} \qquad \textbf{(32)}$$

8.3 APPLICATION OF FOURIER SERIES IN FORCED VIBRATIONS

Let us consider the forced vibrations of a damped oscillator. The equation of motion is

$$m\frac{d^2y}{dt^2} + \Gamma\frac{dy}{dt} + k_0 y = F(t) \qquad \textbf{(33)}$$

where Γ represents the damping constant (see Sec. 7.3) and F represents the external force. It has been shown in Sec. 7.4 that if $\Gamma > 0$ and

$$F(t) = F_0 \cos(pt + \theta) \qquad \textbf{(34)}$$

then the steady-state solution of Eq. (33) is a simple harmonic motion with the frequency of the external force. If $F(t)$ is not a sine or cosine function, a general solution of Eq. (33) is difficult to obtain; however, if $F(t)$ is periodic, then we can apply Fourier's theorem to obtain a solution of Eq. (33). For example, let

$$F(t) = \alpha t \qquad \text{for } -\tau < t < \tau$$

and

$$F(t + 2n\tau) = F(t) \qquad n = 1, 2, \ldots \qquad \textbf{(35)}$$

The Fourier expansion of such a function was discussed in Example 8.1 and is of the form

$$F(t) = \sum_n F_n \sin n\omega t = \frac{2\alpha\tau}{\pi}\sum_{n=1,2,\ldots}\frac{(-1)^{n+1}}{n}\sin n\omega t \qquad \textbf{(36)}$$

We next consider the solution of the differential equation

$$m\frac{d^2y_n}{dt^2} + \Gamma\frac{dy_n}{dt} + k_0 y_n = F_n \sin n\omega t$$

or

$$\frac{d^2y_n}{dt^2} + K\frac{dy_n}{dt} + \omega_0^2 y_n = A_n \sin n\omega t \qquad \textbf{(37)}$$

where

$$K \equiv \frac{\Gamma}{m} \qquad \omega_0^2 \equiv \frac{k_0}{m}$$

and

$$A_n = \frac{F_n}{m} = \frac{(-1)^{n+1}}{n}\frac{2\alpha\tau}{\pi m} \tag{38}$$

The steady-state solution of Eq. (37) will be of the form

$$y_n = C_n \sin n\omega t + D_n \cos n\omega t$$

and the solution of Eq. (33) will be of the form

$$y = \sum_n y_n \tag{39}$$

To determine C_n and D_n, we substitute the above solution in Eq. (37) to obtain

$$-n^2\omega^2(C_n \sin n\omega t + D_n \cos n\omega t)$$
$$+ n\omega K(C_n \cos n\omega t - D_n \sin n\omega t)$$
$$+ \omega_0^2 (C_n \sin n\omega t + D_n \cos n\omega t) = A_n \sin n\omega t$$

Thus

$$\begin{aligned}(\omega_0^2 - n^2\omega^2)C_n - n\omega K D_n &= A_n \\ (\omega_0^2 - n^2\omega^2)D_n + n\omega K C_n &= 0\end{aligned} \tag{40}$$

Solving the above equations, we get

$$D_n = -\frac{n\omega K}{(\omega_0^2 - n^2\omega^2)^2 + n^2\omega^2K^2} A_n$$

and

$$C_n = \frac{\omega_0^2 - n^2\omega^2}{(\omega_0^2 - n^2\omega^2)^2 + n^2\omega^2K^2} A_n$$

Thus the steady-state solution can be written in the form

$$y = \sum_n G_n \sin(n\omega t + \theta_n) \tag{41}$$

where the amplitude G_n is given by

$$\begin{aligned}G_n &= (C_n^2 + D_n^2)^{1/2} \\ &= \frac{A_n}{[(\omega_0^2 - n^2\omega^2)^2 + n^2\omega^2K^2]^{1/2}}\end{aligned} \tag{42}$$

8.4 THE FOURIER INTEGRAL

In Sec. 8.1 we showed that a periodic function can be expanded in the form

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n \cos n\omega t + b_n \sin n\omega t) \tag{43}$$

where

$$a_n = \frac{2}{T}\int_{t_0}^{t_o+T} f(t)\cos n\omega t\, dt \tag{44}$$

$$b_n = \frac{2}{T}\int_{t_0}^{t_o+T} f(t)\sin n\omega t\, dt \tag{45}$$

and

$$T = \frac{2\pi}{\omega} \tag{46}$$

On substituting the above expressions for a_n and b_n into Eq. (43), we get [we must replace t by t' in Eqs. (44) and (45)].

$$f(t) = \frac{1}{T}\int_{-T/2}^{+T/2} f(t')\, dt'$$
$$+ \sum_{n=1}^{\infty}\left[\frac{2}{T}\cos n\omega t \int_{-T/2}^{+T/2} f(t')\cos n\omega t'\, dt'\right.$$
$$\left. + \frac{2}{T}\sin n\omega t \int_{-T/2}^{+T/2} f(t')\sin n\omega t'\, dt'\right] \tag{47}$$

or

$$f(t) = \frac{1}{2\pi}\Delta s \int_{-\pi/\Delta s}^{+\pi/\Delta s} f(t')\, dt'$$
$$+ \sum_{n=1}^{\infty}\frac{\Delta s}{\pi}\int_{-\pi/\Delta s}^{+\pi/\Delta s} f(t')\cos[n\Delta s\,(t-t')]\, dt' \tag{48}$$

where

$$\Delta s \equiv \frac{2\pi}{T} = \omega$$

We let $T \to \infty$ so that $\Delta s \to 0$; notice that when $T \to \infty$, the function is no longer periodic. Thus if the integral

$$\int_{-\infty}^{+\infty} |f(t')|\, dt'$$

exists (i.e., if it has a finite value) then the first term on the RHS of Eq. (48) will go to zero. Further, since

$$\int_0^{\infty} F(s)\, ds = \lim_{\Delta s \to 0}\sum_{n=1}^{\infty} F(n\Delta s)\Delta s \tag{49}$$

we have

$$f(t) = \frac{1}{\pi}\int_0^{\infty}\left[\int_{-\infty}^{+\infty} f(t')\cos[s(t-t')]\,dt'\right]ds \qquad \textbf{(50)}$$

Equation (50) is known as the *Fourier integral*. Since the cosine function inside the integral is an even function of s, we may write

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} f(t')\cos[s(t-t')]\,dt'\right]ds \qquad \textbf{(51)}$$

Further, since $\sin[s(t-t')]$ is an odd function of s,

$$\frac{i}{2\pi}\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} f(t')\sin[s(t-t')]\,dt'\right]ds = 0 \qquad \textbf{(52)}$$

If we add (or subtract) the above two equations, we get

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(t')e^{\pm i\omega(t-t')}\,dt'\,d\omega \qquad \textbf{(53)}$$

where we have replaced s by ω. Equation (53) is usually referred to as the *Fourier integral theorem*. Thus, if

$$F(\omega) \equiv \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(t)e^{\pm i\omega t}\,dt \qquad \textbf{(54)}$$

then

$$f(t) \equiv \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} F(\omega)e^{\mp i\omega t}\,d\omega \qquad \textbf{(55)}$$

The function $F(\omega)$ is known as the *Fourier transform* of $f(t)$. For a time-dependent function $f(t)$, $F(\omega)$ is usually referred to as its frequency spectrum. Equations (54) and (55) are also written in the form

$$F(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} f(t)e^{-i\omega t}\,dt \qquad \textbf{(56)}$$

with

$$f(t) = \int_{-\infty}^{+\infty} F(\omega)e^{i\omega t}\,d\omega \qquad \textbf{(57)}$$

We can also write

$$G(k) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} f(x)e^{-ikx}\,dx \qquad \textbf{(58)}$$

with

$$f(x) = \int_{-\infty}^{+\infty} G(k)e^{+ikx}\,dk \qquad \textbf{(59)}$$

where k is often referred to as spatial frequency—a concept that is extensively used in Fourier optics (see Chap. 19).

In Chap. 9 we will introduce the Dirac delta function, rederive Eqs. (54) to (59), and work out a few examples to illustrate the physics and applications of the Fourier transform. In Chap. 10, we will use Fourier transforms to study the propagation of optical pulse in dispersive and nonlinear media.

Summary

- A periodic function with period T, i.e.,

  $$f(t + nT) = f(t) \qquad n = 0, \pm1, \pm2, \ldots$$

  can be expanded in the form

  $$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n\cos\left(\frac{2\pi n}{T}t\right) + \sum_{n=1}^{\infty} b_n\sin\left(\frac{2\pi n}{T}t\right)$$

  $$= \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n\cos n\omega t + \sum_{n=1}^{\infty} b_n\sin n\omega t$$

  where

  $$\omega = \frac{2\pi}{T}$$

  represents the fundamental frequency. The above infinite series is known as the Fourier series, and the coefficients a_n and b_n are given by

  $$a_n = \frac{2}{T}\int_{t_0}^{t_0+T} f(t)\cos n\omega t \qquad n = 0, 1, 2, 3, \ldots$$

  and

  $$b_n = \frac{2}{T}\int_{t_0}^{t_0+T} f(t)\sin n\omega t \qquad n = 0, 1, 2, 3, \ldots$$

- Transverse vibrations of a plucked string and forced vibrations can be studied by using Fourier series.
- For a time-dependent function $f(t)$, its Fourier transform is defined by the equation

  $$F(\omega) \equiv \frac{1}{2\pi}\int_{-\infty}^{+\infty} f(t)e^{\pm i\omega t}\,dt$$

  Then

  $$F(t) \equiv \int_{-\infty}^{+\infty} f(\omega)e^{\mp i\omega t}\,d\omega$$

Problems

8.1 Consider a periodic force of the form

$$F(t) = \begin{cases} F_0\sin\omega t & \text{for } 0 < t < T/2 \\ 0 & \text{for } T/2 < t < T \end{cases}$$

and

$$F(t + T) = F(t)$$

where

$$\omega = \frac{2\pi}{T}$$

Show that

$$F(t) = \frac{1}{\pi}F_0 + \frac{1}{2}F_0 \sin\omega t - \frac{2}{\pi}F_0\left(\frac{1}{3}\cos 2\omega t + \frac{1}{15}\cos 4\omega t + \cdots\right)$$

One obtains a periodic voltage of the above form in a half wave rectifier. What will be the Fourier expansion corresponding to full wave rectification?

8.2 In quantum mechanics, the solution of the one-dimensional Schrödinger equation for a free particle is given by

$$\Psi(x, t) = \frac{1}{\sqrt{2\pi\hbar}}\int_{-\infty}^{+\infty} a(p)e^{\frac{i}{\hbar}\left(px - \frac{p^2}{2m}t\right)}dp$$

where p is the momentum of the particle of mass m. Show that

$$a(p) = \frac{1}{\sqrt{2\pi\hbar}}\int_{-\infty}^{+\infty} \Psi(x, 0)e^{-\frac{i}{\hbar}px}dx$$

8.3 In continuation of Prob. 8.2, if we assume

$$\Psi(x, 0) = \frac{1}{(\pi\sigma^2)^{1/4}}\exp\left(-\frac{x^2}{2\sigma^2}\right)\exp\left(\frac{i}{\hbar}p_0 x\right)$$

then show that

$$a(p) = \left(\frac{\sigma^2}{\pi\hbar^2}\right)^{1/4}\exp\left[-\frac{\sigma^2}{2\hbar^2}(p - p_0)^2\right]$$

Also show that

$$\int_{-\infty}^{+\infty} |\Psi(x, 0)|^2 dx = 1 = \int_{-\infty}^{+\infty} |a(p)|^2 dp$$

Indeed $|\Psi(x, 0)|^2\,dx$ represents the probability of finding the particle between x and $x + dx$, and $|a(p)|^2\,dp$ represents the probability of finding the momentum between p and $p + dp$, and we would have the uncertainty relation

$$\Delta x\,\Delta p \sim \hbar$$

REFERENCES AND SUGGESTED READINGS

1. H. S. Carslaw, *Introduction to the Theory of Fourier Series and Integrals*, Dover Publications, New York, 1950.
2. E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Oxford University Press, New York, 1937.
3. A. K. Ghatak, I. C. Goyal, and S. J. Chua, *Mathematical Physics*, Macmillan India Ltd., New Delhi, 1995.
4. J. Arsac, *Fourier Transforms and the Theory*, Prentice-Hall, Englewood Cliffs, N. J., 1966.
5. E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Clarendon Press, Oxford, 1959.

Chapter Nine

THE DIRAC DELTA FUNCTION AND FOURIER TRANSFORMS

Strictly of course, $\delta(x)$ is not a proper function of x, but can be regarded only as a limit of a certain sequence of functions. All the same one can use $\delta(x)$ as though it were a proper function for practically all the purposes of quantum mechanics without getting incorrect results. One can also use the differential coefficients of $\delta(x)$, namely, $\delta'(x)$, $\delta''(x)$, ... which are even more discontinuous and less 'proper' than $\delta(x)$ itself.

—P. A. M. Dirac, "The Physical Interpretation of Quantum Dynamics," *Proceedings of the Royal Society of London*, A113, pp. 621–641, 1926.

9.1 INTRODUCTION

The Dirac delta function is defined through the equations

$$\delta(x-a) = 0 \qquad x \neq a \tag{1}$$

$$\int_{a-\alpha}^{a+\beta} \delta(x-a)\,dx = 1 \tag{2}$$

where $\alpha, \beta > 0$. Thus the delta function has an infinite value at $x = a$ such that the area under the curve is unity. For an arbitrary function that is continuous at $x = a$, we have

$$\int_{a-\alpha}^{a+\beta} f(x)\delta(x-a)\,dx = f(a)\int_{a-\alpha}^{a+\beta} \delta(x-a)\,dx \qquad \text{[using Eq. (1)]}$$

$$= f(a) \tag{3}$$

It is readily seen that if x has the dimension of length, $\delta(x-a)$ will have the dimension of inverse length. Similarly, if x has the dimension of time, then $\delta(x-a)$ will have the dimension of $(\text{time})^{-1}$.

9.2 REPRESENTATIONS OF THE DIRAC DELTA FUNCTION

There are many representations of the Dirac delta function. Perhaps the simplest representation is the limiting form of the rectangle function $R_\sigma(x)$ defined through the following equation:

$$R_\sigma(x) = \begin{cases} \dfrac{1}{2\sigma} & \text{for } a-\sigma < x < a+\sigma \\ 0 & \text{for } |x-a| > \sigma \end{cases} \tag{4}$$

The function $R_\sigma(x)$ is plotted in Fig. 9.1 for various values of σ. Now,

$$\int_{-\infty}^{+\infty} R_\sigma(x)\,dx = \frac{1}{2\sigma}\int_{a-\sigma}^{a+\sigma} dx = 1 \qquad \text{(irrespective of value of } \sigma)$$

For $\sigma \to 0$, the function $R_\sigma(x)$ becomes more and more sharply peaked, but the area under the curve remains unity. In the limit of $\sigma \to 0$, the function $R_\sigma(x)$ has all the properties of the delta function, and we may write

$$\delta(x-a) = \lim_{\sigma \to 0} R_\sigma(x)$$

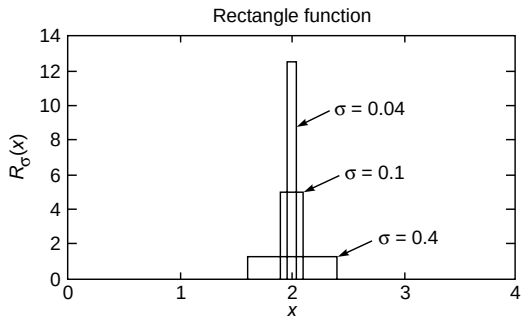


Fig. 9.1 Plots of $R_\sigma(x)$ for $a = 2$ and $\sigma = 0.4$, 0.1, and 0.04. In each case the area under the curve is unity. For $\sigma \to 0$, the function $R_\sigma(x)$ has all the properties of the Dirac delta function.

Now

$$\int_{-\infty}^{+\infty} f(x)\, R_\sigma(x)\, dx = \frac{1}{2\sigma} \int_{a-\sigma}^{a+\sigma} f(x)\, dx \qquad \textbf{(5)}$$

We assume the function $f(x)$ to be continuous at $x = a$. Thus when $\sigma \to 0$, in the infinitesimal interval $a - \sigma < x < a + \sigma$, $f(x)$ may be assumed to be a constant [$= f(a)$] and taken out of the integral. Thus

$$\begin{aligned}\int_{-\infty}^{+\infty} f(x)\delta(x-a)\,dx &= \lim_{\sigma\to 0} \int_{-\infty}^{+\infty} f(x) R_\sigma(x)\,dx \\ &= \lim_{\sigma\to 0} \frac{1}{2\sigma} f(a) \int_{a-\sigma}^{a+\sigma} dx \\ &= f(a)\end{aligned}$$

9.3 INTEGRAL REPRESENTATION OF THE DELTA FUNCTION

An extremely important representation of the Dirac delta function is through the following integral:

$$\delta(x-a) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{\pm ik(x-a)} dk \qquad \textbf{(6)}$$

To prove Eq. (6), we first note that

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{\pm ik(x-a)} dk = \frac{\sin g(x-a)}{\pi(x-a)} \qquad \textbf{(7)}$$

In App. *B*, we have shown that

$$\int_{-\infty}^{+\infty} \frac{\sin gx}{\pi x}\, dx = 1 \qquad g > 0 \qquad \textbf{(8)}$$

irrespective of the value of g, which is assumed to be greater than zero. Further,

$$\lim_{x\to 0} \frac{\sin gx}{x} = g$$

Thus for a large value of g, the function

$$\frac{\sin g(x-a)}{\pi(x-a)}$$

is very sharply peaked around $x = a$ (see Fig. 9.2) and has a unit area under the curve irrespective of the value of g; thus

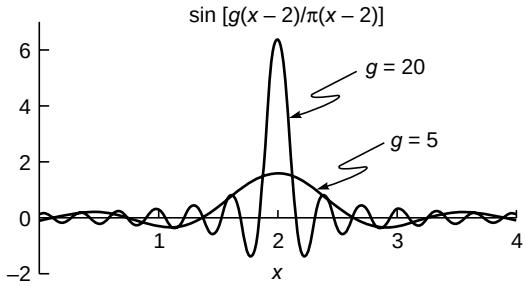


Fig. 9.2 Plots of the function $\dfrac{\sin g(x-a)}{\pi(x-a)}$ for $a = 2$ and $g = 5, 20$. In each case the area under the curve is unity. For $g \to \infty$, the function is very sharply peaked at $x = a$ and has all the properties of Dirac delta function.

in the limit of $g \to \infty$, it has all the properties of the delta function, and we may write

$$\delta(x-a) = \lim_{g\to\infty} \frac{\sin g(x-a)}{\pi(x-a)} = \lim_{g\to\infty} \frac{1}{2\pi} \int_{-g}^{+g} e^{\pm ik(x-a)}\, dk \qquad \textbf{(9)}$$

from which Eq. (6) readily follows.

9.4 DELTA FUNCTION AS A DISTRIBUTION

The delta function is actually a distribution. To understand this, let us consider the Maxwellian distribution

$$N(E)\, dE = N_0 \frac{2}{\sqrt{\pi}} \frac{1}{(kT)^{3/2}} E^{1/2}\, e^{-E/kT}\, dE \qquad \textbf{(10)}$$

where k represents Boltzmann's constant, T is the absolute temperature, and m is the mass of each molecule. In Eq. (10), $N(E)\, dE$ represents the number of molecules whose energies lie between E and $E + dE$. The total number of molecules is given by N_0

$$\begin{aligned}\int_0^\infty N(E)\,dE &= N_0 \frac{2}{\sqrt{\pi}} \frac{1}{(kT)^{3/2}} \int_0^\infty E^{1/2} e^{-E/kT}\, dE \\ &= N_0 \frac{2}{\sqrt{\pi}} \int_0^\infty x^{1/2} e^{-x}\, dx\end{aligned}$$

where $x = E/kT$. The integral is $\Gamma\left(3/2\right) = \dfrac{1}{2}\sqrt{\pi}$ (see Appendix A). Thus

$$\int_0^\infty N(E)\, dE = N_0$$

Whereas N_0 is just a number, the quantity $N(E)$ has dimensions of (energy)$^{-1}$. Obviously, if we ask ourselves how many molecules have the precise speed E_1, the answer is zero. This is a characteristic of a distribution. On the other hand, in addition to the distribution given by Eq. (10), if we do have N_1 molecules, all of them having the same energy E_1, the corresponding distribution function is given by

$$N(E) = N_0 \frac{2}{\sqrt{\pi}} \frac{1}{(kT)^{3/2}} E^{1/2} e^{-E/kT} + N_1 \delta(E - \mathrm{E}_1) \qquad \textbf{(11)}$$

where $\delta(E - E_1)$ represents the Dirac delta function and has the dimensions of inverse energy.

9.5 FOURIER INTEGRAL THEOREM

In Sec. 9.4, we showed the following integral representation of the Dirac delta function:

$$\delta(x - x') = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{\pm ik(x-x')}\, dk \qquad \textbf{(12)}$$

Since

$$f(x) = \int_{-\infty}^{+\infty} \delta(x - x')\, f(x')\, dx' \qquad \textbf{(13)}$$

we may write

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{\pm ik(x-x')} f(x')\, dx'\, dk \qquad \textbf{(14)}$$

Thus if we define

$$F(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) e^{-ikx}\, dx \qquad \textbf{(15)}$$

then

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(k) e^{+ikx}\, dk \qquad \textbf{(16)}$$

The function $F(k)$ is known as the Fourier transform of the function $f(x)$, and Eq. (16) enables us to calculate the original function from the Fourier transform. Equation (14) constitutes what is known as the *Fourier integral theorem* that is valid when the following conditions are satisfied (see, e.g., Refs. 4 and 5 of Chap. 8):

1. The function $f(x)$ must be a single-valued function of the real variable x throughout the range $-\infty < x < \infty$. It may, however, have a finite number of finite discontinuities.
2. The integral $\int_{-\infty}^{+\infty} |f(x)|\, dx$ must exist.

From Eq.(14) it is obvious that in Eqs. (15) and (16) there is no reason why the factors e^{ikx} and e^{-ikx} cannot be interchanged; i.e., we could have defined

$$F(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) e^{+ikx}\, dx \qquad \textbf{(17)}$$

then

$$f(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(x) e^{-ikx}\, dx \qquad \textbf{(18)}$$

However, in all that follows we will use the definitions given by Eqs. (15) and (16).

Example 9.1 As an example we consider a Gaussian function given by

$$f(x) = A \exp\left(-\frac{x^2}{2\sigma^2}\right) \qquad \textbf{(19)}$$

Its Fourier transform is given by

$$F(k) = \frac{A}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2\sigma^2} e^{-ikx}\, dx$$

or

$$F(k) = A\sigma \exp\left(-\frac{1}{2} k^2 \sigma^2\right) \qquad \textbf{(20)}$$

where we have made use of the following integral (see App. A):

$$\int_{-\infty}^{+\infty} e^{-ax^2 + \beta x}\, dx = \sqrt{\frac{\pi}{\alpha}} \exp\left[\frac{\beta^2}{4\alpha}\right];\ \mathrm{Re}\ \alpha > 0 \qquad \textbf{(21)}$$

As can be seen from Eq. (20), the function $F(k)$ is also Gaussian; thus the Fourier transform of a Gaussian is a Gaussian. Note that the Gaussian function given by Eq. (19) has a spatial width given by [see Fig. 9.3(a)]

$$\Delta x \sim \sigma$$

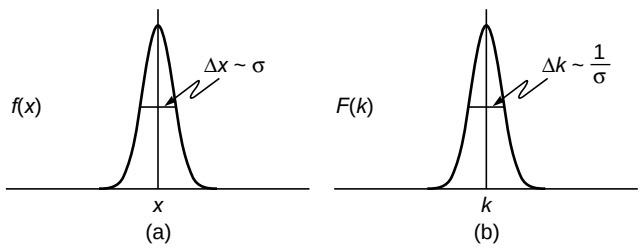


Fig. 9.3 (a) The Gaussian function $f(x)$ as given by Eq. (19). (b) The Fourier transform of the Gaussian function is also a Gaussian in the k-space [see Eq. (20)].

Its Fourier transform $F(k)$ has a width given by [see Fig. 9.3(b)]

$$\Delta k \sim \frac{1}{\sigma} \tag{22}$$

Thus

$$\Delta x\ \Delta k \sim 1 \tag{23}$$

which is a general characteristic of the Fourier transform pair.

Example 9.2 As another example, we calculate the Fourier transform of the rectangle function

$$f(x) = \text{rect}\left(\frac{x}{a}\right) = \begin{cases} 1 & |x| < \frac{1}{2}a \\ 0 & |x| > \frac{1}{2}a \end{cases} \tag{24}$$

Its Fourier transform will be given by (see Fig. 9.4)

$$F(k) = \frac{1}{\sqrt{2\pi}} \int_{-a/2}^{+a/2} e^{-ikx}\, dx$$

$$= \sqrt{\frac{2}{\pi}} \frac{\sin(ka/2)}{k} \tag{25}$$

Once again, the rectangle function has a width $\Delta x = a$, and its Fourier transform has a width

$$\Delta k \sim \frac{\pi}{a}$$

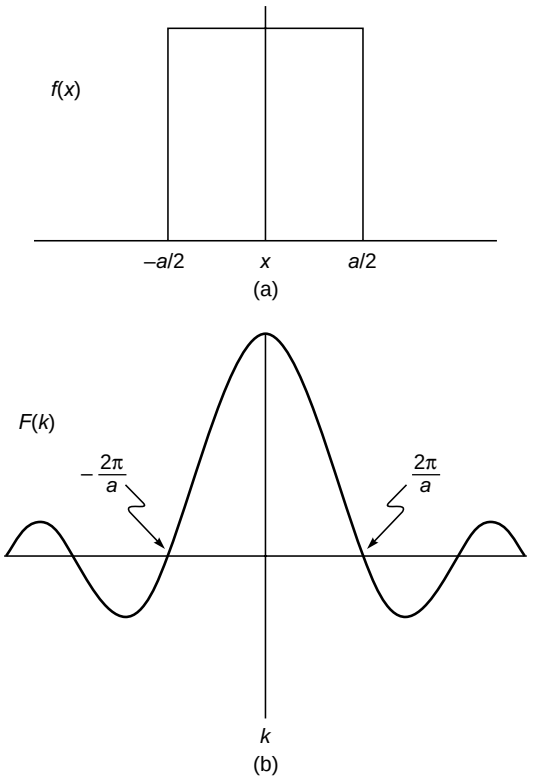


Fig. 9.4 (a) The rectangle function. (b) The Fourier transform of the rectangle function.

giving $\Delta x\ \Delta k \sim 1$. Equation (25) can be written in the form

$$F(k) = \frac{a}{\sqrt{2\pi}} \text{sinc}\ \xi \tag{26}$$

where

$$\xi \equiv \frac{ka}{2} \tag{27}$$

and

$$\text{sinc}\, x \equiv \frac{\sin x}{x} \tag{28}$$

is known as the sinc function. Using Eq. (16), we can write

$$\text{rect}\left(\frac{x}{a}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{a}{\sqrt{2\pi}} \sin\, \xi\, e^{ikx}\, dk$$

or

$$\text{rect}\left(\frac{X}{2}\right) = \sqrt{\frac{2}{\pi}} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \text{sinc}\, \xi\, e^{i\xi X}\, d\xi \right) \tag{29}$$

where

$$X \equiv \frac{x}{a/2} \tag{30}$$

Thus, the Fourier transform of the sinc function is $\sqrt{\pi/2}$ times the rectangle function:

$$\mathbf{F}\left(\frac{\sin x}{x}\right) = \sqrt{\frac{\pi}{2}}\, \text{rect}\left(\frac{k}{2}\right) \tag{31}$$

For a time-dependent function we can write the Fourier transform in the following form (see also Sec. 8.4):

$$\mathbf{F}[f(t)] = F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{\pm i\omega t}\, dt \tag{32}$$

The inverse Fourier transform will then be given by

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\omega) e^{\mp i\omega t}\, d\omega \tag{33}$$

The above equations are nothing but Eqs. (15) and (16) with x and k replaced by t and ω, respectively. The function $F(\omega)$ is usually referred to as the frequency spectrum of the time-dependent function $f(t)$.

Example 9.3 As an example, we consider the Fourier transform of the Gaussian function (see Fig. 9.5)

$$f(t) = A \exp\left(-\frac{t^2}{t_0^2}\right) \tag{34}$$

Thus, the Fourier transform is given by [using Eq. (32)]

$$F(\omega) = \frac{A}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{t^2}{t_0^2}\right) e^{-i\omega t} dt$$

$$= \frac{At_0}{\sqrt{2}} \exp\left(-\frac{\omega^2 t_0^2}{4}\right) \tag{35}$$

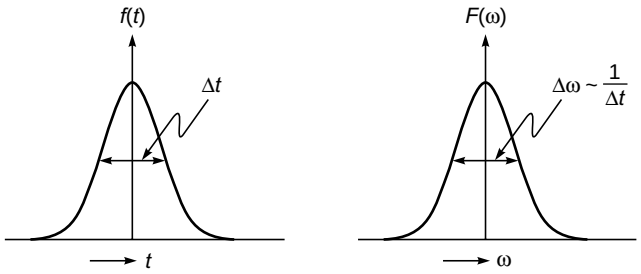


Fig 9.5 The Fourier transform of a Gaussian temporal function is a Gaussian function in the frequency space.

where we have used the integral given by Eq. (21). The function $F(\omega)$ [as given by Eq. (35)] is also plotted in Fig. 9.5. We denote the full width at half maximum (usually abbreviated as FWHM) of $f(t)$ by Δt; thus at $t = \pm\frac{1}{2}\Delta t$, the function $f(t)$ attains one-half of its maximum value:

$$\frac{1}{2}A = A\exp\left[-\frac{(\Delta t)^2}{4t_0^2}\right]$$

Thus

$$\Delta t = 2\sqrt{\ln 2}\, t_0 \approx 1.67 t_0$$

Similarly, if $\Delta\omega$ denotes the FWHM of $F(\omega)$, then (see Fig. 9.5)

$$\Delta\omega = \frac{4\sqrt{\ln 2}}{t_0} \approx \frac{3.34}{t_0} \qquad \textbf{(36)}$$

Thus if a time-dependent function $f(t)$ has a temporal width Δt, then its Fourier transform $F(\omega)$ will have a spectral width

$$\Delta\omega \sim \frac{1}{\Delta t} \qquad \textbf{(37)}$$

giving the uncertainty relation (see also Example 10.4 and Sec. 17.6)

$$\Delta\omega\, \Delta t \sim 1 \qquad \textbf{(38)}$$

The above equation may be compared with the relation $\Delta x\, \Delta k \sim 1$ derived above.

9.6 THE TWO- AND THREE-DIMENSIONAL FOURIER TRANSFORM

One can generalize the analysis of Sec. 9.4 to two or three dimensions. For example, the two-dimensional Fourier transform of a function $f(x,y)$ is defined through the equation

$$F(u,v) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)e^{\pm i(ux+vy)}\,dx\,dy \qquad \textbf{(39)}$$

where u and v are referred to as *spatial frequencies*. The inverse transform is given by

$$f(x,y) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} F(u,v)e^{\mp i(ux+vy)}\,du\,dv \qquad \textbf{(40)}$$

We will use Eqs. (29) and (30) in Sec. 19.6. Similarly, we can define the three-dimensional Fourier transform

$$F(u,v,w) = \frac{1}{(2\pi)^{3/2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y,z)e^{\pm i(ux+vy+wz)}dx\,dy\,dz$$

with its inverse Fourier transform given by **(41)**

$$f(x,y,z) = \frac{1}{(2\pi)^{3/2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} F(u,v,w)e^{\pm i(ux+vy+wz)}du\,dv\,dw$$

(42)

9.6.1 The Convolution Theorem

The *convolution* of two functions $f(x)$ and $g(x)$ is defined by the relation

$$f(x) * g(x) = \int_{-\infty}^{+\infty} f(x')g(x-x')\,dx' = g(x) * f(x) \qquad \textbf{(43)}$$

The convolution has this important property: The Fourier transform of the convolution of two functions is $\sqrt{2\pi}$ times the product of their Fourier transforms. The proof is as follows:

$$\mathbf{F}(f(x) * g(x)) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} dx e^{-ikx}\left[\int_{-\infty}^{+\infty} dx'\, f(x')g(x-x')\right]$$

$$= \sqrt{2\pi}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} dx'\, f(x')e^{-ikx'}\right]$$

$$\times\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} dx\, g(x-x')e^{-ik(x-x')}\right]$$

In the second equation, we substitute $x - x'$ by ξ to obtain

$$\mathbf{F}(f(x) * g(x)) = \sqrt{2\pi}\, F(k)G(k)$$

where $F(k)$ and $G(k)$ are Fourier transforms of $f(x)$ and $g(x)$, respectively. The convolution can be used to obtain the Fourier transforms of the product of two functions:

$$\mathbf{F}(f(x)g(x)) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(x)g(x)e^{-ikx}dx$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} dx\, g(x)e^{-ikx}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} F(k')e^{ik'x}dk'\right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dk' F(k') \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dx\, g(x) e^{-i(k-k')x} \right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(k') G(k-k')\, dk'$$

Thus,

$$\mathbf{F}(f(x)g(x)) = \frac{1}{\sqrt{2\pi}} F(k) * G(k)$$

The above result tells us that the Fourier transform of the product of two functions is $\frac{1}{\sqrt{2\pi}}$ times the convolution of their Fourier transforms.

Summary

- The Dirac delta function is defined through the equations

  $$\delta(x-a) = 0 \qquad x \neq a$$

  and for a well-behaved function $f(x)$, which is continuous at $x = a$,

  $$\int_{-\infty}^{+\infty} f(x)\delta(x-a)\, dx = f(a)$$

- For a time-dependent function $f(t)$, its Fourier transform is defined by the equation

  $$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{\pm i\omega t} dt$$

  Then

  $$F(t) = \int_{-\infty}^{+\infty} f(\omega)\, e^{\mp i\omega t}\, dt$$

- The Fourier transform of the Gaussian function

  $$f(t) = A \exp\left(-\frac{t^2}{t_0^2}\right)$$

  is given by

  $$F(\omega) = \frac{At_0}{2\sqrt{\pi}} e^{-\omega^2 t_0^2/4}$$

- In general, if a function has a temporal spread of $\Delta\mathbf{t}$, then its Fourier transform $\mathbf{F}(\omega)$ will have a spectral spread $\Delta\omega \approx 1/\Delta\mathbf{t}$.
- The two-dimensional Fourier transform of a function $f(x,y)$ is defined through the equation

  $$F(u, v) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) e^{\pm i(ux+vy)} dx\, dy$$

  where u and ν are referred to as *spatial frequencies*. The inverse transform would be given by

  $$f(x,y) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F(u,v) e^{\pm i(ux+vy)} du\, dv$$

- The convolution of two functions $f(x)$ and $g(x)$ is defined by the relation

  $$f(x) * g(x) = \int_{-\infty}^{+\infty} f(x') g(x-x')\, dx' = g(x) * f(x)$$

  The Fourier transform of the convolution of two functions is $\sqrt{2\pi}$ times the product of their Fourier transforms.

Problems

9.1 Consider the Gaussian function

$$G_\sigma(x) \equiv \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right] \qquad \sigma > 0$$

Using Eq. (21), show that $\int_{-\infty}^{+\infty} G_\sigma(x)\, dx = 1$. Plot $G_\sigma(x)$ for $a = 2$ and $\sigma = 1.0$, 5.0, and 10.0. Hence show that

$$\delta(x-a) = \lim_{\sigma\to 0} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right] \qquad \textbf{(44)}$$

which is the Gaussian representation of the delta function.

9.2 Consider the ramp function defined by the following equation:

$$F_\sigma(x) = \begin{cases} 0 & \text{for } x < a-\sigma \\ \dfrac{1}{2\sigma}(x-a+\sigma) & \text{for } |x-a| < \sigma \\ 1 & \text{for } x > a+\sigma \end{cases} \qquad \textbf{(45)}$$

Show that $dF_\sigma/dx = R_\sigma(x)$, where $R_\sigma(x)$ is the rectangle function defined by Eq. (4). Taking the limit $\sigma \to 0$, show that $\delta(x-a) = \dfrac{d}{dx} H(x-a)$ where $H(x-a)$ is the unit step function. Thus we get the following important result:

If a function has a discontinuity of α at $x = a$, then its derivative (at $x = a$) is $\alpha\delta(x-a)$.

9.3 Consider the symmetric function

$$\psi(x) = A \exp(-K|x|)$$

Show that

$$\psi''(x) = K^2\psi(x) - 2AK\delta(x)$$

9.4 Consider the function $f(t) = Ae^{-t^2/2\tau^2}e^{i\omega_0 t}$. Calculate its Fourier spectrum $F(\omega) = \dfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(t)e^{-i\omega t}dt$

and evaluate approximately $\Delta\omega\Delta t$. Evaluate $f(t)$, using the expression for $F(\omega)$.

9.5 Calculate the Fourier transform of the following functions

(a) $f(x) = \begin{cases} Ae^{ik_0 x} & |x| < L/2 \\ 0 & |x| < L/2 \end{cases}$

(b) $f(x) = A\exp\left(-\dfrac{|x|}{L}\right)$

In each case make an estimate of Δx and Δk and interpret physically.

9.6 Show that the convolution of two Gaussian functions is another Gaussian function:

$$\exp\left(-\frac{x^2}{a^2}\right) * \exp\left(-\frac{x^2}{b^2}\right) = ab\left(\frac{\pi}{a^2+b^2}\right)^{1/2}\exp\left(-\frac{x^2}{a^2+b^2}\right)$$

Chapter Ten

GROUP VELOCITY AND PULSE DISPERSION

In a perfect wave, you cannot say when it *starts*, so you cannot use it for a timing signal. In order to send a *signal* you have to change the wave somehow, make a notch in it, make it a little bit fatter or thinner. That means that you have to have more than one frequency in the wave, and it can be shown that the speed at which *signals* travel is not dependent upon the index alone, but upon the way that the index changes with the frequency.

—Richard Feynman, *Feynman Lectures on Physics*, Vol. I

Important Milestone

1672 *Isaac Newton reported to the Royal Society his observations on the dispersion of sunlight as it passed through a prism. From this experiment, Newton concluded that sunlight is composed of light of different colors which are refracted by glass to different extents.*

10.1 INTRODUCTION

When we switch a light source on and off, we produce a pulse. This pulse propagates through a medium with what is known as the group velocity, which will be discussed in this chapter. In addition, as the pulse propagates, it undergoes distortion which will also be discussed.[1] A study of this distortion of optical pulses is a subject of great importance in many areas; in particular, it has very important significance in fiber-optic communication systems, which will be briefly discussed in Chap. 27 and 29.

10.2 GROUP VELOCITY

Let us consider two plane waves (having the same amplitude A) with slightly different frequencies $\omega + \Delta\omega$ and $\omega - \Delta\omega$ propagating along the $+z$ direction:

$$\Psi_1(z, t) = A\cos\left[(\omega + \Delta\omega)t - (k + \Delta k)z\right] \quad \textbf{(1)}$$

$$\Psi_2(z, t) = A\cos\left[(\omega - \Delta\omega)t - (k - \Delta k)z\right] \quad \textbf{(2)}$$

where $k + \Delta k$ and $k - \Delta k$ are the wave numbers corresponding to the frequencies $\omega + \Delta\omega$ and $\omega - \Delta\omega$, respectively. The superposition of the two waves is given by

$$\Psi(z, t) = A\cos[(\omega + \Delta\omega)t - (k + \Delta k)z] + A\cos[(\omega - \Delta\omega)t - (k - \Delta k)z]$$

or

$$\Psi(z, t) = 2A\cos(\omega t - kz)\cos[(\Delta\omega)t - (\Delta k)z] \quad \textbf{(3)}$$

In Fig. 10.1(a) we have shown the variation of the rapidly varying $\cos(\omega t - kz)$ term at $t = 0$; the distance between two consecutive peaks is $2\pi/k$. In Fig. 10.1(b) we have shown the variation of the slowly varying envelope term, represented by $\cos[(\Delta\omega)t - (\Delta k)z]$ at $t = 0$; the distance between two consecutive peaks is $2\pi/\Delta k$. In Fig. 10.2(a) and (b) we plotted $\Psi(z, t)$ at

$$t = 0 \qquad \text{and} \qquad t = \Delta t$$

Obviously the rapidly varying first term moves with velocity

$$v_p = \frac{\omega}{k} \quad \textbf{(4)}$$

[1] This chapter assumes a knowledge of waves, which will be discussed in Chap. 11. The reader might prefer to go through Chap. 11 first, before going through this chapter.

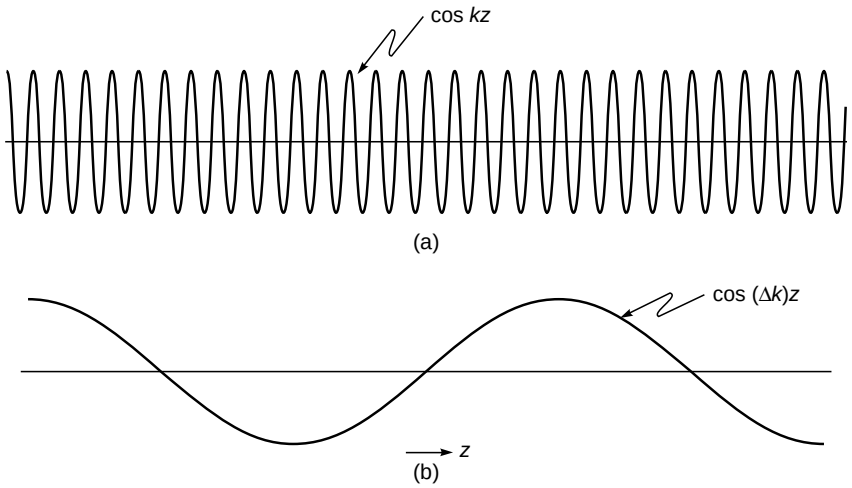


Fig. 10.1 (a) Variation of the rapidly varying $\cos(\omega t - kz)$ term at $t = 0$; the distance between two consecutive peaks is $2\pi/k$. (b) Variation of the slowly varying envelope term, represented by $\cos[(\Delta\omega)t - (\Delta k)z]$, at $t = 0$. The distance between two consecutive peaks is $2\pi/\Delta k$.

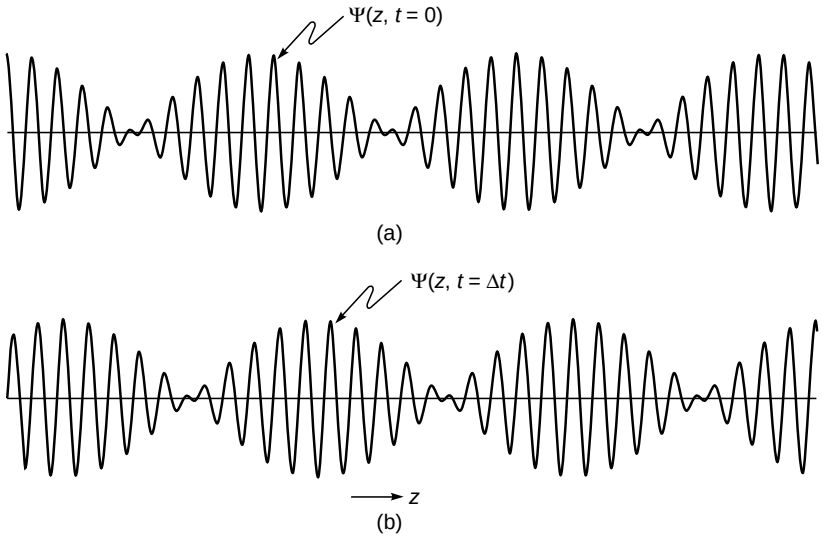


Fig. 10.2 (a) and (b) Variation of $\Psi(z, t)$ at $t = 0$ and at $t = \Delta t$; the envelope moves with the group velocity $\Delta\omega/\Delta k$.

and the slowly varying envelope [which is represented by the second term in Eq. (3)] moves with velocity

$$v_g = \frac{\Delta\omega}{\Delta k} \qquad \textbf{(5)}$$

The quantities v_p and v_g are known as the *phase velocity* and the *group velocity*, respectively. The group velocity is a concept of great importance; indeed in the next section we will rigorously show that a temporal pulse travels with the group velocity given by

$$v_g = \frac{1}{dk/d\omega} \qquad \textbf{(6)}$$

Now, in a medium characterized by the refractive index variation $n(\omega)$,

$$k(\omega) = \frac{\omega}{c} n(\omega) \qquad \textbf{(7)}$$

Thus

$$\frac{1}{v_g} = \frac{dk}{d\omega} = \frac{1}{c}\left(n(\omega) + \omega\frac{dn}{d\omega}\right) \qquad \textbf{(8)}$$

In free space $n(\omega) = 1$ at all frequencies; hence

$$v_g = v_p = c \qquad \textbf{(9)}$$

Returning to Eq. (8), we note that it is customary to express in terms of the free space wavelength λ_0 which is related to ω through

$$\omega = \frac{2\pi c}{\lambda_0} \tag{10}$$

Thus

$$\frac{dn}{d\omega} = \frac{dn}{d\lambda_0}\frac{d\lambda_0}{d\omega} = -\frac{\lambda_0^2}{2\pi c}\frac{dn}{d\lambda_0} \tag{11}$$

or

$$\frac{1}{v_g} = \frac{1}{c}\left[n(\lambda_0) - \lambda_0 \frac{dn}{d\lambda_0}\right] \tag{12}$$

The group index n_g is defined as

$$n_g = \frac{c}{v_g} = n(\lambda_0) - \lambda_0 \frac{dn}{d\lambda_0} \tag{13}$$

In Table 10.1 we tabulated $n(\lambda_0)$, $dn/d\lambda_0$, and $n_g(\lambda_0)$ for pure silica as a function of the free space wavelength λ_0. In Fig. 10.3 we have plotted (for pure silica) the wavelength variations of the group velocity v_g; note that the group velocity attains a maximum value at $\lambda_0 \approx 1.27$ μm. As we will show later in this chapter (and in Chap. 27), this wavelength is of great significance in optical communication systems.

Example 10.1 For pure silica the refractive index variation in the wavelength domain 0.5 μm < λ_0 < 1.6 μm can be assumed to be given by the approximate empirical formula

$$n(\lambda_0) \approx C_0 - a\lambda_0^2 + \frac{a}{\lambda_0^2} \tag{14}$$

where $C_0 \approx 1.451$, $a \approx 0.003$, and λ_0 is measured in μm. [A more accurate expression for $n(\lambda_0)$ is given in Prob. 10.6.] Simple algebra shows

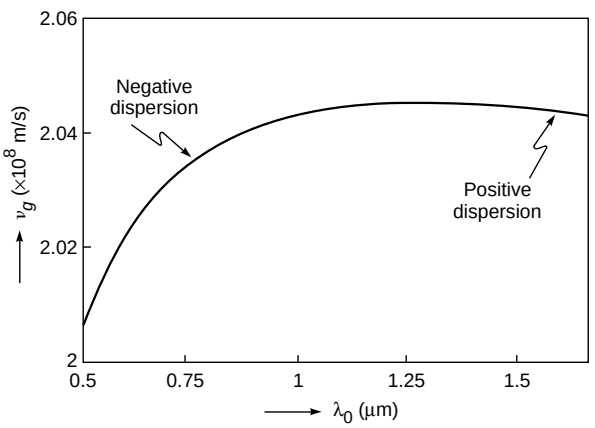


Fig. 10.3 Variation of the group velocity v_g with wavelength for pure silica.

$$n_g(\lambda_0) = C_0 + a\lambda_0^2 + \frac{3a}{\lambda_0^2} \tag{15}$$

Thus at $\lambda_0 = 1$ μm,

$$n(\lambda_0) \approx 1.451$$

and

$$n_g(\lambda_0) \approx 1.463$$

indicating that the difference between group and phase velocities is about 0.8%. More accurate values of $n(\lambda_0)$ and $n_g(\lambda_0)$ (as obtained by using the expression given in Prob. 10.6) are given in Table 10.1.

Using Table 10.1 we find that in pure silica, for

$$\lambda_0 = 0.80\ \mu\text{m} \qquad v_g = c/n_g = 2.0444 \times 10^8\ \text{m s}^{-1}$$

and for

$$\lambda_0 = 0.85\ \mu\text{m} \qquad v_g = c/n_g = 2.0464 \times 10^8\ \text{m s}^{-1}$$

implying that (for $\lambda_0 < 1.27$ μm) higher-wavelength components travel faster; similarly for $\lambda_0 > 1.27$ μm, lower- wavelength components travel faster. Now, every source of light has a certain wavelength spread, which is usually referred to as the spectral width of the source. Thus a white light source (such as coming from the Sun) has a spectral width of about 3000 Å; on the other hand, a light-emitting diode (usually abbreviated as LED) has a spectral width of about 25 nm, and a typical laser diode (usually abbreviated as LD) operating around 1.3 μm has a spectral width of about 2 nm; this spectral width is usually denoted by $\Delta\lambda_0$. Since each wavelength component (of a pulse) will travel with a slightly different group velocity, it will, in general, result in the broadening of the pulse. To calculate this broadening, we note that the time taken by a pulse to traverse a length L of the dispersive medium is

$$\tau = \frac{L}{v_g} = \frac{L}{c}\left[n(\lambda_0) - \lambda_0 \frac{dn}{d\lambda_0}\right] \tag{16}$$

Since the RHS depends on λ_0, the above equation implies that different wavelengths will travel with different group velocities in propagating through a certain length of the dispersive medium. Thus the pulse broadening is given by

$$\begin{aligned}\Delta\tau_m &= \tau(\lambda_0 + \Delta\lambda_0) - \tau(\lambda_0) \\ &= \frac{d\tau}{d\lambda_0}\Delta\lambda_0 \\ &= -\frac{L\Delta\lambda_0}{\lambda_0 c}\left(\lambda_0^2 \frac{d^2 n}{d\lambda_0^2}\right)\end{aligned} \tag{17}$$

The quantity $\Delta\tau_m$ is usually referred as material dispersion because it is due to the material properties of the medium—hence the subscript m. In Eq. (17), the quantity inside the

Table 10.1 Values of n, n_g, and D_m for pure silica. The numerical values in the table have been calculated using the refractive index variation as given in Ref. 2 (see Prob. 10.6).

λ_0 (μm)	$n(\lambda_0)$	$\frac{dn}{d\lambda_0}$ (μm^{-1})	$n_g(\lambda_0)$	$\frac{d^2n}{d\lambda_0^2}$ (μm^{-2})	D_m (ps nm^{-1} km^{-1})
0.70	1.45561	–0.02276	1.47154	0.0741	–172.9
0.75	1.45456	–0.01958	1.46924	0.0541	–135.3
0.80	1.45364	–0.01725159	1.46744	0.0400	–106.6
0.85	1.45282	–0.01552236	1.46601	0.0297	–84.2
0.90	1.45208	–0.01423535	1.46489	0.0221	–66.4
0.95	1.45139	–0.01327862	1.46401	0.0164	–51.9
1.00	1.45075	–0.01257282	1.46332	0.0120	–40.1
1.05	1.45013	–0.01206070	1.46279	0.0086	–30.1
1.10	1.44954	–0.01170022	1.46241	0.0059	–21.7
1.15	1.44896	–0.01146001	1.46214	0.0037	–14.5
1.20	1.44839	–0.01131637	1.46197	0.0020	–8.14
1.25	1.44783	–0.01125123	1.46189	0.00062	–2.58
1.30	1.44726	–0.01125037	1.46189	–0.00055	2.39
1.35	1.44670	–0.01130300	1.46196	–0.00153	6.87
1.40	1.44613	–0.01140040	1.46209	–0.00235	10.95
1.45	1.44556	–0.01153568	1.46229	–0.00305	14.72
1.50	1.44498	–0.01170333	1.46253	–0.00365	18.23
1.55	1.44439	–0.01189888	1.46283	–0.00416	21.52
1.60	1.44379	–0.01211873	1.46318	–0.00462	24.64

parentheses is dimensionless. Indeed, after propagating through a length L of the dispersive medium, a pulse of temporal width τ_0 will get broadened to τ_f where

$$\tau_f^2 \approx \tau_0^2 + (\Delta\tau_m)^2 \qquad \textbf{(18)}$$

In the next section we will explicitly show this for a Gaussian pulse. From Eq. (17) we see that the broadening of the pulse is proportional to the length L traversed in the medium and also to the spectral width of the source $\Delta\lambda_0$. We assume

$$\Delta\lambda_0 = 1 \text{ nm} = 10^{-9}\text{m} \qquad \text{and} \qquad L = 1 \text{ km} = 1000 \text{ m}$$

Thus

$$\Delta\tau_m = -\frac{L\,\Delta\lambda_0}{\lambda_0 c}\left(\lambda_0^2\frac{d^2n}{d\lambda_0^2}\right)$$

$$= -\frac{(1000\text{ m})\times(10^{-9}\text{ m})}{\left[\lambda_0(\mu\text{m})\times(10^{-6})\right]\left(3\times10^8\text{ m s}^{-1}\right)}\left(\lambda_0^2\frac{d^2n}{d\lambda_0^2}\right) \text{ s}$$

$$= -\frac{1}{3\lambda_0(\mu\text{m})}\left(\lambda_0^2\frac{d^2n}{d\lambda_0^2}\right)\times10^{-8} \text{ s}$$

$$= \frac{1}{3\lambda_0(\mu\text{m})}\left(\lambda_0^2\frac{d^2n}{d\lambda_0^2}\right)\times10^{4} \text{ ps}$$

where λ_0 is measured in μm and we have assumed $c \approx 3\times10^8\text{ m s}^{-1}$. We define the dispersion coefficient as

$$D_m = \frac{\Delta\tau_m}{L\,\Delta\lambda_0} \approx -\frac{1}{3\lambda_0}\left(\lambda_0^2\frac{d^2n}{d\lambda_0^2}\right)\times10^4 \quad \text{ps km}^{-1}\text{ nm}^{-1} \qquad \textbf{(19)}$$

The quantity D_m is usually referred to as the material dispersion coefficient (because it is due to the material properties of the medium), hence the subscript m on D. Thus $D_m = 15$ ps km^{-1} nm^{-1} implies that the pulse will broaden by 15 ps per kilometer length of the fiber per nanometer spectral width of the source. A medium is said to be characterized by positive dispersion when D_m is positive, and it is said to be characterized by negative dispersion when D_m is negative.

The spectral width of a pulse is usually due to the intrinsic spectral width of the source—which for a typical LED is about 25 nm and for a commercially available laser diode is about 1 to 2 nm. On the other hand, for a nearly monochromatic source, the intrinsic spectral width could be extremely small, and the actual spectral width of a pulse is determined from its finite duration (such a pulse is often referred to as a Fourier transformed pulse). Thus a 20 ps (Fourier transformed) pulse will have a spectral width

$$\Delta\nu \approx \frac{1}{20\times10^{-12}} \approx 5\times10^{11}\text{ Hz}$$

implying

$$\Delta\lambda_0 \approx \frac{\lambda_0^2\,\Delta\nu}{c} \approx 0.4 \text{ nm}$$

We may see that

$$\frac{d^2n}{d\lambda_0^2} \approx 0$$

around $\lambda_0 \approx 1.27$ μm. Indeed the wavelength $\lambda_0 \approx 1270$ nm is usually referred to as the zero material dispersion wavelength, and it is because of low material dispersion; the second- and third-generation optical communication systems operated around $\lambda_0 \approx 1300$ nm; more details will be given in Chap. 27.

Example 10.2 In first-generation optical communication system, one used LEDs with $\lambda_0 \approx 0.85$ μm and $\Delta\lambda_0 \approx 25$ nm. Now at $\lambda_0 \approx 0.85$ μm

$$\frac{d^2n}{d\lambda_0^2} \approx 0.030 \text{ (μm)}^{-2}$$

giving

$$D_m \approx -85 \text{ ps km}^{-1}\text{nm}^{-1}$$

the negative sign indicating that higher wavelengths travel faster than lower wavelengths. Thus for $\Delta\lambda_0 \approx 25$ nm, the actual broadening of the pulse will be

$$\Delta\tau_m \approx 2.1 \text{ ns km}^{-1}$$

implying that the pulse will broaden by 2.1 ns after traversing through 1 km of the silica fiber.

Example 10.3 In the fourth-generation optical communication systems, one uses laser diodes with $\lambda_0 = 1.55$ μm and $\Delta\lambda_0 \approx 2$ nm. Now at $\lambda_0 \approx 1.55$ μm

$$\frac{d^2n}{d\lambda_0^2} \approx 0.0042 \text{ (μm)}^{-2}$$

giving

$$D_m \approx +21.7 \text{ ps km}^{-1}\text{nm}^{-1}$$

the positive sign indicating that higher wavelengths travel slower than lower wavelengths. (Notice from Table 10.1 that for $\lambda_0 \geq 1.27$ μm, n_g increases with λ_0.) Thus for $\Delta\lambda_0 \approx 2$ nm, the actual broadening of the pulse will be

$$\Delta\tau_m \approx 43 \text{ ps km}^{-1}$$

implying that the pulse will broaden by about 43 ps after traversing through 1 km of the silica fiber.

10.3 GROUP VELOCITY OF A WAVE PACKET

The displacement corresponding to a one-dimensional plane wave propagating in the +z direction can be written in the form

$$E(z, t) = Ae^{i(\omega t - kz)} \tag{20}$$

where A represents the amplitude of the wave and

$$k(\omega) = \frac{\omega}{c} n(\omega) \tag{21}$$

with n being the refractive index of the medium. The wave described by Eq. (20) is said to describe a monochromatic wave which propagates with the phase velocity given by

$$v_p = \frac{\omega}{k} = \frac{c}{n} \tag{22}$$

We note here that, in general, A may be complex and if we write

$$A = |A|e^{i\phi}$$

then Eq. (20) becomes

$$E = |A|e^{i(\omega t - kz + \phi)}$$

The actual displacement is the real part of E and is, therefore, given by

$$\text{Actual electric field} = \text{Re}(E) = |A| \cos(\omega t - kz + \phi) \tag{23}$$

The plane wave represented by Eq. (20) is a practical impossibility because at an arbitrary value of z, the displacement is finite for *all* values of t; for example,

$$E(z = 0, t) = Ae^{+i\omega t}; \qquad -\infty < t < \infty \tag{24}$$

which corresponds to a sinusoidal variation for *all* values of time. In practice, the displacement is finite only over a certain domain of time, and we have what is known as a *wave packet*. A wave packet can always be expressed as a superposition of plane waves of different frequencies:

$$E(z, t) = \int_{-\infty}^{+\infty} A(\omega)e^{i(\omega t - kz)}\,d\omega \tag{25}$$

Obviously

$$E(z = 0, t) = \int_{-\infty}^{+\infty} A(\omega)e^{+i\omega t}\,d\omega \tag{26}$$

Thus, $E(z = 0, t)$ is the Fourier transform of $A(\omega)$, and using the results of Chap. 9, we obtain

$$A(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} E(z = 0, t)e^{-i\omega t}\,dt \tag{27}$$

Thus given $E(z = 0, t)$, we know we can determine $E(z, t)$ using the following recipe:

We determine $A(\omega)$ from Eq. (27), substitute it in Eq. (25), and carry out the resulting integration.

Example 10.4 Gaussian pulse: As an example, we consider a Gaussian pulse for which we may write

$$E(z = 0, t) = E_0 e^{-t^2/\tau_0^2} e^{+i\omega_0 t} \tag{28}$$

If we substitute Eq. (28), in Eq. (27), we obtain

$$A(\omega) = \frac{E_0}{2\pi}\int e^{-t^2/\tau_0^2} e^{-i(\omega-\omega_0)t}\, dt$$

$$= \frac{E_0\tau_0}{2\sqrt{\pi}}\exp\left[-\frac{1}{4}(\omega-\omega_0)^2\tau_0^2\right] \tag{29}$$

where we have used

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2+\beta x}\, dx = \sqrt{\frac{\pi}{\alpha}}\, e^{\beta^2/4\alpha} \tag{30}$$

(see App. A). In general, $A(\omega)$ can be complex, and as such one defines the power spectral density

$$S(\omega) = |A(\omega)|^2 \tag{31}$$

For the Gaussian pulse

$$S(\omega) = \frac{E_0^2\tau_0^2}{4\pi}\exp\left[-\frac{1}{2}(\omega-\omega_0)^2\tau_0^2\right] \tag{32}$$

In Fig. 10.4(a) we have plotted the function

$$E_0 e^{-t^2/\tau_0^2}\cos(\omega_0 t)$$

[which is the real part of Eq. (28)] for a 20 fs pulse ($\tau_0 = 20 \times 10^{-15}$s) corresponding to $\lambda_0 = 1$ μm ($\omega_0 \approx 6\pi \times 10^{14}$ Hz); the corresponding spectral density function $S(\omega)$ is plotted in Fig. 10.4(b). As can be seen, $S(\omega)$ is a very sharply peaked function of ω around $\omega = \omega_0$. The full width at half maximum of $S(\omega)$ (usually abbreviated as FWHM) is denoted by $\Delta\omega$; thus at

$$\omega = \omega_0 \pm \frac{1}{2}\Delta\omega$$

$S(\omega)$ attains one-half of its maximum value; the value of $\Delta\omega$ is obtained from the following equation

$$\frac{1}{2} = \exp\left[-\frac{(\Delta\omega)^2\tau_0^2}{8}\right]$$

or

$$\text{FWHM} = \Delta\omega = \frac{2\sqrt{2\ln 2}}{\tau_0} \approx \frac{2.35}{\tau_0} \tag{33}$$

Thus the Gaussian pulse of temporal width 20 fs has a frequency spread $\Delta\omega$ given by

$$\Delta\omega \approx 1.18 \times 10^{14} \text{ Hz} \tag{34}$$

So

$$\frac{\Delta\omega}{\omega_0} \approx 0.06$$

Note that to have clarity in the figure we have chosen a very small value of τ_0; usually τ_0 has a much larger value. A larger value of τ_0 will imply a much smaller value of $\Delta\omega$ (resulting in greater monochromaticity of the pulse), and obviously Fig. 10.4(b) will be much more sharply peaked; we will discuss this in greater detail in the chapter on coherence (Chap. 17).

Returning to Eq. (25), we consider the following cases:

10.3.1 Propagation in a Nondispersive Medium

For electromagnetic waves, the free space is a nondispersive medium in which all frequencies propagate with the sam velocity c; thus

$$k(\omega) = \frac{\omega}{c}$$

and Eq. (25) can be written in the form

$$\text{E}(z, t) = \int_{-\infty}^{+\infty} A(\omega) e^{-i\frac{\omega}{c}(z-ct)}\, d\omega \tag{35}$$

The right-hand side is a function of $z - ct$, and thus any pulse will propagate with velocity c without undergoing any distortion. Thus, for the Gaussian pulse given by Eq. (28),

$$E(z, t) = E_0 \exp\left[-\frac{(z-ct)^2}{c^2\tau_0^2}\right]\exp\left[-i\frac{\omega_0}{c}(z-ct)\right] \tag{36}$$

which represents a distortionless propagation of a Gaussian pulse in a nondispersive medium[2]. In 10.5 we have shown Fig. 10.5 the distortionless propagation of a 20 fs pulse.

10.3.2 Propagation in a Dispersive Medium

For a wave propagating in a medium characterized by the refractive index variation $n(\omega)$, we will have

$$k(\omega) = \frac{\omega}{c}\, n(\omega)$$

Now, in most problems, $A(\omega)$ is a very sharply peaked function [see, e.g., Fig. 10.4(b)] so that we may write

$$E(z, t) \approx \int_{\omega_0-\Delta\omega}^{\omega_0+\Delta\omega} A(\omega) e^{i[\omega t - k(\omega)z]}\, d\omega \tag{37}$$

[2] Whereas Eq. (36) follows directly from Eq. (35), it is left as an exercise to the reader to show that if we substitute $A(\omega)$ from Eq. (29) into Eq. (35), we readily get Eq. (36).

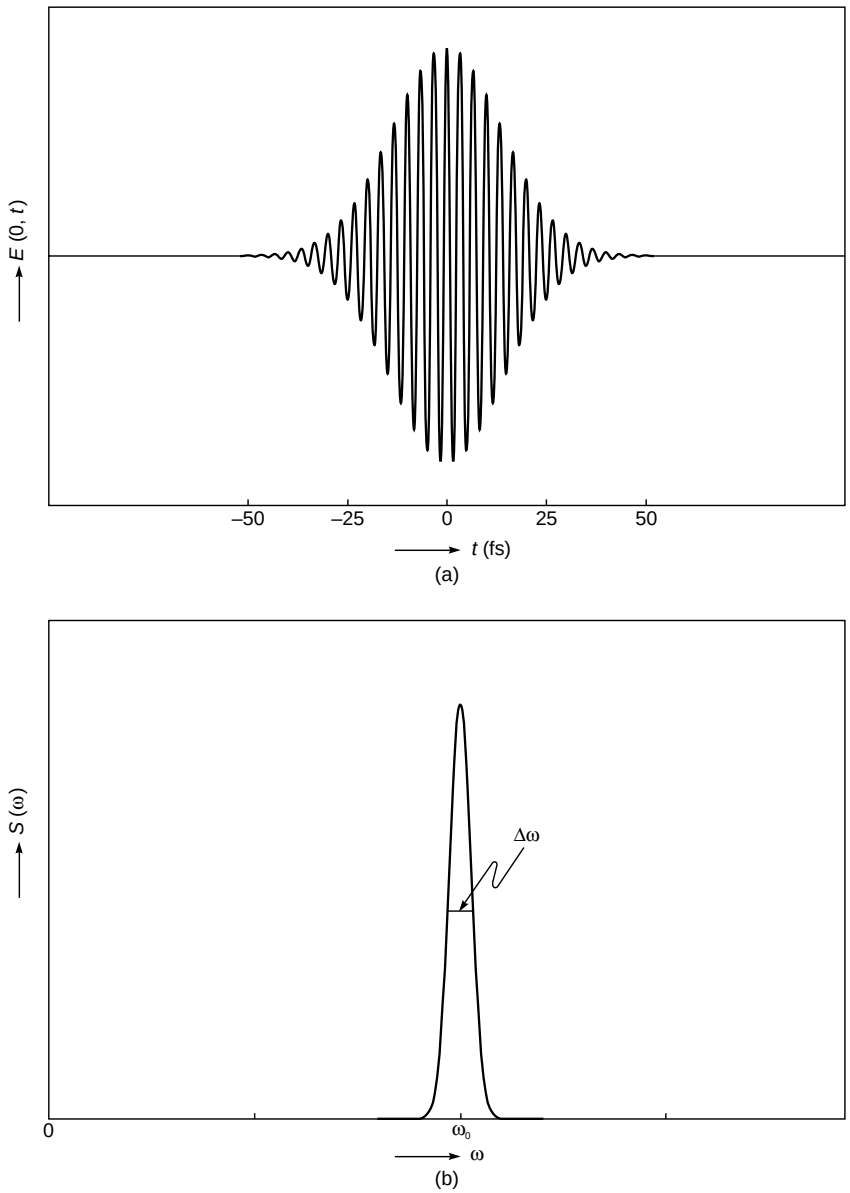


Fig. 10.4 (a) A 20 fs (= 20×10^{-15} s) Gaussian pulse corresponding to $\lambda_0 = 1$ μm; (b) the corresponding frequency spectrum, which is usually a very sharply peaked function around $\omega = \omega_0$.

because for $\omega > \omega_0 + \Delta\omega$ and for $\omega < \omega_0 - \Delta\omega$, the function $A(\omega)$ is negligibly small. In this tiny domain of integration, we may make a Taylor series expansion of $k(\omega)$

$$k(\omega) = k(\omega_0) + (\omega - \omega_0)\left.\frac{dk}{d\omega}\right|_{\omega=\omega_0} + \frac{1}{2}(\omega - \omega_0)^2 \left.\frac{d^2k}{d\omega^2}\right|_{\omega=\omega_0} + \cdots \quad \textbf{(38)}$$

or $$k(\omega) = k_0 + \frac{1}{v_g}(\omega - \omega_0) + \frac{1}{2}(\omega - \omega_0)^2 \gamma \quad \textbf{(39)}$$

where $$k_0 \equiv k(\omega_0) \quad \textbf{(40)}$$

$$\frac{1}{v_g} \equiv \left.\frac{dk}{d\omega}\right|_{\omega=\omega_0} \quad \textbf{(41)}$$

and $$\gamma \equiv \left.\frac{d^2k}{d\omega^2}\right|_{\omega=\omega_0} \quad \textbf{(42)}$$

We have *defined* v_g through Eq. (41)—we will show below that the envelope of the pulse moves with velocity v_g which

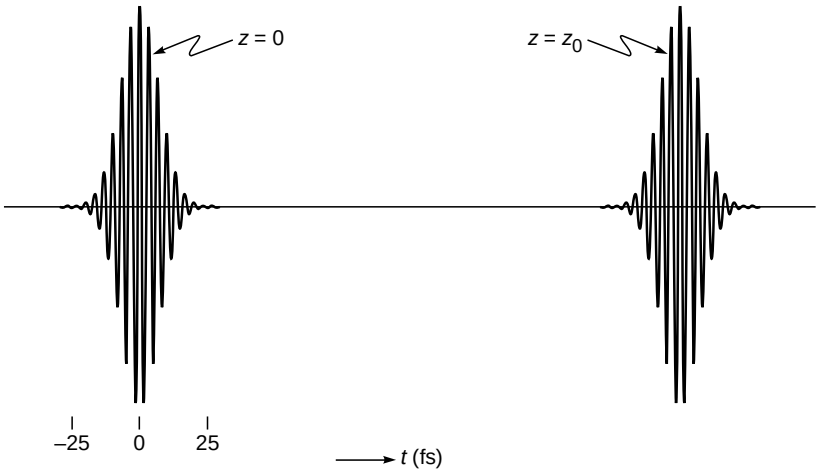


Fig. 10.5 Distortionless propagation of a Gaussian pulse in a nondispersive medium.

is the group velocity. Now, if we retain only the first two terms on the RHS of Eq. (39), then Eq. (37) gives

$$E(z,t) \approx \int_{-\infty}^{+\infty} A(\omega)\exp\left[-i\left(k_0 z+\frac{\omega-\omega_0}{v_g}z-\omega t\right)\right]d\omega \qquad \textbf{(43)}$$

where we have replaced the limits from $-\infty$ to $+\infty$ because, in any case, the contribution from the region $|\omega - \omega_0| > \Delta\omega$ is going to be extremely small. By writing

$$\omega t = (\omega - \omega_0)t + \omega_0 t \qquad \textbf{(44)}$$

Eq. (43) can be rewritten in the form

$$E(z,t) \approx \underbrace{e^{i(\omega_0 t-k_0 z)}}_{\text{Phase term}} \underbrace{\int_{-\infty}^{+\infty} A(\Omega)e^{-\frac{i\Omega}{v_g}(z-v_g t)}d\Omega}_{\text{Envelope term}} \qquad \textbf{(45)}$$

where

$$\Omega \equiv \omega - \omega_0 \qquad \textbf{(46)}$$

We see that in the envelope term, z and t do not appear independently but only as $z - v_g t$; thus, the envelope of the pulse moves undistorted with the group velocity

$$v_g = \frac{1}{(dk/d\omega)_{\omega_0}} \qquad \textbf{(47)}$$

Thus if we neglect γ [and other higher-order terms in Eq. (39)], the pulse moves undistorted with group velocity v_g.

Next, if we take into account all three terms in Eq. (39), we obtain

$$E(z,t) \approx \underbrace{e^{i(\omega_0 t-k_0 z)}}_{\text{Phase term}} \underbrace{\int_{-\infty}^{+\infty} A(\Omega)\exp\left[i\Omega\left(t-\frac{z}{v_g}\right)-\frac{i}{2}\Omega^2\gamma z\right]d\Omega}_{\text{Envelope term}} \qquad \textbf{(48)}$$

For the Gaussian pulse [see Eq. (28)], $A(\omega)$ is given by Eq. (29); if we now substitute $A(\omega)$ in the above equation and use Eq. (30) to carry out the integration, we readily obtain

$$E(z,t) = \frac{E_0}{\sqrt{1+ip}}e^{i(\omega_0 t-k_0 z)}\exp\left[-\frac{\left(t-z/v_g\right)^2}{\tau_0^2(1+ip)}\right] \qquad \textbf{(49)}$$

where

$$p \equiv \frac{2\gamma z}{\tau_0^2} \qquad \textbf{(50)}$$

The corresponding intensity distribution is given by

$$I(z,t) = \frac{I_0}{\tau(z)/\tau_0}\exp\left[-\frac{2\left(t-z/v_g\right)^2}{\tau^2(z)}\right] \qquad \textbf{(51)}$$

where

$$\tau^2(z) \equiv \tau_0^2(1+p^2) \qquad \textbf{(52)}$$

In Fig. 10.6 we have plotted the time variation of the intensity at different values of z. From Eq. (52) we find that as the pulse propagates, it undergoes temporal broadening. We define the pulse broadening $\Delta\tau$ as

$$\Delta\tau = \sqrt{\tau^2(z)-\tau_0^2} = |p|\tau_0 = \frac{2|\gamma|z}{\tau_0} \qquad \textbf{(53)}$$

Now

$$\gamma = \frac{d^2k}{d\omega^2} = \frac{d}{d\omega}\left[\frac{1}{c}\left(n-\lambda_0\frac{dn}{d\lambda_0}\right)\right] = \frac{1}{c}\frac{d}{d\lambda_0}\left[n(\lambda_0)-\lambda_0\frac{dn}{d\lambda_0}\right]\frac{d\lambda_0}{d\omega} = \frac{\lambda_0}{2\pi c^2}\left[\lambda_0^2\frac{d^2n}{d\lambda_0^2}\right] \qquad \textbf{(54)}$$

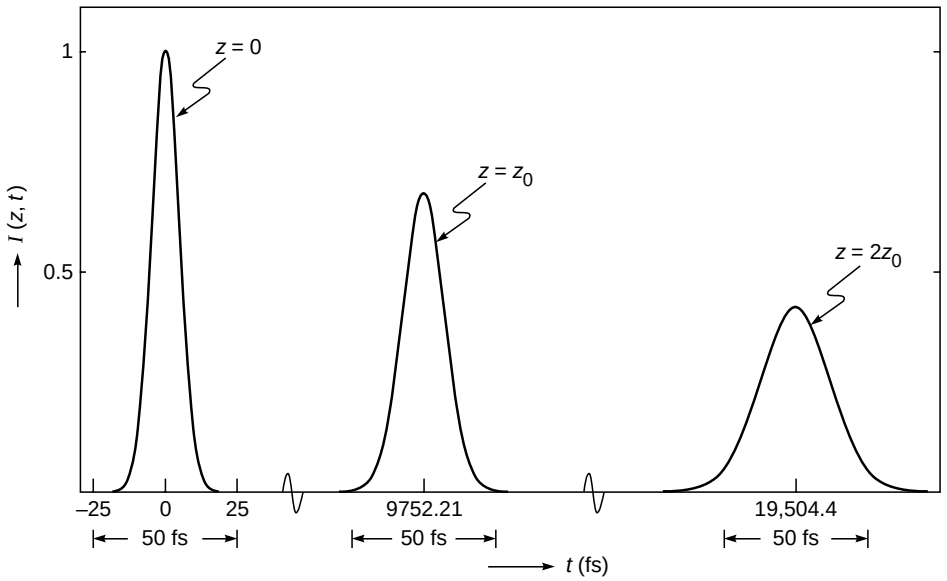


Fig. 10.6 The time variation of the intensity at different values of z; notice the temporal broadening of the pulse.

where the quantity inside the square brackets is dimensionless. Further, since the spectral width of the Gaussian pulse is given by [see Eq. (33)]

$$\Delta\omega \approx \frac{2}{\tau_0} \tag{55}$$

we may write

$$\frac{1}{\tau_0} \approx \frac{1}{2}\Delta\omega \approx \frac{1}{2}\frac{2\pi c}{\lambda_0^2}\left|\Delta\lambda_0\right| \tag{56}$$

Substituting for τ_0 from Eq. (56) and for γ from Eq. (54) in Eq. (53), we get

$$\Delta\tau = \frac{z}{\lambda_0 c}\left|\lambda_0^2 \frac{d^2n}{d\lambda_0^2}\right| \Delta\lambda_0 \tag{57}$$

which is identical to the result obtained earlier section [see Eq. (17)].

Example 10.5 As an example, we assume $\lambda_0 = 1.55$ μm. For pure silica, at this wavelength (see Table 10.1)

$$\frac{d^2n}{d\lambda_0^2} \approx -0.004165\ (\mu\text{m})^{-2}$$

Thus

$$\gamma \approx -\frac{1.55\times10^{-6}}{2\pi\times9\times10^{16}}(1.55\times1.55\times0.004165)$$

$$\approx -2.743\times10^{-26}\ \text{m}^{-1}\text{s}^2$$

For a 100 ps pulse propagating through a 2 km long fiber

$$\Delta\tau \approx \frac{2\times2.743\times10^{-26}\times2\times10^3}{10^{-10}} \approx 1.1\ \text{ps}$$

On the other hand, for a 10 fs pulse, at $z = 2z_0 = 4$ mm we will have

$$\Delta\tau \approx 22\ \text{fs}$$

implying

$$\tau_f \approx [\tau_0^2 + (\Delta\tau)^2]^{1/2} \approx 25\ \text{ps}$$

showing that a 10 fs pulse doubles its temporal width after propagating through a very small distance (see Figs. 10.7 and 10.8).

10.3.3 The Chirping of the Dispersed Pulse

If we carry out simple manipulations, Eq. (49) can be written in the form

$$E(z,t) = \frac{E_0}{[\tau(z)/\tau_0]^{1/2}} \exp\left[-\frac{\left(t - z/v_g\right)^2}{\tau^2(z)}\right]$$

$$\times \exp\left[i\left(\Phi(z,t) - k_0 z\right]\right. \tag{58}$$

where the phase term is given by

$$\Phi(z,t) = \omega_0 t + \kappa\left(t - \frac{z}{v_g}\right)^2 - \frac{1}{2}\tan^{-1}p \tag{59}$$

and

$$\kappa(z) = \frac{p}{\tau_0^2(1+p^2)} \tag{60}$$

Equation (59) represents the phase term, and the instantaneous frequency is given by

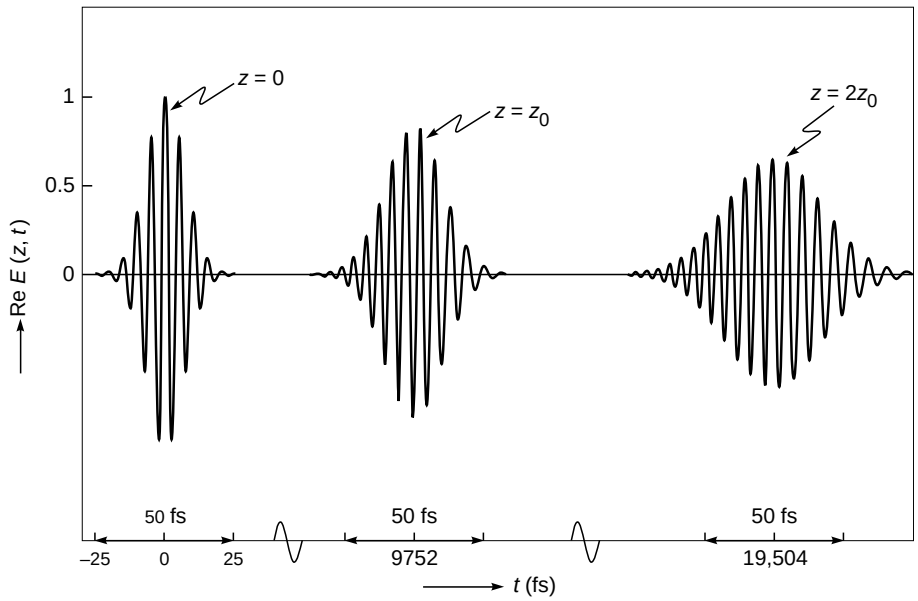


Fig. 10.7 The temporal broadening of a 10 fs unchirped Gaussian pulse (λ_0 = 1.55 μm) propagating through silica. Notice that since dispersion is positive, the pulse gets down-chirped.

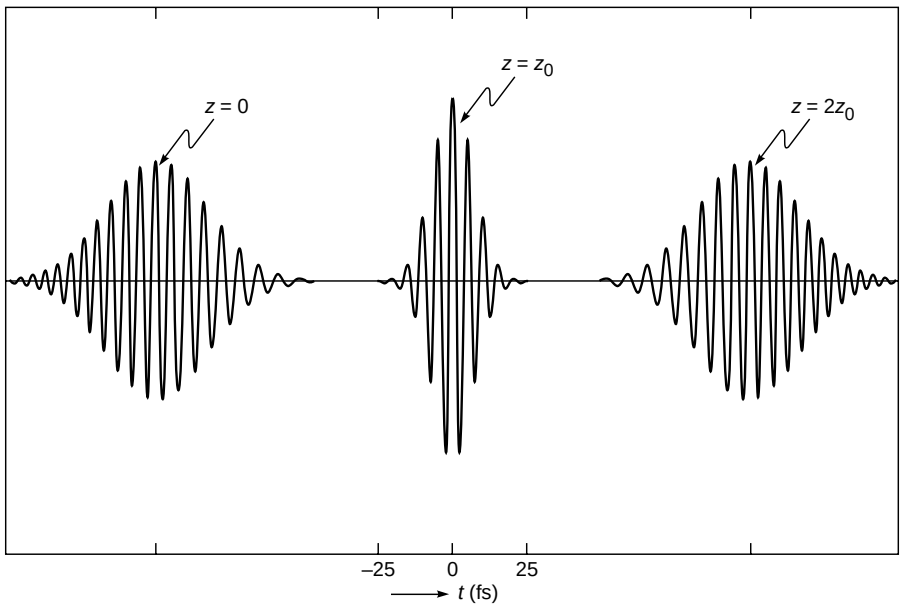


Fig. 10.8 If a down-chirped pulse is passed through a medium characterized by negative dispersion, it will get compressed until it becomes unchirped, and then it will broaden again with opposite chirp.

$$\omega(t) = \frac{\partial \Phi}{\partial t} = \omega_0 + 2\kappa\left(t - \frac{z}{v_g}\right) \qquad \textbf{(61)}$$

showing that $\omega(t)$ changes within the pulse. The frequency chirp is therefore given by

$$\Delta\omega = \omega(t) - \omega_0 = 2\kappa\left(t - \frac{z}{v_g}\right) \qquad \textbf{(62)}$$

Example 10.6 In continuation of Example 10.5, we assume λ_0 = 1.55 μm and consider the chirping produced in a 100 ps pulse propagating in pure silica at z = 2 km. Now

$$p = \frac{2\gamma z}{\tau_0^2} = -\frac{2 \times 2.743 \times 10^{-26} \times 2 \times 10^3}{(100 \times 10^{-12})^2}$$

$$\approx -0.011$$

At

$$t - \frac{z}{v_g} = -50 \text{ ps}$$

(i.e., at the front end of the pulse)

$$\Delta\omega \approx \frac{2p}{\tau_0^2(1+p^2)}(-50\times10^{-12})$$

$$\approx +\frac{2\times0.011\times50\times10^{-12}}{(100\times10^{-12})^2}$$

$$= +1.1 \times 10^8 \text{ Hz}$$

Thus at the leading edge of the pulse, the frequencies are slightly higher which is usually referred to as *blue-shifted*. Notice

$$\frac{\Delta\omega}{\omega_0} \approx 9 \times 10^{-8}$$

At

$$t = \frac{z}{v_g}, \qquad \Delta\omega = 0$$

and at

$$t - \frac{z}{v_g} = +50 \text{ ps}$$

(i.e., at the trailing edge of the pulse)

$$\Delta\omega \approx -1.1 \times 10^8 \text{ Hz}$$

Thus, at the trailing edge of the pulse, the frequencies are slightly lower which is usually referred to as *red-shifted*.

From Example 10.6, we can conclude the following:

For positive dispersion (i.e., negative value of γ), p and κ will also be negative, implying that the instantaneous frequency (within the pulse) decreases with time (we are of course assuming $z > 0$); this is known as a *down-chirped pulse* in which the *leading edge* of the pulse ($t < z/v_g$) is *blue-shifted* (i.e., it has frequency higher than ω_0) and the *trailing edge* of the pulse ($t > z/v_g$) is *red-shifted* (i.e., it has frequency lower than ω_0).

This is shown in Fig. 10.7 where at $t = 0$ we have an unchirped pulse. As the pulse propagates farther, it will get further broadened and also get further *down-chirped*.

From Eq. (61) it can be readily seen that at negative values of z, p (and therefore κ) will be positive and the *leading edge* of the pulse ($t < z/v_g$) will be *red-shifted* (i.e., it will have frequency lower than ω_0) and the *trailing edge* of the pulse ($t > z/v_g$) will be *blue-shifted* (i.e., it will have frequency higher than ω_0).

This implies that we will have an up-chirped pulse. Thus if an up-chirped pulse is passed through a medium characterized by positive dispersion, it will get compressed until it becomes unchirped, and then it will broaden again with opposite chirp.

Similarly we can discuss the case of negative dispersion (implying a positive value of γ). If a down-chirped pulse is passed through a medium characterized by negative dispersion, it will get compressed until it becomes unchirped, and then it will broaden again with opposite chirp (see Fig. 10.8).

10.4 SELF PHASE MODULATION

As a pulse propagates through a dispersive medium, the frequency spectrum remains the same—i.e., no new frequencies are generated. Different frequencies superpose with different phases to distort the temporal shape of the pulse (see Prob. 10.10). New frequencies are generated when the medium is nonlinear — we briefly discuss this here.

The refractive index of any material is a constant only for small intensities of the propagating laser beam. If the intensities are large, the refractive index variation is approximately given by

$$n \simeq n_0 + n_2 I \qquad \textbf{(63)}$$

where n_2 is a constant and I represents the intensity of the beam. For example, for fused silica, $n_0 \simeq 1.47$ and $n_2 \simeq 3.2 \times 10^{-20}$ m^2 W^{-1}. Further, if the effective area of the light beam is A_{eff}, then the intensity is given by

$$I = \frac{P}{A_{\text{eff}}} \qquad \textbf{(64)}$$

where P is the power associated with the light beam. Now in a single mode fiber, the spot size w_0 of the beam is about 5 μm (see Examples 29.8 and 29.9). Thus the effective[3] cross-sectional area of the beam, $A_{\text{eff}} \approx \pi w_0^2 \approx 50$ μm^2. For a 5 mW laser beam propagating through such a fiber, the resultant intensity is given by

$$I = \frac{P}{A_{\text{eff}}} \approx \frac{5\times10^{-3}\text{ W}}{50\times10^{-12}\text{ m}^2} = 10^8 \text{ W m}^{-2} \qquad \textbf{(65)}$$

Thus the change in refractive index is given by

$$\Delta n = n_2 I \simeq 3.2 \times 10^{-12} \qquad \textbf{(66)}$$

Although this is very small, but when the beam propagates over an optical fiber over long distances (a few hundred to a few thousand kilometers), the accumulated nonlinear effects can be significant. That is the great advantage of the optical fiber—the beam remains confined to a very small area for long distances!

We consider a laser pulse (of frequency ω_0) propagating through an optical fiber; the effective propagation constant

[3] Values adapted from Ref. 2.

is given by

$$k = \frac{\omega_0}{c}(n_0 + n_2 I) = \frac{\omega_0}{c}\left[n_0 + n_2 \frac{P(t)}{A_{\text{eff}}}\right] \quad \textbf{(67)}$$

Thus, for such a propagating beam, the phase term is approximately given by

$$e^{+i(\omega_0 t - kz)} = \exp\left\{+i\left[\omega_0 t - \frac{\omega_0}{c}\left(n_0 + n_2 \frac{P(t)}{A_{\text{eff}}}\right)z\right]\right\} = e^{+i\Phi}$$

where the phase Φ is defined as

$$\Phi(z, t) \equiv \omega_0 t - \frac{\omega_0}{c}\left(n_0 + n_2 \frac{P(t)}{A_{\text{eff}}}\right)z \quad \textbf{(68)}$$

We can define an instantaneous frequency as cf. [Eq. (61)]

$$\omega(t) \equiv \frac{\partial \Phi}{\partial t} = \omega_0 - g\,\frac{dP(t)}{dt}z$$

where

$$g = \frac{n_2 \omega_0}{c A_{\text{eff}}} = \frac{2\pi n_2}{\lambda_0 A_{\text{eff}}} \quad \textbf{(69)}$$

For $A_{\text{eff}} \approx 50\ \mu\text{m}^2$, $\lambda_0 \approx 1.55\ \mu\text{m}$ and $n_2 \approx 3.2 \times 10^{-20}\ \text{m}^2\ \text{W}^{-1}$, $g \approx 2.6 \times 10^{-3}\ \text{W}^{-1}\ \text{m}^{-1}$.

Now, for a Gaussian pulse propagating with group velocity v_g [see Eq. (51)]

$$P(z,t) = P_0 \exp\left[-\frac{2\left(t - z/v_g\right)^2}{\tau_0^2}\right]$$

where we have neglected dispersion [i.e., $p = 0$ in Eqs. (49) and (52)]. Thus

$$\omega(t) = \omega_0\left[1 + \frac{2gz}{\omega_0 \tau_0^2} P_0\left(t - \frac{z}{v_g}\right)\exp\left[-\frac{2\left(t - z/v_g\right)^2}{\tau_0^2}\right]\right]$$

For $\lambda_0 = 1.55\ \mu\text{m}$

$$\omega_0 = \frac{2\pi c}{\lambda_0} = \frac{2\pi \times 3 \times 10^8}{1.55 \times 10^{-6}} \approx 1.22 \times 10^{15}\ \text{s}^{-1}$$

Further, for $P_0 = 15$ mW, $\tau_0 = 20$ fs, and $z = 200$ km,

$$\frac{2gzP_0}{\omega_0 \tau_0^2}\left(t - \frac{z}{v_g}\right) = \frac{2 \times 2.6 \times 10^{-3} \times 2 \times 10^5 \times 15 \times 10^{-3}}{1.22 \times 10^{15} \times (20 \times 10^{-15})^2}\left(t - \frac{z}{v_g}\right)$$

$$\approx 3.2 \times 10^{13}\left(t - \frac{z}{v_g}\right)$$

$$= \begin{cases} +0.64 & \text{for } t - \dfrac{z}{v_g} \approx 20\ \text{fs (trailing edge of pulse)} \\ -0.64 & \text{for } t - \dfrac{z}{v_g} \approx -20\ \text{fs (front end of pulse)} \end{cases}$$

Thus the instantaneous frequency within the pulse changes with time leading to chirping of the pulse, as shown in Fig. 10.9; this is known as self phase modulation (usually abbreviated as SPM). Note that since the pulse width has not changed, but the pulse is chirped, the frequency content of the pulse has increased. Thus SPM leads to generation of

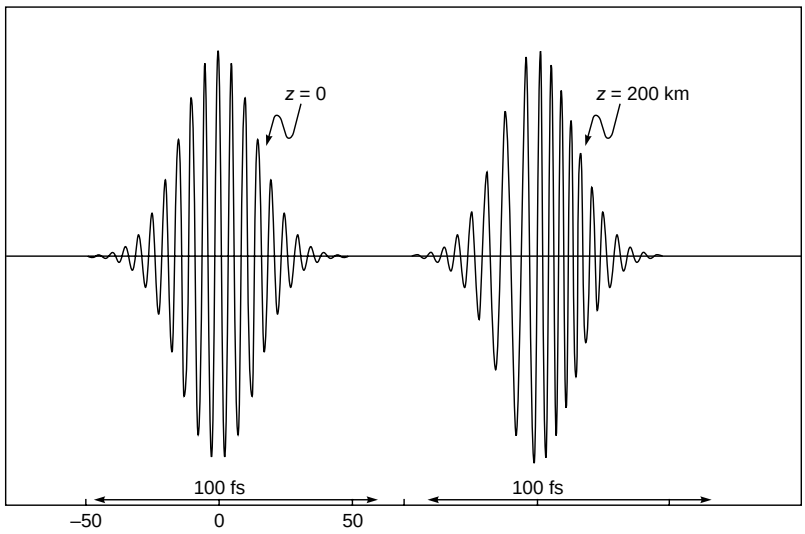


Fig. 10.9 Due to self phase modulation, the instantaneous frequency within the pulse changes with time leading to chirping of the pulse. Calculations correspond to $P_0 = 15$ mW, $\lambda_0 = 1550$ nm, $\tau_0 = 20$ fs, $A_{\text{eff}} = 5\ \mu\text{m}^2$, and $v_g = 2 \times 10^8$ m/s.

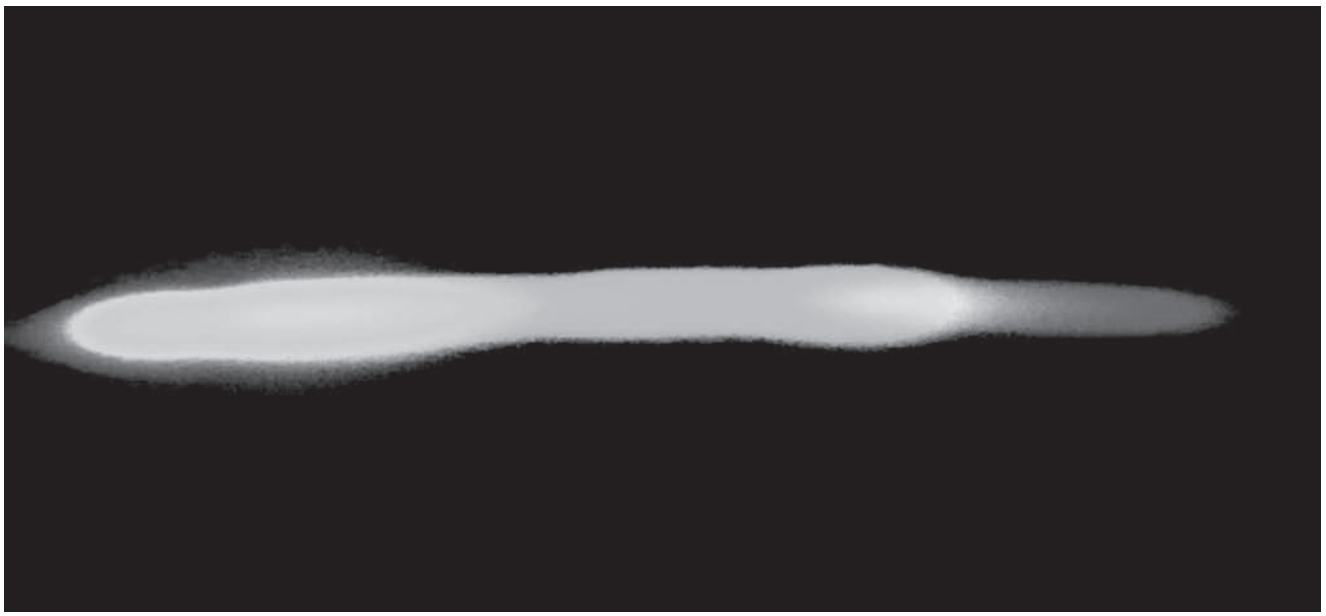


Fig. 10.10 Laser pulses of 80 fs duration having a wavelength 800 nm (and total energy of 1.6 nJ) are incident on a special optical fiber known as a holey fiber, in which a silica core is surrounded by a periodic lattice of air holes; holey fibers are characterized by very small mode field diameters, which leads to very high intensities. Because of the high intensities, SPM (self phase modulation) and other nonlinear effects can be observed; these nonlinear effects result in the generation of new frequencies. In this experiment, the entire visible spectrum gets generated which can be observed by passing the light coming out of the optical fiber through a prism. The repetition rate of the laser pulses is 82 MHz. The special fibers were fabricated by Dr. Shyamal Bhadra and Dr. Kamal Dasgupta and their group at CGCRI, Kolkata, and the supercontinuum generation was observed by Prof. Ajoy Kar and Dr. Henry Bookey at Heriot Watt University, Edinburgh. A color photograph appears on the cover of the book. Photograph courtesy Prof. Ajoy Kar.

new frequencies. Indeed by passing a pulse through a fiber characterized by very small cross-sectional area (so that the value of g is large), it is possible to generate the entire visible spectrum (see Fig. 10.10).

Summary

- When we switch a light source on and off, we produce a pulse. This pulse propagates through a medium with what is known as the group velocity, which is given by

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

For a medium characterized by the refractive index variation $n(\omega)$

$$k(\omega) = \frac{\omega}{c} n(\omega)$$

the group velocity is given by

$$\frac{1}{v_g} = \frac{1}{c}\left[n(\lambda_0) - \lambda_0 \frac{dn}{d\lambda_0}\right]$$

where λ_0 is the wavelength in free space and $c \approx 3 \times 10^8$ m/s is the speed of light in free space.

- After traversing through a distance L in a dispersive medium, a pulse will broaden by an amount

$$\Delta t_m = -\frac{L\,\Delta\lambda_0}{\lambda_0 c}\left(\lambda_0^2 \frac{d^2 n}{d\lambda_0^2}\right)$$

where $\Delta\lambda_0$ is the spectral width of the source; the subscript m denotes that the fact we are considering material dispersion. The dispersion coefficient is given by

$$D_m = \frac{\Delta t_m}{L\,\Delta\lambda_0} \approx -\frac{1}{3\lambda_0}\left(\lambda_0^2 \frac{d^2 n}{d\lambda_0^2}\right) \times 10^4 \text{ ps km}^{-1}\text{nm}^{-1}$$

where λ_0 is measured in μm and we have assumed $c \approx 3 \times 10^8$ m/s. For example, for silica, at $\lambda_0 = 1.55$ μm, $d^2n/d\lambda_0^2 \approx -0.00416$ (μm^{-2}) and $D_m \approx +22$ ps per kilometer (length of the medium) per nanometer (spectral width of the source). On the other hand, for silica $d^2n/d\lambda_0^2 \approx 0$ around $\lambda_0 \approx 1.27$ μm. Indeed the wavelength $\lambda_0 \approx 1.27$ μm is usually referred to as the zero material dispersion wavelength, and it is because of

low material dispersion; the second- and third- generation optical communication systems operated around $\lambda_0 \approx 1.3$ μm.

◆ For a Gaussian pulse

$$E(z = 0, t) = E_0 \exp\left(-\frac{t^2}{\tau_0^2}\right) e^{+i\omega_0 t}$$

the temporal width after propagating through a distance z is given by $\tau(z) = \tau_0\sqrt{1+p^2}$; thus the temporal broadening is given by

$$\Delta\tau = \sqrt{\tau^2(z) - \tau_0^2} = |p|\tau_0$$

where

$$p = \frac{2}{\tau_0^2}\cdot\frac{\lambda_0}{2\pi c^2}\left(\lambda_0^2\frac{d^2n}{d\lambda_0^2}\right) z$$

Thus at $\lambda_0 \approx 1.55$ μm, for a $\tau_0 \approx 100$ ps pulse (propagating in pure silica), $\Delta\tau \approx 0.55$ ps km^{-1}.

Problems

10.1 Using the empirical formula given by Eq. (14) calculate the phase and group velocities in silica at $\lambda_0 = 0.7$, 0.8, 1.0, 1.2, and 1.4 μm. Compare with the (more accurate) values given in Table 10.1.
[**Ans:** $n(\lambda_0) \approx 1.456$, 1.454, 1.451, 1.449, 1.455;
$n_g(\lambda_0) \approx 1.4708$, 1.4670, 1.4630, 1.4616, 1.4615].

10.2 For pure silica we may assume the empirical formula

$$n(\lambda_0) \approx 1.451 - 0.003\left(\lambda_0^2 - \frac{1}{\lambda_0^2}\right)$$

where λ_0 is measured in μm.
(a) Calculate the zero dispersion wavelength.
(b) Calculate the material dispersion at 800 nm in ps km^{-1} nm^{-1}.
[**Ans:** 1.32 μm; –101 ps km^{-1} nm^{-1}]

10.3 Let

$$n(\lambda_0) = n_0 + A\lambda_0$$

where λ_0 is the free space wavelength. Derive expressions for phase and group velocities.
[**Ans:** $v_g = c/n_0$]

10.4 Consider a LED source emitting light of wavelength 850 nm and having a spectral width of 50 nm. Using Table 10.1, calculate the broadening of a pulse propagating in pure silica.
[**Ans:** 4.2 ns km^{-1}]

10.5 In 1836 Cauchy gave the following approximate formula to describe the wavelength dependence of refractive index in glass in the visible region of the spectrum:

$$n(\lambda) = A + \frac{B}{\lambda_0^2}$$

Now (see also Table 12.2)

$$\left.\begin{aligned} n(\lambda_1) &= 1.50883 \\ n(\lambda_2) &= 1.51690 \end{aligned}\right\} \quad \text{for borosilicate glass}$$

$$\left.\begin{aligned} n(\lambda_1) &= 1.45640 \\ n(\lambda_2) &= 1.46318 \end{aligned}\right\} \quad \text{for vitreous quartz}$$

where $\lambda_1 = 0.6563$ μm and $\lambda_2 = 0.4861$ μm.
(a) Calculate the values of A and B.
(b) Using the Cauchy formula, calculate the refractive index at 0.5890 and 0.3988 μm and compare with the corresponding experimental values:
(i) 1.51124 and 1.52546 for borosilicate glass
(ii) 1.45845 and 1.47030 for vitreous quartz
[**Ans: (a)** For borosilicate glass $A = 1.499$, $B \approx 4.22 \times 10^{-15}$ m^2 giving $n = 1.51120$ at $\lambda = 0.5890$ μm, and $n = 1.52557$ at $\lambda = 0.3988$ μm; **(b)** for vitreous quartz $A = 1.44817$, $B \approx 3.546 \times 10^{-15}$ m^2]

10.6 The refractive index variation for pure silica in the wavelength region $0.5\ \mu\text{m} < \lambda_0 < 1.6\ \mu\text{m}$ is accurately described by the empirical formula

$$n(\lambda_0) = C_0 + C_1\lambda_0^2 + C_2\lambda_0^4 + \frac{C_3}{\lambda_0^2 - l} + \frac{C_4}{(\lambda_0^2 - l)^2} + \frac{C_5}{(\lambda_0^2 - l)^3}$$

where $C_0 = 1.4508554$, $C_1 = -0.0031268$, $C_2 = -0.0000381$, $C_3 = 0.0030270$, $C_4 = -0.0000779$, $C_5 = 0.0000018$, $l = 0.035$, and λ_0 is measured in μm. Develop a simple program to calculate and plot $n(\lambda_0)$ and $d^2n/d\lambda_0^2$ in the wavelength domain $0.5 < \lambda_0 < 1.6$ μm, and compare with the results given in Table 10.1.

10.7 **(a)** For a Gaussian pulse given by

$$E = E_0 e^{-t^2/\tau_0^2} e^{i\omega_0 t}$$

the spectral width is approximately given by

$$\Delta\omega \approx \frac{1}{\tau_0}$$

Assume $\lambda_0 = 8000$ Å.
Calculate $\Delta\omega/\omega_0$ for $\tau_0 = 1$ ns and for $\tau_0 = 1$ ps.
(b) For such a Gaussian pulse, the pulse broadening is given by $\Delta\tau = 2z/\tau_0|\gamma|$, where $\gamma = d^2k/d\omega^2$. Using Table 10.1, calculate $\Delta\tau$ and interpret the result physically.
[**Ans: (a)** $\frac{\Delta\omega}{\omega_0} \approx 4 \times 10^{-7}$ and 4×10^{-4};
(b) $\gamma \approx 3.62 \times 10^{-26}$ m^{-1} s^2; $\Delta\tau \approx 0.072$ and ≈ 72 ps km^{-1} for $\tau_0 = 1$ ns and 1 ps, respectively]

10.8 As a Gaussian pulse propagates, the frequency chirp is given by

$$\Delta\omega = \frac{2p}{\tau_0^2(1+p^2)}\left(t - \frac{z}{v_g}\right)$$

where p is defined in Eq. (50). Assume a 100 ps (= τ_0) pulse at $\lambda_0 = 1$ μm. Calculate the frequency chirp $\Delta\omega/\omega_0$ at $t - z/v_g = -100$, -50, $+50$, and $+100$ ps. Assume $z = 1$ km and other values from Table 10.1.

[**Ans:** $\frac{\Delta\omega}{\omega_0} \approx -4.5 \times 10^{-8}, -2.25 \times 10^{-8}, +2.25 \times 10^{-8}$, and $+4.5 \times 10^{-8}$ at $t - z/v_g = -100, -50, +50$ and $+100$ ps, respectively.

10.9 Repeat Prob. 10.8 for $\lambda_0 = 1.5$ μm; the values of τ_0 and z remain the same. Show that the qualitative difference in the results obtained in the previous and in the present problem is the fact that at $\lambda = 1$ μm we have negative dispersion and the front end is red-shifted ($\Delta\omega$ is negative) and the trailing end is blue-shifted. The converse is true at $\lambda = 1.5$ μm where we have positive dispersion.

10.10 The frequency spectrum of $E(0, t)$ is given by the function $A(\omega)$. Show that the frequency spectrum of $E(z, t)$ is simply

$$A(\omega)e^{-ik(\omega)z}$$

implying that no new frequencies are generated—different frequencies superpose with different phases at different values of z.

10.11 The time evolution of a Gaussian pulse in a dispersive medium is given by

$$E(z, t) = \frac{E_0}{\sqrt{1+ip}}\, e^{i(\omega_0 t - k_0 z)} \exp\left[-\frac{\left(t - z/v_g\right)^2}{\tau_0^2(1+ip)}\right]$$

where $p \equiv 2\gamma z/\tau_0^2$. Calculate explicitly the frequency spectrum of $E(0, t)$ and $E(z, t)$, and show that the results agree with those of Prob. 10.10.

REFERENCES AND SUGGESTED READINGS

1. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley Publishing Co., Reading, Mass., 1964.
2. U. C. Paek, G. E. Peterson, and A. Carnevale, "Dispersionless Single Mode Light Guides with a Index Profiles," *Bell System Technical Journal*, Vol. 60, p. 583, 1981.
3. A. Ghatak and K. Thyagarajan, *Introduction to Fiber Optics*, Cambridge University Press, Cambridge, 1998. Reprinted in India by Foundation Books, New Delhi.

Chapter Eleven

WAVE PROPAGATION AND THE WAVE EQUATION

If you are dropping pebbles into a pond and do not watch the spreading rings, your occupation should be considered as useless', said the fictional Russian philosopher, Kuzma Prutkoff. And, indeed we can learn much by observing these graceful circles spreading out from the punctured surface of calm water.

—Gamow and Cleveland

11.1 INTRODUCTION

In this chapter we will discuss the phenomenon of waves. A wave is propagation of a disturbance. For example, when we drop a small stone in a calm pool of water, a circular pattern spreads out from the point of impact. The impact of the stone creates a disturbance which propagates outward. In this propagation, the water molecules do not move outward with the wave; instead they move in nearly circular orbits about an equilibrium position. Once the disturbance has passed a certain region, every drop of water is left at its original position. This fact can easily be verified by placing a small piece of wood on the surface of water. As the wave passes, the piece of wood makes oscillations, and once the disturbance has passed, the wood comes back to its original position. Further, with time the circular ripples spread out; i.e., the disturbance (which is confined to a particular region at a given time) produces a similar disturbance at a neighboring point at a slightly later time with the pattern of disturbance roughly remaining the same. Such propagation of disturbances (without any translation of the medium in the direction of propagation) is termed a *wave*. It is also seen that the wave carries energy; in this case the energy is in the form of kinetic energy of water molecules.

We will first consider the simplest example of wave propagation, i.e., the propagation of a transverse wave on a string. Consider yourself holding one end of a string, with the other end being held tightly by another person so that the string does not sag. If you move the end of the string up and down a few times, then a disturbance is created which propagates toward the other end of the string. Thus, if we take a snapshot of the string at $t = 0$ and at a slightly later time Δt, then the snapshots will roughly[1] look like the ones shown in Fig. 11.1(a) and (b). The figure shows that the disturbances have identical shapes except for the fact that one is displaced from the other by distance $v\Delta t$, where v represents the speed of the disturbance. Such a propagation of a disturbance without its change in form is a characteristic of a wave. The following points may, however, be noted:

1. A certain amount of work is done when the wave is generated, and as the wave propagates through the string, it carries with it a certain amount of energy which is felt by the person holding the other end of the string.
2. The wave is transverse; i.e., the displacement of the particles of the string is at right angles to the direction of propagation of the wave.

Referring to Fig. 11.1(a) and (b), we note that the shape of the string at the instant Δt is similar to its shape at $t = 0$, except for the fact that the whole disturbance has traveled through a certain distance. If v represents the speed of the wave, then this distance is simply $v\Delta t$. Consequently, if the equation describing the rope at $t = 0$ is $y(x)$, then at a later instant t, the equation of the curve is $y(x - vt)$, which simply implies a shift of the origin by a distance vt. Similarly, for a disturbance propagating in the $-x$ direction, if the equation describing the rope at $t = 0$ is $y(x)$, then at a later instant t the equation of the curve is $y(x + vt)$.

[1] We are assuming here that as the disturbance propagates through the string, there is negligible attenuation and also no change in the shape of the disturbance.

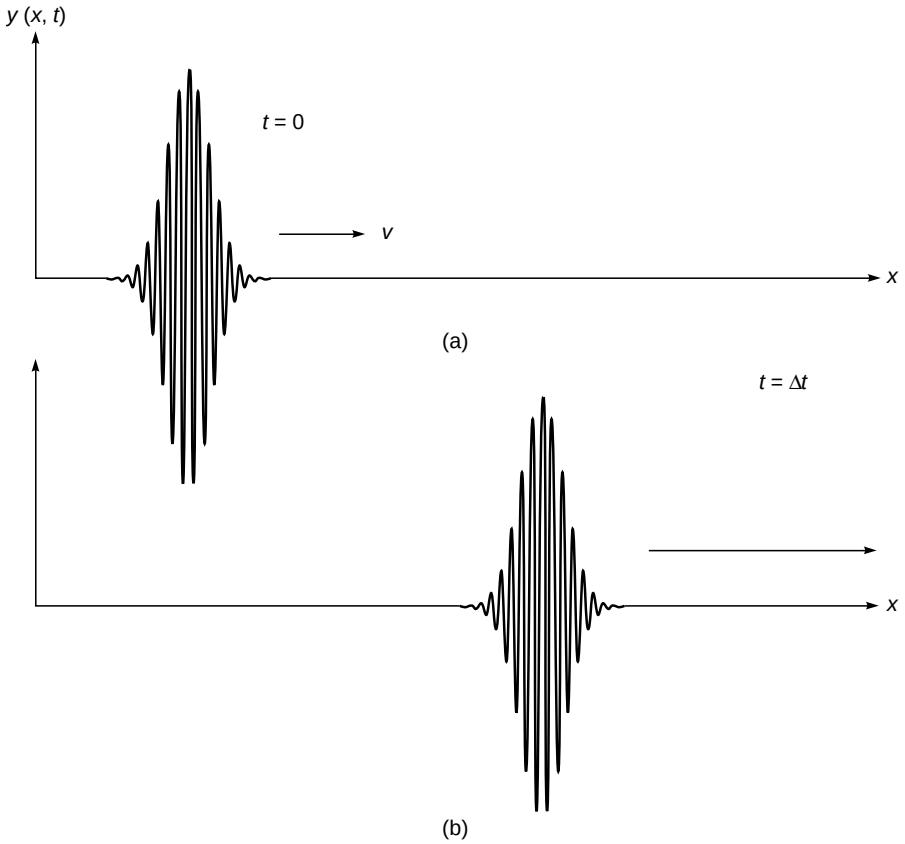


Fig. 11.1 A transverse wave is propagating along the $+x$ axis on a string. (a) and (b) Displacements at $t = 0$ and $t = \Delta t$, respectively.

Example 11.1 Study the propagation of a semicircular pulse in the $+x$ direction whose displacement at $t = 0$ is given by the following equations:

$$y(x, t=0) = \begin{cases} (R^2 - x^2)^{1/2} & |x| \le R \\ 0 & |x| \ge R \end{cases} \quad \textbf{(1)}$$

Solution: For a wave propagating in the $+x$ direction the dependence of $y(x,t)$ on x and t should be through the function $x - vt$. Consequently,

$$y(x, t) = \begin{cases} [R^2 - (x - vt)^2]^{1/2} & |x - vt| \le R \\ 0 & |x - vt| \ge R \end{cases} \quad \textbf{(2)}$$

The shape of the pulse at $t = 0$ and at a later time t_0 is shown in Fig. 11.2. Equation (2) immediately follows from the fact that $y(x, t)$ has to be of the form $y(x - vt)$ and at $t = 0$, $y(x, t)$ must be given by Eq. (1).

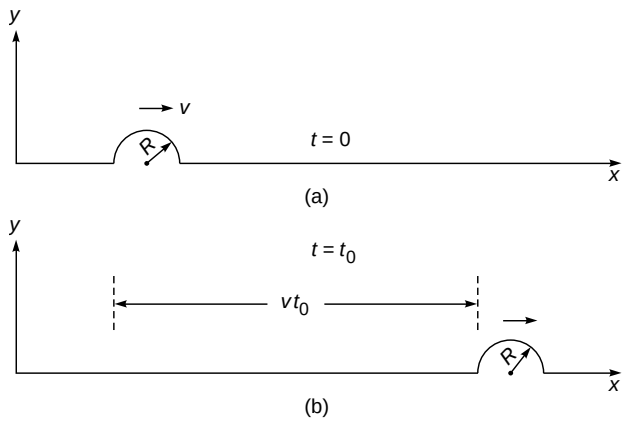


Fig. 11.2 The propagation of a semicircular pulse along the $+x$ axis. (a) and (b) Shape of the pulse at $t = 0$ and at a later time t_0, respectively.

Example 11.2 Consider a pulse propagating in the $-x$ direction with speed v. The shape of the pulse at $t = t_0$ is given by

$$y(x, t = t_0) = \frac{b^2}{a^2 + (x - x_0)^2} \quad \textbf{(3)}$$

(Such a pulse is known as a Lorentzian pulse.) Determine the shape of the pulse at an arbitrary time t.

Solution: The shape of the pulse at $t = t_0$ is shown in Fig. 11.3(a). The maximum of the displacement occurs at $x = x_0$. Since the pulse is propagating in the $-x$ direction, at a later time t, the maximum will

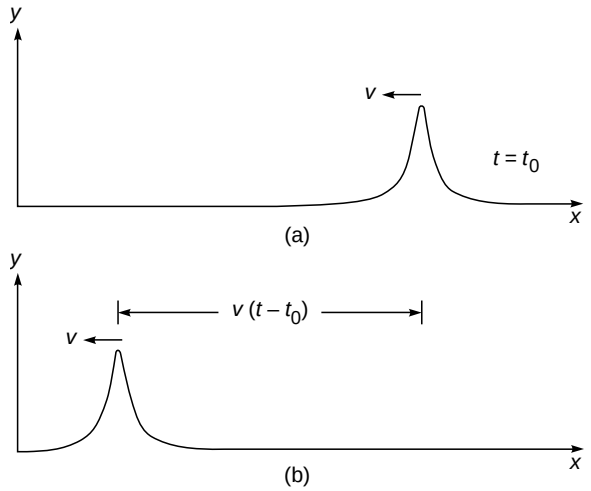


Fig. 11.3 The propagation of a Lorentzian pulse along the minus x axis; (a) and (b) show the shape of the pulse at $t = t_0$ and at a later instant t respectively.

occur at $x_0 - v(t - t_0)$. Consequently, the shape of the pulse at an arbitrary time t is given by

$$y(x, t) = \frac{b^2}{a^2 + [x - x_0 + v(t - t_0)]^2} \qquad \textbf{(4)}$$

Equation (4) could have been written directly from Eq. (3) by replacing x by $x + v(t - t_0)$.

11.2 SINUSOIDAL WAVES: CONCEPT OF FREQUENCY AND WAVELENGTH

Until now we have been considering the propagation of a pulse which lasts for a finite amount of time. We now consider a periodic wave in which the displacement $y(x, t)$ has the form

$$y(x, t) = a \cos [k(x \mp vt) + \phi] \qquad \textbf{(5)}$$

where the upper and lower signs correspond to waves propagating in the $+x$ and $-x$ directions, respectively. Such a displacement is indeed produced in a long stretched string at the end of which a continuously vibrating tuning fork is placed. The quantity ϕ is known as the phase of the wave (see Chap. 7). We may, without loss of generality, assume $\phi = 0$. Thus for a wave propagating in the $+x$ direction,

$$y(x, t) = a \cos k(x - vt) \qquad \textbf{(6)}$$

In Fig. 11.4 we have plotted the dependence of the displacement y on x at $t = 0$ and at $t = \Delta t$. These are given by

$$y(x) = a \cos kx \qquad \text{at } t = 0 \qquad \textbf{(7a)}$$

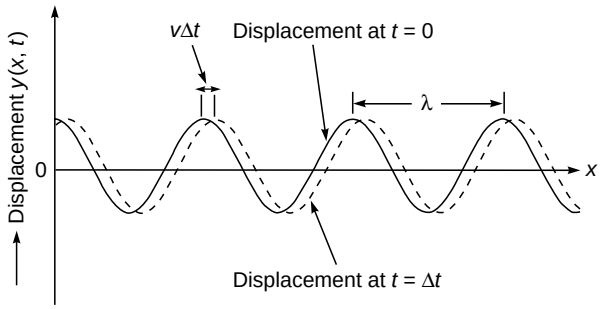


Fig. 11.4 The curves represent the displacement of a string at $t = 0$ and $t = \Delta t$, respectively, when a sinusoidal wave is propagating in the $+x$ direction.

and

$$y(x) = a \cos k(x - v\Delta t) \qquad \text{at } t = \Delta t \qquad \textbf{(7b)}$$

The two curves are the snapshots of the string at the two instants. It can be seen from the figure that, at a particular instant, any two points separated by a distance

$$\lambda = \frac{2\pi}{k} \qquad \textbf{(8)}$$

have identical displacements. This distance is known as the wavelength. Further, the displaced curve (which corresponds to the instant $t = \Delta t$) can be obtained by displacing the curve corresponding to $t = 0$ by a distance $v\Delta t$; this shows that the wave is propagating in the $+x$ direction with speed v. It can also be seen that the maximum displacement of the particle (from its equilibrium position) is a, which is known as the amplitude of the wave.

In Fig. 11.5 we have plotted the time dependence of the displacement of the points characterized by $x = 0$ and $x = \Delta x$. These are given by

$$\begin{aligned} y(t) &= a \cos \omega t \qquad && \text{at } x = 0 \\ y(t) &= a \cos (\omega t - k\Delta x) \qquad && \text{at } x = \Delta x \end{aligned} \qquad \textbf{(9)}$$

where

$$\omega = kv \qquad \textbf{(10)}$$

The curves correspond to the time variation of the displacement of the two points. Corresponding to a particular point, the displacement repeats itself after a time

$$T = 2\pi/\omega \qquad \textbf{(11)}$$

which is known as the time period of the wave. The quantity

$$\nu = \frac{1}{T} \qquad \textbf{(12)}$$

is known as the frequency of the wave and represents the number of oscillations that a particle carries out in 1 s. It can

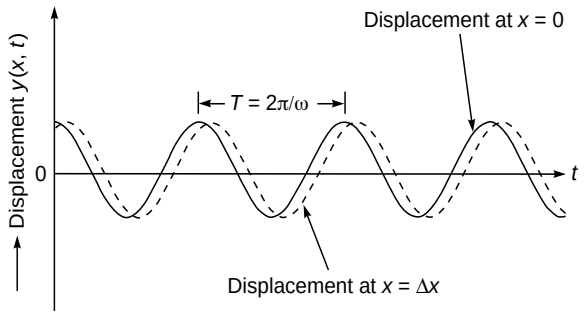


Fig. 11.5 The curves represent the time variation of the displacement of a string at $x = 0$ and $x = \Delta x$, respectively, when a sinusoidal wave is propagating in the $+x$ direction.

be seen from the two curves in Fig. 11.5 that the two points $x = 0$ and $x = \Delta x$ execute exactly similar vibrations except for a phase difference of $k\Delta x$. In fact any two points on the string execute simple harmonic motions with the same amplitude and same frequency but with a phase difference of kx_0, where x_0 represents the distance between the two points. Clearly if this distance is a multiple of the wavelength, i.e.,

$$x_0 = m\lambda \qquad m = 1, 2, \ldots$$

then

$$kx_0 = \frac{2\pi}{\lambda} m\lambda = 2m\pi$$

which implies that two points separated by a distance that is a multiple of the wavelength vibrate with the same phase. Similarly, two points separated by a distance $\frac{1}{2}\lambda, \frac{3}{2}\lambda, \ldots$ vibrate in opposite phase. In general, a path difference of x_0 corresponds to a phase difference of $(2\pi/\lambda)x_0$.

Using Eqs. (10) to (12), we get

$$\nu = \frac{1}{T} = \frac{\omega}{2\pi} = \frac{kv}{2\pi} = \frac{v}{\lambda}$$

or
$$v = \nu\lambda \qquad \textbf{(13)}$$

Notice the similarity in the variation of the displacement with respect to x (at a given value of time) and with respect to t (at a given value of x); see Figs. 11.4 and 11.5. The similarity can be expressed by writing Eq. (6) in the form

$$y(x, t) = a \cos\left(\frac{2\pi}{\lambda}x - \frac{2\pi}{T}t\right) \qquad \textbf{(14)}$$

which shows that the wavelength λ in Fig. 11.4 plays the same role as the time period T in Fig. 11.5. Equation (14) is often written in the form

$$y(x,t) = a\cos(kx - \omega t) \qquad \textbf{(15)}$$

Note that the entire discussion given above would remain valid for an arbitrary value of the phase factor ϕ.

11.3 TYPES OF WAVES

As mentioned earlier, when a wave is propagating through a string, the displacement is at right angles to the direction of propagation. Such a wave is known as a transverse wave.[2] Similarly, when a sound wave propagates through air, the displacement of the air molecules is along the direction of propagation of the wave; such waves are known as longitudinal waves. However, there are waves which are neither longitudinal nor transverse in character; for example, when a wave propagates through the surface of water, the water molecules move approximately in circular orbits.

11.4 ENERGY TRANSPORT IN WAVE MOTION

A wave carries energy; for example, when a transverse wave propagates through a string, the particles execute simple harmonic motions about their equilibrium positions, and associated with this motion is a certain amount of energy. As the wave propagates through, the energy gets transported from one end of the string to the other. We consider the time variation of the displacement of a particle, which can be written as

$$y = a \cos(\omega t + \phi) \qquad \textbf{(16)}$$

The instantaneous velocity of the particle is

$$v = \frac{dy}{dt} = -a\omega \sin(\omega t + \phi) \qquad \textbf{(17)}$$

Thus, the kinetic energy T is given by

$$T = \frac{1}{2} m \left(\frac{dy}{dt}\right)^2$$

$$= \frac{1}{2} ma^2\omega^2 \sin^2(\omega t + \phi) \qquad \textbf{(18)}$$

The total energy E is the maximum value of T

$$E = (T)_{\max}$$

$$= \frac{1}{2} ma^2\omega^2 [\sin^2(\omega t + \phi)]_{\max}$$

$$= \frac{1}{2} ma^2\omega^2 \qquad \textbf{(19)}$$

[2] Electromagnetic waves are also transverse in character. However, the electromagnetic waves have also a longitudinal component near the source which dies off rapidly at large distances (see Sec. 23.4).

For a sound wave propagating through a gas, the energy per unit volume ε is given by

$$\varepsilon = \frac{1}{2} mna^2\omega^2$$
$$= \frac{1}{2} \rho a^2\omega^2$$
$$= 2\pi^2\rho a^2\nu^2 \quad \textbf{(20)}$$

where m represents the mass of gas molecules, n represents the number of molecules per unit volume, and ρ (= nm) is the density of the gas. With such a wave, we can associate the intensity which is defined as the energy flow per unit time across a unit area perpendicular to the direction of propagation. Since the speed of propagation of the wave is v, the intensity I is given by[3]

$$I = 2\pi^2\rho v a^2\nu^2 \quad \textbf{(21)}$$

Thus the intensity is proportional to the square of the amplitude and square of the frequency.

Let us consider a wave emanating from a point source in a uniform isotropic[4] medium. Let W represent the power of the source and we assume that there is no absorption. We consider a sphere of radius r whose center is at the point source. Clearly, W (measured in Joules per second) will cross the spherical surface whose area is $4\pi r^2$. Thus, the intensity I will be given by

$$I = \frac{W}{4\pi r^2} \quad \textbf{(22)}$$

which is nothing but the inverse square law. Using Eqs. (21) and (22), we obtain

$$\frac{W}{4\pi r^2} = 2\pi^2\rho v a^2\nu^2$$

or

$$a = \left(\frac{W}{8\pi^3\rho v \nu^2}\right)^{1/2} \frac{1}{r} \quad \textbf{(23)}$$

showing that the amplitude falls off as $1/r$. Indeed, for a spherical wave[5] emanating from a point source, the displacement is given by

$$f = \frac{a_0}{r} \sin(kr - \omega t)$$

where a_0 represents the amplitude of the wave at unit distance from the source.

Example 11.3 A source of sound is vibrating with a frequency of 256 vibrations per second in air and propagating energy uniformly in all directions at the rate of 5 J s^{-1}. Calculate the intensity and the amplitude of the wave at a distance of 25 m from the source. Assume that there is no absorption (speed of sound waves in air = 330 m s^{-1}; density of air = 1.29 kg m^{-3}).

Solution:

$$\text{Intensity } I = \frac{5 \text{ J/s}}{4\pi \times (25)^2 \text{ m}^2}$$
$$\approx 6.4 \times 10^{-4} \text{ J s}^{-1} \text{ m}^{-2}$$

$$\text{Thus} \qquad a = \left(\frac{5}{8\pi^3 \times 1.29 \times 330 \times 256 \times 256}\right)^{1/2} \frac{1}{25}$$
$$\approx 1.0 \times 10^{-6} \text{ m}$$

Example 11.4 Show that when a transverse wave propagates through a string, the energy transmitted per unit time is $\frac{1}{2}\rho\omega^2 a^2 v$, where ρ is the mass per unit length, a the amplitude of the wave, and v is the speed of propagation of the wave.

Solution: The energy associated per unit length of the string is $\frac{1}{2}\rho\omega^2 a^2$; since the speed of the wave is v, the result follows.

11.5 THE ONE-DIMENSIONAL WAVE EQUATION

In Sec. 11.1 we showed that the displacement ψ of a one-dimensional wave is always of the form

$$\psi = f(x - vt) + g(x + vt) \quad \textbf{(24)}$$

where the first term on the RHS of the above equation represents a disturbance propagating in the $+x$ direction with speed v and similarly the second term represents a disturbance propagating in the $-x$ direction with speed v. The questions now arise as to how we can predict the existence of waves and what would be the velocity of propagation of these waves. The answer is as follows: If we can derive an equation of the form

$$\frac{\partial^2\psi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2\psi}{\partial t^2} \quad \textbf{(25)}$$

[3] This can be easily understood from the fact that if we have N particles per unit volume, each moving with the same velocity v, then the number of particles crossing a unit area (normal to v) per unit time is Nv.

[4] Isotropic media are the ones in which physical properties (such as velocity of propagation of a particular wave) are the same in all directions. In Chap. 22 we will consider anisotropic media.

[5] When waves emanate from a point source in an isotropic medium, all the points on the surface of a sphere (whose center is at the point source) have the same amplitude and the same phase; in other words, the locus of points which have the same amplitude and the same phase is a sphere. Such waves are known as spherical waves. Far away from the source, over a small area, the spherical waves are essentially plane waves.

from physical considerations, then we can be sure that waves will result and ψ will represent the displacement associated with the wave. This follows from the fact that the general solution of Eq. (25) is of the form

$$\psi = f(x - vt) + g(x + vt) \tag{26}$$

where f and g are arbitrary functions of their argument (see Sec. 11.9). Consequently, if we ever obtain an equation of the form of Eq. (25) from physical considerations, we can predict the existence of waves, the speed of which would be v.

The simplest particular solutions of the wave equation correspond to sinusoidal variation:

$$\psi = A \sin [k(x \pm vt) + \phi] \tag{27}$$

or

$$\psi = A \cos [k(x \pm vt) + \phi] \tag{28}$$

As shown in Sec. 11.2,

$$k = \frac{2\pi}{\lambda} \qquad \text{and} \qquad kv = \omega = 2\pi\nu \tag{29}$$

where λ is the wavelength and ν the frequency of the wave. Instead of sinusoidal variation it is often more convenient to write the solution in the form

$$\psi = A \exp [i(kx \pm \omega t + \phi)] \tag{30}$$

where, as before, A and ϕ represent the amplitude and initial phase of the wave, respectively. In writing Eq. (30), it is implied that the actual displacement is just the real part of ψ which is

$$A \cos (kx \pm \omega t + \phi)$$

In the next three sections we will derive the wave equation for some simple cases.[6] In Sec. 11.9 we will discuss the general solution of the wave equation.

11.6 TRANSVERSE VIBRATIONS OF A STRETCHED STRING

Let us consider a stretched string having a tension T. In its equilibrium position the string is assumed to lie on the x axis. If the string is pulled in the y direction, then forces will act on the string which will tend to bring it back to its equilibrium position. Let us consider a small length AB of the string and calculate the net force acting on it in the y direction. Due to the tension T, the endpoints A and B experience force in the direction of the arrows shown in Fig. 11.6. The force at A in the upward direction is

$$-T \sin \theta_1 \approx -T \tan \theta_1 = -T \left.\frac{\partial y}{\partial x}\right|_x \tag{31}$$

Similarly, the force at B in the upward direction is

$$T \sin \theta_2 \approx T \tan \theta_2 = T \left.\frac{\partial y}{\partial x}\right|_{x+dx} \tag{32}$$

where we have assumed θ_1 and θ_2 to be small. Thus the net force acting on AB in the y direction is

$$T\left[\left(\frac{\partial y}{\partial x}\right)_{x+dx} - \left(\frac{\partial y}{\partial x}\right)_x\right] = T\frac{\partial^2 y}{\partial x^2}dx \tag{33}$$

where we have used the Taylor series expansion of $(\partial y/\partial x)_{x+dx}$ about the point x

$$\left(\frac{\partial y}{\partial x}\right)_{x+dx} = \left(\frac{\partial y}{\partial x}\right)_x + \left.\frac{\partial}{\partial x}\left(\frac{\partial y}{\partial x}\right)\right|_x dx$$

and have neglected higher-order terms because dx is infinitesimal. The equation of motion is therefore

$$\Delta m \frac{\partial^2 y}{\partial t^2} = T\frac{\partial^2 y}{\partial x^2}dx \tag{34}$$

where Δm is the mass of element AB. If ρ is the mass per unit length, then

$$\Delta m = \rho\, dx$$

and we get

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{T/\rho}\frac{\partial^2 y}{\partial t^2} \tag{35}$$

which is the one-dimensional wave equation. Thus we may conclude that transverse waves can propagate through a stretched string, and if we compare the above equation with Eq. (25), we obtain the following expression for the speed of the transverse waves:

$$v = \sqrt{\frac{T}{\rho}} \tag{36}$$

[6] In Chap. 23 we will derive the wave equation from Maxwell's equations and thereby obtain an expression for the speed of electromagnetic waves.

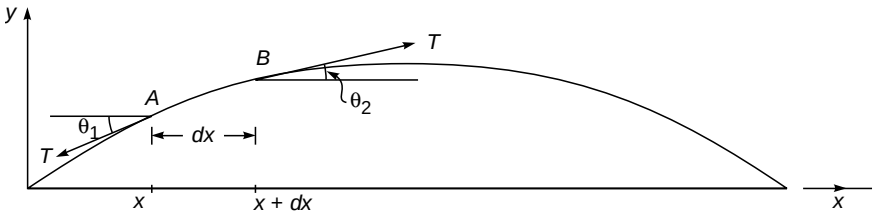


Fig. 11.6 Transverse vibrations of a stretched string.

The vibrations of a clamped string will be discussed in Sec. 13.2. In an actual string, the displacement is not rigorously of the form given by Eq. (24); this is a consequence of the various approximations made in the derivation of the wave equation. There is, in general, an attenuation of the wave, and also the shape does not remain unaltered.

11.7 LONGITUDINAL SOUND WAVES IN A SOLID

In this section we will derive an expression for the velocity of longitudinal sound waves propagating in an elastic solid. Let us consider a solid cylindrical rod of cross-sectional area A. Let PQ and RS be two transverse sections of the rod at distances x and $x + \Delta x$, respectively, from a fixed point O, where we have chosen the x axis to be along the length of the rod (see Fig. 11.7).

Let the longitudinal displacement of a plane be denoted by $\xi(x)$. Thus the displacements of the planes PQ and RS are $\xi(x)$ and $\xi(x + \Delta x)$, respectively. In the displaced position, the distance between the planes $P'Q'$ and $R'S'$ is

$$\xi(x + \Delta x) - \xi(x) + \Delta x = \xi(x) + \frac{\partial \xi}{\partial x}\Delta x - \xi(x) + \Delta x$$

$$= \Delta x + \frac{\partial \xi}{\partial x}\Delta x$$

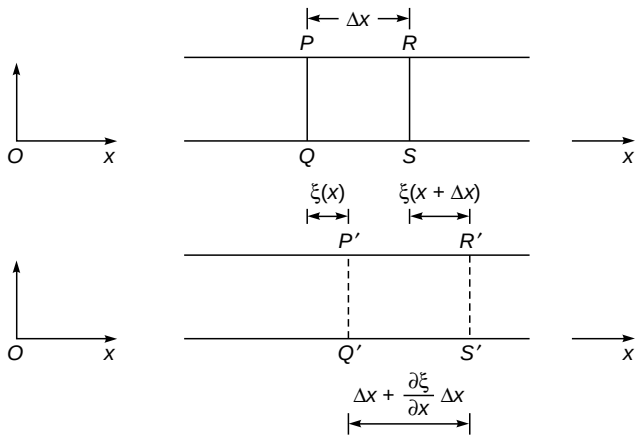


Fig. 11.7 Propagation of longitudinal sound waves through a cylindrical rod.

The elongation of the element is $(\partial \xi / \partial x)\Delta x$, and therefore, the longitudinal strain is

$$\frac{\text{Increase in length}}{\text{Original length}} = \frac{(\partial \xi / \partial x)\Delta x}{\Delta x} = \frac{\partial \xi}{\partial x} \quad \textbf{(37)}$$

Since Young's modulus Y is defined as the ratio of the longitudinal stress to the longitudinal strain, we have

$$\text{Longitudinal stress} = \frac{F}{A} = Y \times \text{strain}$$

$$= Y \frac{\partial \xi}{\partial x} \quad \textbf{(38)}$$

where F is the force acting on the element $P'Q'$. Thus

$$F(x) = YA \frac{\partial \xi}{\partial x} \quad \textbf{(39)}$$

and, therefore,

$$\frac{\partial F}{\partial x} = YA \frac{\partial^2 \xi}{\partial x^2} \quad \textbf{(40)}$$

Now, if we consider the volume $P'Q'S'R'$, then a force F is acting on the element $P'Q'$ *in the negative* x direction, and a force $F(x + \Delta x)$ is acting on the plane $R'S'$ *along the positive* x direction. Thus the resultant force acting on the element $P'Q'S'R'$ will be

$$F(x + \Delta x) - F(x) = \frac{\partial F}{\partial x}\Delta x$$

$$= YA \frac{\partial^2 \xi}{\partial x^2}\Delta x \quad \textbf{(41)}$$

If ρ represents the density, then the mass of the element is $\rho A \Delta x$. Thus the equation of motion will be

$$\rho A \Delta x \frac{\partial^2 \xi}{\partial t^2} = YA \Delta x \frac{\partial^2 \xi}{\partial x^2}$$

or

$$\frac{\partial^2 \xi}{\partial x^2} = \frac{1}{v_l^2}\frac{\partial^2 \xi}{\partial t^2} \quad \textbf{(42)}$$

where

$$v_l = \left(\frac{Y}{\rho}\right)^{1/2} \quad \textbf{(43)}$$

represents the velocity of the waves and the subscript l refers to the fact that we are considering longitudinal waves.[7]

The above derivation is valid when the transverse dimension of the rod is small compared with the wavelength of the disturbance so that one may assume that the longitudinal displacement at all points on any transverse section (such as PQ) is the same. In general, if one carries out a rigorous analysis of the vibrations of an extended isotropic elastic solid, one can show that the velocities of the longitudinal and transverse waves are given by[8]

$$v_l = \left[\frac{Y}{\rho}\frac{1-\sigma}{(1+\sigma)(1-2\sigma)}\right]^{1/2} = \left(\frac{K+(4/3)\eta}{\rho}\right)^{1/2} \qquad \textbf{(44)}$$

$$v_t = \left[\frac{Y}{\rho}\frac{1}{2(1+\sigma)}\right]^{1/2} = \left(\frac{\eta}{\rho}\right)^{1/2} \qquad \textbf{(45)}$$

where σ, η, and K represent the Poisson ratio, modulus of rigidity, and bulk modulus, respectively. In this case, the transverse wave [whose velocity is given by Eq. (45)] is due to the restoring forces arising because of the elastic properties of the material, whereas corresponding to the transverse waves discussed in Sec. 11.6, the string moved as a whole and the restoring force was due to the externally applied tension.

11.8 LONGITUDINAL WAVES IN A GAS

To determine the speed of propagation of longitudinal sound waves in a gas, we consider a column $PQSR$ as shown in Fig. 11.8(a). Once again, because of a longitudinal displacement, the plane PQ gets displaced by $\xi(x)$ and the plane RS gets displaced by a distance $\xi(x + \Delta x)$ (see Fig. 11.8). Let the pressure of the gas in the absence of any disturbance be P_0. Let $P_0 + \Delta P(x)$ and $P_0 + \Delta P(x + \Delta x)$ denote the pressures at the planes $P'Q'$ and $R'S'$, respectively. Now, if we consider the column $P'Q'S'R'$, then the pressure $P_0 + \Delta P(x)$ on the face $P'Q'$ acts in the $+x$ direction whereas the pressure $P_0 + \Delta P(x + \Delta x)$ on the face $R'S'$ acts in the $-x$ direction. Thus the force acting on the column $P'Q'S'R'$ is

$$[\Delta P(x) - \Delta P(x+\Delta x)]A = -\frac{\partial}{\partial x}(\Delta P)\Delta x A \qquad \textbf{(46)}$$

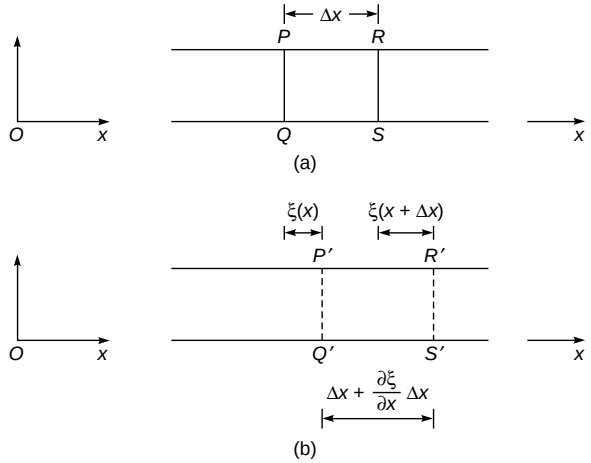


Fig. 11.8 Propagation of longitudinal sound waves through air.

where A represents the cross-sectional area. Consequently, the equation of motion for the column $P'Q'S'R'$ is

$$-\frac{\partial}{\partial x}(\Delta P)A\Delta x = \rho A\,\Delta x\frac{\partial^2\xi}{\partial t^2}$$

where ρ represents the density of the gas. Thus

$$-\frac{\partial}{\partial x}(\Delta P) = \rho\frac{\partial^2\xi}{\partial t^2} \qquad \textbf{(47)}$$

Now, a change in pressure gives rise to a change in volume, and if the frequency of the wave is large ($\gtrsim 20$ Hz), the pressure fluctuations will be rapid and one may assume the process to be adiabatic. Thus, we may write

$$PV^{\gamma} = \text{constant} \qquad \textbf{(48)}$$

where $\gamma = C_p/C_v$ represents the ratio of the two specific heats. If we differentiate the above expression, we get

$$\Delta P\,V^{\gamma} + \gamma V^{\gamma-1}P\Delta V = 0$$

$$\Delta P = -\frac{\gamma P}{V}\Delta V \qquad \textbf{(49)}$$

The change in the length of the column $PQSR$ is

$$[\xi(x+\Delta x) - \xi(x) + \Delta x] - \Delta x = \frac{\partial\xi}{\partial x}\Delta x$$

Thus, the change in the volume is

$$\Delta V = \frac{\partial\xi}{\partial x}A\Delta x$$

[7] In a similar manner one can consider transverse waves propagating through an elastic solid, the velocity of which is given by [see, for example, Ref. 8]

$$v_t = \sqrt{\eta/\rho}$$

where η represents the modulus of rigidity.

[8] See, for example, Ref. 5.

The original volume V of the element is $A\Delta x$. Thus

$$\Delta P = -\frac{\gamma P}{A\Delta x}\frac{\partial \xi}{\partial x}A\Delta x$$

$$= -\gamma P\frac{\partial \xi}{\partial x} \quad \textbf{(50)}$$

or $$\frac{\partial}{\partial x}(\Delta P) = -\gamma P\frac{\partial^2 \xi}{\partial x^2} \quad \textbf{(51)}$$

Using Eqs. (47) and (51), we obtain

$$\frac{\partial^2 \xi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 \xi}{\partial t^2} \quad \textbf{(52)}$$

where

$$v = \left(\frac{\gamma P}{\rho}\right)^{1/2} \quad \textbf{(53)}$$

represents the velocity of propagation of longitudinal sound waves in a gas. For air, if we assume $\gamma = 1.40$, $P = 1.01 \times 10^5$ Nm^{-2} and $\rho = 1.3 \times 10^{-3}$ kg m^{-3}, then we obtain

$$v \approx 330 \text{ m s}^{-1}$$

The adiabatic compressibility of a gas is given by

$$\kappa_S = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_S = \frac{1}{\gamma P} \quad \textbf{(54)}$$

where the subscript S refers to the adiabatic condition (constant entropy). The bulk modulus K of a gas is the inverse of κ_S

$$K = \frac{1}{\kappa_S} = \gamma P \quad \textbf{(55)}$$

and if we substitute this expression for K in Eq. (44), we obtain Eq. (53) where we have used the fact that the modulus of rigidity η for a gas is zero.

11.9 THE GENERAL SOLUTION OF THE ONE-DIMENSIONAL WAVE EQUATION[9]

To obtain a general solution of the equation

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 \psi}{\partial t^2} \quad \textbf{(56)}$$

we introduce two new variables

$$\xi = x - vt \quad \textbf{(57)}$$

$$\eta = x + vt \quad \textbf{(58)}$$

and write Eq. (56) in terms of these variables. Now,

$$\frac{\partial \psi}{\partial x} = \frac{\partial \psi}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial \psi}{\partial \eta}\frac{\partial \eta}{\partial x} \quad \textbf{(59)}$$

or $$\frac{\partial \psi}{\partial x} = \frac{\partial \psi}{\partial \xi} + \frac{\partial \psi}{\partial \eta} \quad \textbf{(60)}$$

where we have used the fact that

$$\frac{\partial \xi}{\partial x} = 1 \quad \text{and} \quad \frac{\partial \eta}{\partial x} = 1$$

Differentiating Eq. (60) with respect to x, we get

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial \psi}{\partial \xi}\right) + \frac{\partial}{\partial x}\left(\frac{\partial \psi}{\partial \eta}\right)$$

$$= \frac{\partial}{\partial \xi}\left(\frac{\partial \psi}{\partial \xi}\right)\frac{\partial \xi}{\partial x} + \frac{\partial}{\partial \eta}\left(\frac{\partial \psi}{\partial \xi}\right)\frac{\partial \eta}{\partial x}$$

$$+ \frac{\partial}{\partial \xi}\left(\frac{\partial \psi}{\partial \eta}\right)\frac{\partial \xi}{\partial x} + \frac{\partial}{\partial \eta}\left(\frac{\partial \psi}{\partial \eta}\right)\frac{\partial \eta}{\partial x}$$

or $$\frac{\partial^2 \psi}{\partial x^2} = \frac{\partial^2 \psi}{\partial \xi^2} + 2\frac{\partial^2 \psi}{\partial \eta \partial \xi} + \frac{\partial^2 \psi}{\partial \eta^2} \quad \textbf{(61)}$$

Similarly

$$\frac{\partial \psi}{\partial t} = \frac{\partial \psi}{\partial \xi}\frac{\partial \xi}{\partial t} + \frac{\partial \psi}{\partial \eta}\frac{\partial \eta}{\partial t}$$

$$= -v\frac{\partial \psi}{\partial \xi} + v\frac{\partial \psi}{\partial \eta}$$

and

$$\frac{\partial^2 \psi}{\partial t^2} = -v\left[\frac{\partial}{\partial \xi}\left(\frac{\partial \psi}{\partial \xi}\right)\frac{\partial \xi}{\partial t} + \frac{\partial}{\partial \eta}\left(\frac{\partial \psi}{\partial \xi}\right)\frac{\partial \eta}{\partial t}\right]$$

$$+ v\left[\frac{\partial}{\partial \xi}\left(\frac{\partial \psi}{\partial \eta}\right)\frac{\partial \xi}{\partial t} + \frac{\partial}{\partial \eta}\left(\frac{\partial \psi}{\partial \eta}\right)\frac{\partial \eta}{\partial t}\right]$$

or $$\frac{\partial^2 \psi}{\partial t^2} = v^2\left(\frac{\partial^2 \psi}{\partial \xi^2} - 2\frac{\partial^2 \psi}{\partial \eta \partial \xi} + \frac{\partial^2 \psi}{\partial \eta^2}\right) \quad \textbf{(62)}$$

Substituting the expressions for $\partial^2\psi/\partial x^2$ and $\partial^2\psi/\partial t^2$ from Eqs. (61) and (62) into Eq. (56), we obtain

$$\frac{\partial^2 \psi}{\partial \xi^2} + 2\frac{\partial^2 \psi}{\partial \eta \partial \xi} + \frac{\partial^2 \psi}{\partial \eta^2} = \frac{\partial^2 \psi}{\partial \xi^2} - 2\frac{\partial^2 \psi}{\partial \eta \partial \xi} + \frac{\partial^2 \psi}{\partial \eta^2}$$

[9] This section may be skipped in the first reading.

or
$$\frac{\partial}{\partial \eta}\left(\frac{\partial \psi}{\partial \xi}\right) = 0 \qquad \textbf{(63)}$$

Thus $\partial\psi/\partial\xi$ has to be independent of η; however, it can be an arbitrary function of ξ:

$$\frac{\partial \psi}{\partial \xi} = F(\xi) \qquad \textbf{(64)}$$

or
$$\psi = \int F(\xi) d\xi + \text{ constant of integration}$$

The constant of integration can be an arbitrary function of η, and since the integral of an arbitrary function is again an arbitrary function, we obtain as the most general solution of the wave equation

$$\begin{aligned}\psi &= f(\xi) + g(\eta) \\ &= f(x - vt) + g(x + vt)\end{aligned} \qquad \textbf{(65)}$$

where f and g are arbitrary functions of their argument. The function $f(x - vt)$ represents a disturbance propagating in the $+x$ direction with speed v, and the function $g(x + vt)$ represents a disturbance propagating in the $-x$ direction.

Example 11.5 Solve the one-dimensional wave equation [Eq. (25)] by the method of separation of variables,[10] and show that the solution can indeed be expressed in the form given by Eqs. (27) and (28).

Solution: In the method of separation of variables, we try a solution of the wave equation

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 \psi}{\partial t^2} \qquad \textbf{(66)}$$

of the form

$$\psi(x, t) = X(x)\, T(t) \qquad \textbf{(67)}$$

where $X(x)$ is a function of x alone and $T(t)$ is a function of t alone. Substituting in Eq. (66), we get

$$T(t)\frac{d^2X}{dx^2} = \frac{1}{v^2}X(x)\frac{\partial^2 T}{dt^2}$$

or

$$\frac{1}{X(x)}\frac{d^2X}{dx^2} = \frac{1}{v^2 T(t)}\frac{d^2T}{dt^2} \qquad \textbf{(68)}$$

Notice that partial derivatives have been replaced by total derivatives.

The LHS is a function of x alone and the RHS is a function of t alone. This implies that a function of one independent variable x is equal to a function of another independent variable t for all values of x and t. This is possible only when each side is equal to a constant; we set this constant equal to $-k^2$. Thus

$$\frac{1}{X(x)}\frac{d^2X}{dx^2} = \frac{1}{v^2}\frac{1}{T(t)}\frac{d^2T}{dt^2} = -k^2 \qquad \textbf{(69)}$$

or
$$\frac{d^2X}{dx^2} + k^2X(x) = 0 \qquad \textbf{(70)}$$

and
$$\frac{d^2T}{dt^2} + \omega^2 T(t) = 0 \qquad \textbf{(71)}$$

where
$$\omega = kv = \frac{2\pi v}{\lambda} \qquad \textbf{(72)}$$

represents the angular frequency of the wave. The solutions of Eqs. (70) and (71) can be easily written as

$$X(x) = A\cos kx + B\sin kx$$

and

$$T(t) = C\cos \omega t + D\sin \omega t$$

Thus

$$\begin{aligned}\psi(x, t) = &(A\cos kx + B\sin kx) \\ &(C\cos \omega t + D\sin \omega t)\end{aligned} \qquad \textbf{(73)}$$

Suitable choice of the constants A, B, C, and D gives

$$\psi(x, t) = a\cos(kx - \omega t + \phi)$$

or
$$\psi(x, t) = a\cos(kx + \omega t + \phi)$$

representing waves propagating in the $+x$ and $-x$ directions, respectively. One can also have

$$\psi(x, t) = a\exp[\pm i(kx \pm \omega t + \phi)]$$

as a solution.

In general, all values of the frequencies are possible, but the frequency and wavelength have to be related through Eq. (72). However, there are systems (such as a string under tension and fixed at both ends) where only certain values of frequencies are possible (see Sec. 8.2).

Example 11.6 Until now we have confined our discussion to waves in one dimension. The three-dimensional wave equation is of the form

$$\nabla^2\psi = \frac{1}{v^2}\frac{\partial^2 \psi}{\partial t^2} \qquad \textbf{(74)}$$

[10] The method of separation of variables is a powerful method for solving certain kinds of partial differential equations. According to this method, the solution is assumed to be a product of functions, each function depending only on one independent variable [see Eq. (67)]. On substituting this solution, if the variables separate out, then the method is said to work and the general solution is a linear sum of all possible solutions; see, e.g., the analysis given in Sec. 8.2. If the variables do not separate out, one has to try some other method to solve the equation.

where $$\nabla^2\psi \equiv \frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2} \quad (75)$$

Solve the three-dimensional wave equation by the method of separation of variables and interpret the solution physically.

Solution: Using the method of separation of variables, we write

$$\psi(x, y, z, t) = X(x)Y(y)Z(z)T(t) \quad (76)$$

where $X(x)$ is a function of x alone, etc. Substituting in Eq. (74), we obtain

$$YZT\frac{d^2X}{dx^2} + XZT\frac{d^2Y}{dy^2} + XYT\frac{d^2Z}{dz^2} = \frac{1}{v^2}XYZ\frac{d^2T}{dt^2}$$

or dividing throughout by ψ

$$\left(\frac{1}{X}\frac{d^2X}{dx^2}\right)+\left(\frac{1}{Y}\frac{d^2Y}{dy^2}\right)+\left(\frac{1}{Z}\frac{d^2Z}{dz^2}\right) = \frac{1}{v^2}\left[\frac{1}{T}\frac{d^2T}{dt^2}\right] \quad (77)$$

Since the first term on the LHS is a function of x alone, the second term is a function of y alone, etc., each term must be set equal to a constant. We write

$$\begin{aligned}\frac{1}{X}\frac{d^2X}{dx^2} &= -k_x^2\\ \frac{1}{Y}\frac{d^2Y}{dy^2} &= -k_y^2\\ \frac{1}{Z}\frac{d^2Z}{dz^2} &= -k_z^2\end{aligned} \quad (78)$$

where k_x^2, k_y^2, and k_z^2 are constants. Thus

$$\frac{1}{v^2}\left(\frac{1}{T}\frac{d^2T}{dt^2}\right) = -(k_x^2 + k_y^2 + k_z^2)$$

or

$$\frac{d^2T}{dt^2} + \omega^2T(t) = 0 \quad (79)$$

where

$$\omega^2 = k^2v^2 \quad (80)$$

and

$$k^2 = k_x^2 + k_y^2 + k_z^2$$

The solutions of Eqs. (78) and (79) could be written in terms of sine and cosine functions; it is more convenient to write them in terms of the exponentials:

$$\begin{aligned}\psi &= A\exp[i(k_x x + k_y y + k_z z \pm \omega t + \phi)]\\ &= A\exp[i(\mathbf{k}\cdot\mathbf{r} \pm \omega t + \phi)]\end{aligned} \quad (81)$$

where the vector $\mathbf{k}$ is defined such that its x, y, and z components are k_x, k_y, and k_z, respectively. One could have also written

$$\psi = A\cos(\mathbf{k}\cdot\mathbf{r} - \omega t + \phi) \quad (82)$$

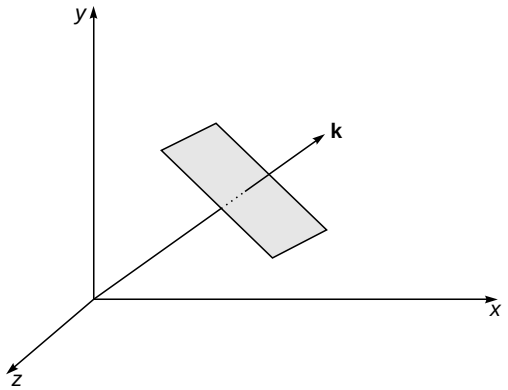


Fig. 11.9 Propagation of a plane wave along the direction $\mathbf{k}$. $\left(k_x = k_y = \frac{k}{\sqrt{2}},\ k_z = 0\right)$

Consider a vector $\mathbf{r}$ which is normal to $\mathbf{k}$; thus $\mathbf{k}\cdot\mathbf{r} = 0$; consequently at a given time the phase of the disturbance is constant on a plane normal to $\mathbf{k}$. The direction of propagation of the disturbance is along $\mathbf{k}$, and the phase fronts are planes normal to $\mathbf{k}$; such waves are known as plane waves (see Fig. 11.9). Notice that for a given value of the frequency, the value of k^2 is fixed [see Eq. (80)]; however, we can have waves propagating in different directions depending on the values of k_x, k_y, and k_z. For example, if

$$k_x = k \quad \text{and} \quad k_y = k_z = 0 \quad (83a)$$

we have a wave propagating along the x axis, and the phase fronts are parallel to the yz plane. Similarly, for

$$k_x = \frac{k}{\sqrt{2}}, \qquad k_y = \frac{k}{\sqrt{2}}, \qquad k_z = 0 \quad (83b)$$

the waves are propagating in a direction which makes equal angles with x and y axes (see Fig. 11.9).

Example 11.7 For a spherical wave, the displacement ψ depends only on r and t, where r is the magnitude of the distance from a fixed point. Obtain a general solution of the wave equation for a spherical wave.

Solution: When ψ depends only on r and t,

$$\nabla^2\psi = \frac{\partial^2\psi}{\partial r^2}+\frac{2}{r}\frac{\partial\psi}{\partial r} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) \quad (84)$$

Thus, the wave equation for a spherical wave simplifies to

$$\nabla^2\psi = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) = \frac{1}{v^2}\frac{\partial^2\psi}{\partial t^2} \quad (85)$$

If we make the substitution

$$\psi = \frac{u(r, t)}{r} \quad (86)$$

then

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial \psi}{\partial r}\right) = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r} - u\right)$$

$$= \frac{1}{r}\frac{\partial^2 u}{\partial r^2}$$

Thus Eq. (85) becomes

$$\frac{1}{r}\frac{\partial^2 u}{\partial r^2} = \frac{1}{v^2}\frac{1}{r}\frac{\partial^2 u}{\partial t^2}$$

or

$$\frac{\partial^2 u}{\partial r^2} = \frac{1}{v^2}\frac{\partial^2 u}{\partial t^2} \qquad \textbf{(87)}$$

which is of the same form as the one-dimensional wave equation. The general solution of Eq. (87) is therefore given by

$$\psi = \frac{f(r - vt)}{r} + \frac{g(r + vt)}{r} \qquad \textbf{(88)}$$

the first and the second terms (on the RHS) representing an outgoing spherical wave and an incoming spherical wave, respectively. For time dependence of the form exp ($\pm i\omega t$), one obtains

$$\psi = \frac{A}{r}\exp\left[i(kr \pm \omega t)\right] \qquad \textbf{(89)}$$

Notice that the factor $1/r$ term implies that the amplitude of a spherical wave decreases inversely with r, and therefore the intensity will fall off as $1/r^2$.

Summary

- For a sinusoidal wave, the displacement is given by

  $$\Psi = a\cos\left[kx \pm \omega t + \phi\right]$$

  where a represents the amplitude of the wave, ω ($= 2\pi\nu$) is the angular frequency of the wave, k ($= 2\pi/\lambda$) is the wave number, and λ represents the wavelength associated with the wave. The upper and lower signs correspond to waves propagating in the $-x$ and $+x$ directions, respectively. Such a displacement is indeed produced in a long stretched string at the end of which a continuously vibrating tuning fork is placed. The quantity ϕ is known as the phase of the wave.
- The most general solution of the wave equation

  $$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 \psi}{\partial t^2}$$

  is of the form

  $$\psi = f(x - vt) + g(x + vt)$$

  where f and g are arbitrary functions of their argument. The first term on the RHS of the above equation represents a disturbance propagating in the $+x$ direction with speed v, and similarly, the second term represents a disturbance propagating in the $-x$ direction with speed v. Thus if we can derive the wave equation from physical considerations, then we can be sure that waves will result and ψ will represent the displacement associated with the wave.
- For a spherical wave, the displacement is given by

  $$\Psi = \frac{A}{r}e^{i(kr \pm \omega t)}$$

  where the + and – signs correspond to incoming and outgoing waves, respectively. Notice that the factor $1/r$ term implies that the amplitude of a spherical wave decreases inversely with r, and therefore, the intensity will fall off as $1/r^2$.

Problems

11.1 The displacement associated with a wave is given by
(a) $y(x, t) = 0.1\cos(0.2x - 2t)$
(b) $y(x, t) = 0.2\sin(0.5x + 3t)$
(c) $y(x, t) = 0.5\sin 2\pi(0.1x - t)$
where in each case x and y are measured in centimeters and t in seconds. Calculate the wavelength, amplitude, frequency, and velocity in each case.

[**Ans.:** (a) $\nu \approx 0.32$ s^{-1}; $v = 10$ cm s^{-1}; (b) $\nu \approx 0.48$ s^{-1}; $v = 6$ cm s^{-1}; (c) $\nu = 1$ s^{-1}; $v = 10$ cm s^{-1}]

11.2 A transverse wave ($\lambda = 15$ cm, $\nu = 200$ s^{-1}) is propagating on a stretched string in the $+x$ direction with an amplitude of 0.5 cm. At $t = 0$, the point $x = 0$ is at its equilibrium position moving in the upward direction. Write the equation describing the wave, and if $\rho = 0.1$ g cm^{-1}, calculate the energy associated with the wave per unit length of the wire.

[**Ans.:** Energy associated with the wave $\approx 1.97 \times 10^4$ erg cm^{-1}]

11.3 Assuming that the human ear can hear in the frequency range $20 < \nu < 20{,}000$ Hz, what will be the corresponding wavelength range?

[**Ans.:** 16.5 m $> \lambda >$ 0.0165 m]

11.4 Calculate the speed of longitudinal waves at NTP in **(a)** argon ($\gamma = 1.67$), **(b)** hydrogen ($\gamma = 1.41$).

[**Ans.:** **(a)** 308 m s^{-1}; **(b)** 1.26×10^5 cm s^{-1}]

11.5 Consider a wave propagating in the $+x$ direction with speed 100 cm s^{-1}. The displacement at $x = 10$ cm is given by the following equation:

$$y(x = 10, t) = 0.5\sin(0.4\,t)$$

where x and y are measured in centimeters and t in seconds. Calculate the wavelength and the frequency associated with the wave, and obtain an expression for the time variation of the displacement at $x = 0$.

[**Ans.:** $\lambda \approx 1571$ cm: $y(x, t) = 0.5\sin[0.4\,t - 0.004(x - 10)]$]

11.6 Consider a wave propagating in the $-x$ direction whose frequency is 100 s^{-1}. At $t = 5$ s the displacement associated with the wave is given by

$$y(x, t = 5) = 0.5\cos(0.1x)$$

where x and y are measured in centimeters and t in seconds. Obtain the displacement (as a function of x) at $t = 10$ s. What are the wavelength and the velocity associated with the wave?

[**Ans.:** $y(x, t) = 0.5 \cos[0.1x + 200\pi(t - 5]]$

11.7 Repeat the above problem corresponding to

$$y(x, t = 5) = 0.5 \cos (0.1x) + 0.4 \sin (0.1x + \pi/3)$$

11.8 A Gaussian pulse is propagating in the $+x$ direction and at $t = t_0$ the displacement is given by

$$y(x, t = t_0) = a \exp\left[-\frac{(x-b)^2}{\sigma^2}\right]$$

Find $y(x, t)$.

$$\left[\textbf{Ans.:}\ y(x,t) = a \exp\left\{-\frac{[x - b - v(t - t_0)^2]}{\sigma^2}\right\}\right]$$

11.9 A sonometer wire is stretched with a tension of 1 N. Calculate the velocity of transverse waves if $\rho = 0.2$ g cm^{-1}.

[**Ans.:** $v \approx 707$ cm s^{-1}]

11.10 The displacement associated with a three-dimensional wave is given by

$$\psi(x, y, z, t) = a \cos\left(\frac{\sqrt{3}}{2}kx + \frac{1}{2}ky - \omega t\right)$$

Show that the wave propagates along a direction making an angle of 30° with the x axis.

11.11 Obtain the unit vector along the direction of propagation for a wave, the displacement of which is given by

$$\psi(x, y, z, t) = a \cos (2x + 3y + 4z - 5t)$$

where, x, y, and z are measured in centimeters and t is in seconds. What will be the wavelength and the frequency of the wave?

$$\left[\textbf{Ans.:}\ \frac{2}{\sqrt{29}}\hat{\mathbf{x}} + \frac{3}{\sqrt{29}}\hat{\mathbf{y}} + \frac{4}{\sqrt{29}}\hat{\mathbf{z}}\right]$$

REFERENCES AND SUGGESTED READINGS

1. H. J. J. Braddick, *Vibrations, Waves and Diffraction*, McGraw-Hill Publishing Co., London, 1965.
2. F. S. Crawford, *Waves and Oscillations, Berkeley Physics Course*, Vol. III, McGraw-Hill Book Co., New York, 1968.
3. C. A. Coulson, *Waves*, 7th ed., Oliver & Boyd Ltd., Edinburgh, Scotland, 1955.
4. W. C. Elmore and M. A. Heald, *Physics of Waves*, McGraw-Hill Publishing Co., Maidenhead, 1969.
5. G. Joos, *Theoretical Physics* (trans. by I. M. Freeman), Blackie & Son Ltd., London, 1955.
6. H. J. Pain, *The Physics of Vibrations and Waves*, John Wiley & Sons, London, 1968.
7. Physical Science Study Committee, *Physics*, D.C. Heath and Co., Boston, Mass., 1967.
8. J. C. Slater and N. H. Frank, *Electromagnetism*, Dover Publications, New York, 1969.
9. R. A. Waldson, *Waves and Oscillations*, Van Nostrand Publishing Co., New York, 1964.

Chapter Twelve HUYGENS' PRINCIPLE AND ITS APPLICATIONS

Christian Huygens, a Dutch physicist, in a communication to the Academie des Sciences in Paris, propounded his wave theory of light (published in his Traite de Lumiere in 1690). He considered that light is transmitted through an all-pervading aether that is made up of small elastic particles, each of which can act as a secondary source of wavelets. On this basis, Huygens explained many of the known propagation characteristics of light, including the double refraction in calcite discovered by Bartholinus.

—From the Internet

12.1 INTRODUCTION

The wave theory of light was first put forward by Christian Huygens in 1678. During that period, everyone believed in Newton's corpuscular theory, which had satisfactorily explained the phenomena of reflection and refraction, the rectilinear propagation of light, and the fact that light could propagate through vacuum. So empowering was Newton's authority that the scientists around Newton believed in the corpuscular theory much more than Newton himself; as such, when Huygens put forward his wave theory, no one really believed him. On the basis of his wave theory, Huygens explained satisfactorily the phenomena of reflection, refraction, and total internal reflection and also provided a simple explanation of the then recently discovered birefringence (see Chap. 22). As we will see later, Huygens' theory predicted that the velocity of light in a medium (such as water) should be less than the velocity of light in free space, which is just the converse of the prediction made from Newton's corpuscular theory (see Sec. 2.2).

The wave character of light was not really accepted until the interference experiments of Young and Fresnel (in the early part of the nineteenth century) which could be explained only on the basis of a wave theory. At a later date, the data on the speed of light through transparent media were also available which were consistent with the results obtained by using the wave theory. Huygens did not know whether the light waves were longitudinal or transverse and also how they propagate through vacuum. It was only in the later part of the nineteenth century, when Maxwell propounded his famous electromagnetic theory, that the nature of light waves could be understood properly.

12.2 HUYGENS' THEORY

Huygens' theory is essentially based on a geometrical construction which allows us to determine the shape of the wave front at any time, if the shape of the wave front at an earlier time is known. A wave front is the locus of the points which are in the same phase; for example, if we drop a small stone in a calm pool of water, circular ripples spread out from the point of impact, each point on the circumference of the circle (whose center is at the point of impact) oscillates with the same amplitude and same phase, and thus we have a circular wave front. On the other hand, if we have a point source emanating waves in a uniform isotropic medium, the locus of points which have the same amplitude and are in the same phase is spheres. In this case we have spherical wave fronts, as shown in Fig. 12.1(a). At large distances from the source, a small portion of the sphere can be considered as a plane, and we have what is known as a plane wave [see Fig. 12.1(b)].

Now, according to Huygens' principle, each point of a wave front is a source of secondary disturbance, and the wavelets emanating from these points spread out in all directions with the speed of the wave. The envelope of these wavelets gives the shape of the new wave front. In Fig. 12.2, S_1S_2 represents the shape of the wave front (emanating from the point O) at a particular time which we denote as $t = 0$.

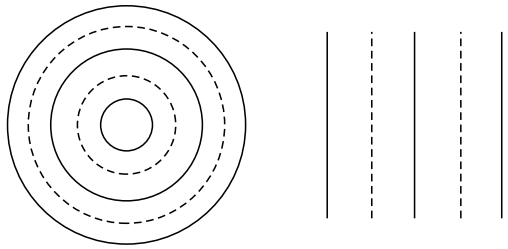


Fig. 12.1 (a) A point source emitting spherical waves. (b) At large distances, a small portion of the spherical wave front can be approximated to a plane wave front, thus resulting in plane waves.

The medium is assumed to be homogeneous and isotropic; i.e., the medium is characterized by the same property at all points, and the speed of propagation of the wave is the same in all directions. Let us suppose we want to determine the shape of the wave front after a time interval of Δt. Then with each point on the wave front as center, we draw spheres of radius $v\Delta t$, where v is the speed of the wave in that medium. If we draw a common tangent to all these spheres, then we obtain the envelope which is again a sphere centered at O. Thus the shape of the wave front at a later time Δt is the sphere $S_1'S_2'$.

There is, however, one drawback with the above model, because we also obtain a back wave which is not present in practice. This back wave is shown as $S_1''S_2''$ in Fig. 12.2. In Huygens' theory, the presence of the back wave is avoided by assuming that the amplitude of the secondary wavelets is not uniform in all directions; it is maximum in the forward direction and zero in the backward direction[1]. The absence of the back wave is really justified through the more rigorous wave theory.

In the next section we will discuss the original argument of Huygens to explain the rectilinear propagation of light. In Sec. 12.4 we will derive the laws of refraction and reflection by using Huygens' principle. Finally, in Sec. 12.5 we will show how Huygens' principle can be used in inhomogeneous media.

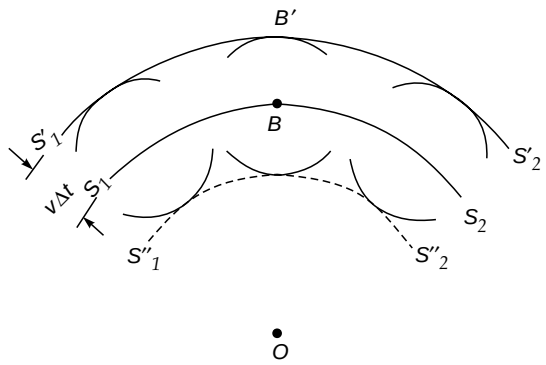


Fig. 12.2 Huygens' construction for the determination of the shape of the wave front, given the shape of the wave front at an earlier time. S_1S_2 is a spherical wave front centered at O at a time, say, $t = 0$. $S_1'S_2'$ corresponds to the state of the wave front at a time Δt, which is again spherical and centered at O. The dashed curve represents the back wave.

12.3 RECTILINEAR PROPAGATION

Let us consider spherical waves emanating from the point source O and striking the obstacle A (see Fig. 12.3). According to the rectilinear propagation of light (which is also predicted by corpuscular theory), one should obtain a shadow in the region PQ of the screen. As we will see in a later chapter, this is not quite true and one does obtain a finite intensity in the region of the geometrical shadow. However, at the time of Huygens, light was known to travel in straight lines, and Huygens explained this by assuming that the secondary wavelets do not have any amplitude at any point not enveloped by the wave front. Thus, referring to Fig. 12.2, the secondary wavelets emanating from a typical point B will give rise to a finite amplitude at B' only and not at any other point.

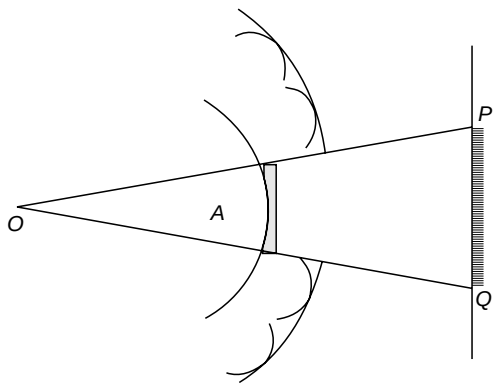


Fig. 12.3 Rectilinear propagation of light. Point O is a point source emitting spherical waves, and A is an obstacle which forms a shadow in the region PQ of the screen.

[1] Indeed it can be shown from diffraction theory that one does obtain (under certain approximations) an obliquity factor, which is of the form $\frac{1}{2}(1 + \cos\theta)$, where θ is the angle between the normal to the wave front and the direction under consideration. Clearly when $\theta = 0$, the obliquity factor is 1 (thereby giving rise to maximum amplitude in the forward direction) and when $\theta = \pi$, the obliquity factor is zero (thereby giving rise to zero amplitude in the backward direction).

The above explanation of the rectilinear propagation of light is indeed unsatisfactory and is incorrect. Further, as pointed out earlier, one does observe a finite intensity of light in the geometrical shadow. A satisfactory explanation was put forward by Fresnel, who postulated that the secondary wavelets mutually interfere. Huygens' principle, along with the fact that the secondary wavelets mutually interfere, is known as the Huygens–Fresnel principle. If a plane wave is allowed to fall on a tiny hole,[2] then the hole approximately acts as a point source and spherical waves emanate from it [see Fig. 12.4(a) and (b)]. This fact is in direct contradiction to the original proposition of Huygens[3] according to which the secondary wavelets do not have any amplitude at any point not enveloped by the wave front; however, as we will see in the chapter on diffraction, it can be explained satisfactorily on the basis of the Huygens–Fresnel principle.

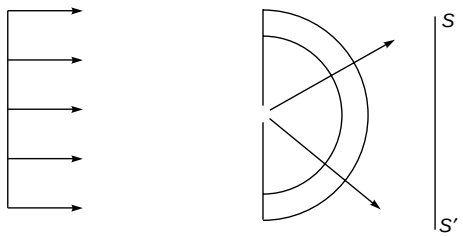


Fig. 12.4 (a) A plane wave front is incident on a pinhole. If the diameter of the pinhole is small (compared to the wavelength), the entire screen SS' will be illuminated; see also Fig. 17 in the insert at the back of the book.

12.4 APPLICATION OF HUYGENS' PRINCIPLE TO STUDY REFRACTION AND REFLECTION

12.4.1 Refraction of a Plane Wave at a Plane Interface

We will first derive the laws of refraction. Let S_1S_2 be a surface separating two media with different speeds of propagation of light v_1 and v_2 as shown in Fig. 12.5. Let A_1B_1 be a plane wave front incident on the surface at an angle i; A_1B_1 represents the position of the wave front at an instant $t = 0$.

Let τ be the time taken for the wave front to travel the distance B_1B_3. Then $B_1B_3 = v_1\tau$. In the same time the light would have traveled a distance $A_1A_3 = v_2\tau$ in the second medium. (Note that the lines A_1A_3, B_1B_3, etc. are always normal to the wave front; these represent rays in isotropic media—see Chap. 4.) It can be easily seen that the incident and

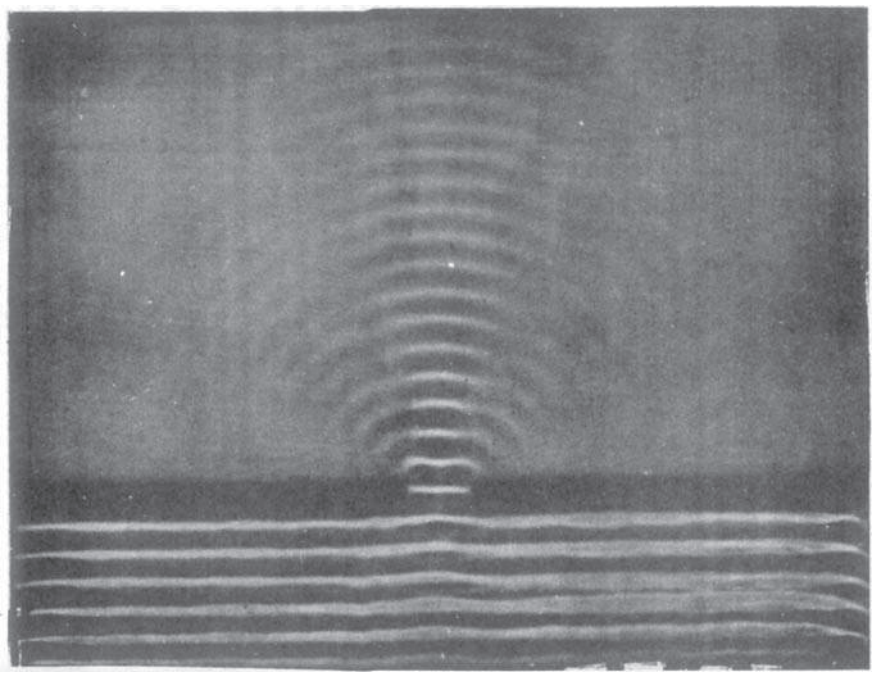


Fig. 12.4 (b) Diffraction of straight water wave when it passes through an opening (Adapted from Ref. 6).

[2] By a tiny hole we imply that the diameter of the hole should be of the order of 0.1 mm or less.

[3] Use of Huygens' principle in determining the shape of the wave front in anisotropic media will be discussed in Chap. 22.

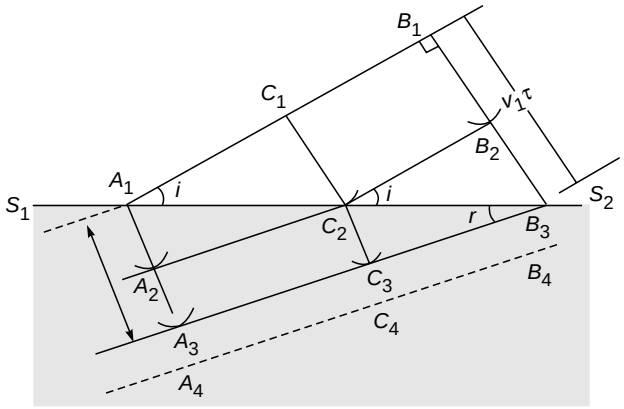


Fig. 12.5 Refraction of a plane wave front A_1B_1 by a plane interface S_1S_2 separating two media with different velocities of propagation of light v_1 and v_2 ($< v_1$); i and r are the angles of incidence and refraction, respectively. $A_2C_2B_2$ corresponds to the shape of the wave front at an intermediate time τ_1. Notice that $r < i$.

refracted rays make angles i and r, respectively, with the normal. To determine the shape of the wave front at the instant $t = \tau$, we consider an arbitrary point C_1 on the wave front. Let the time taken for the disturbance to travel the distance C_1C_2 be τ_1. Thus $C_1C_2 = v_1\tau_1$. From point C_2 we draw a secondary wavelet of radius $v_2(\tau - \tau_1)$. Similarly from point A_1 we draw a secondary wavelet of radius $v_2\tau$. The envelope of these secondary wavelets is shown as $A_3C_3B_3$. The shape of the wave front at the intermediate time τ_1 is shown as $A_2C_2B_2$, and clearly $B_1B_2 = C_1C_2 = v_1\tau_1$ and $A_1A_2 = v_2\tau_1$. In the right-angle triangles $B_2C_2B_3$ and $C_3C_2B_3$, $\angle B_2C_2B_3 = i$ (the angle of incidence) and $\angle C_2B_3C_3 = r$ (the angle of refraction). Clearly,

$$\frac{\sin i}{\sin r} = \frac{B_2B_3/C_2B_3}{C_2C_3/C_2B_3} = \frac{B_2B_3}{C_2C_3} = \frac{v_1(\tau - \tau_1)}{v_2(\tau - \tau_1)} = \frac{v_1}{v_2} \qquad \textbf{(1)}$$

which is known as Snell's law. It is observed that when light travels from a rarer to a denser medium, the angle of incidence is greater than the angle of refraction, and consequently

$$\frac{\sin i}{\sin r} > 1$$

which implies $v_1 > v_2$; thus, Huygens' theory predicts that the speed of light in a rarer medium is greater than the speed of light in a denser medium. This prediction is contradictory to that made by Newton's corpuscular theory (see Sec. 2.2), and as later experiments showed, the prediction of the wave theory was indeed correct.

If c represents the speed of light in free space, then the ratio c/v (where v represents the speed of light in the particular medium) is called the refractive index n of the medium. Thus if $n_1 (= c/v_1)$ and $n_2 (= c/v_2)$ are the refractive indices of the two media, then Snell's law can also be written as

$$n_1 \sin i = n_2 \sin r \qquad \textbf{(2)}$$

Let $A_1C_1B_1$, $A_2C_2B_2$, $A_3C_3B_3$, and $A_4C_4B_4$ denote the successive positions of crests. If λ_1 and λ_2 denote the wavelength of light in medium 1 and medium 2, respectively, then the distance B_1B_2 (= B_2B_3 = C_1C_2) will be equal to λ_1 and the distance A_1A_2 (= A_2A_3 = C_2C_3) will be equal to λ_2. From Fig. 12.5 it is obvious that

$$\frac{\lambda_1}{\lambda_2} = \frac{\sin i}{\sin r} = \frac{v_1}{v_2} \qquad \textbf{(3)}$$

or

$$\frac{v_1}{\lambda_1} = \frac{v_2}{\lambda_2} \qquad \textbf{(4)}$$

Thus, when a wave gets refracted into a denser medium ($v_1 > v_2$), the wavelength and the speed of propagation decrease, but the frequency (= v/λ) remains the same; when refracted into a rarer medium, the wavelength and the speed of propagation will increase. In Table 12.1 we have given the indices of refraction of several materials with respect to vacuum. In Table 12.2, the wavelength dependence of the

Table 12.1 Refractive Indices of Various Materials Relative to Vacuum (Adapted from Ref. 1)

(For light of wavelength $\lambda = 5.890 \times 10^{-5}$ cm)

Material	*n*	*Material*	*n*
Vacuum	1.0000	Quartz (fused)	1.46
Air	1.0003	Rock salt	1.54
Water	1.33	Glass (ordinary crown)	1.52
Quartz (crystalline)	1.54	Glass (dense flint)	1.66

Table 12.2 Refractive Indices of Telescope Crown Glass and Vitreous Quartz for Various Wavelengths (Adapted from Ref. 7)

	Wavelength	*Telescope crown*	*Vitreous quartz*
1	6.562816×10^{-5} cm	1.52441	1.45640
2	5.889953×10^{-5} cm	1.52704	1.45845
3	4.861327×10^{-5} cm	1.53303	1.46318

Note: The wavelengths specified at serial numbers 1, 2, and 3 correspond roughly to the red, yellow, and blue colors. The table shows the accuracy with which the wavelengths and refractive indices can be measured; see also Prob. 10.5.

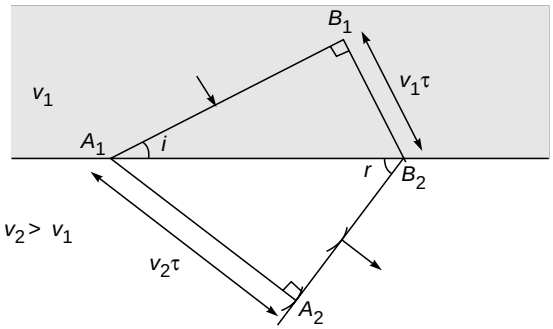


Fig. 12.6 Refraction of a plane wave front incident on a rarer medium (i.e., $v_2 > v_1$). Notice that the angle of refraction r is greater than the angle of incidence i. The value of i, when r is equal to $\pi/2$, gives the critical angle.

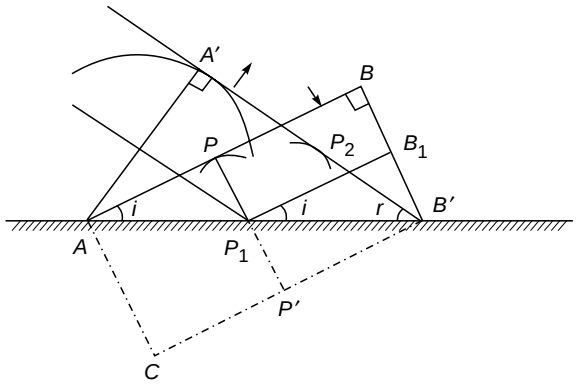


Fig. 12.7 Reflection of a plane wave front AB incident on a plane mirror. $A'B'$ is the reflected wave front; i and r correspond to angles of incidence and reflection, respectively.

refractive index for crown glass and vitreous quartz is given. The three wavelengths correspond roughly to the red, yellow, and blue colors. Notice the accuracy with which the wavelength and the refractive index can be measured.

12.4.2 Total Internal Reflection

In Fig. 12.5 the angle of incidence has been shown to be greater than the angle of refraction. This corresponds to the case when $v_2 < v_1$, i.e., the light wave is incident on a denser medium. However, if the second medium is a rarer medium (i.e., $v_1 < v_2$), then the angle of refraction will be greater than the angle of incidence, and a typical refracted wave front has the form shown in Fig. 12.6, where $B_1B_2 = v_1\tau$ and $A_1A_2 = v_2\tau$. Clearly, if the angle of incidence is such that $v_2\tau$ is greater than A_1B_2, then the refracted wave front will be absent and we will have what is known as total internal reflection. The critical angle will correspond to

$$A_1B_2 = v_2\tau$$

Thus

$$\sin i_c = \frac{B_1B_2}{A_1B_2} = \frac{v_1}{v_2} = n_{12} \qquad \textbf{(5)}$$

where i_c denotes the critical angle and n_{12} represents the refractive index of the second medium with respect to the first. For all angles of incidence greater than i_c, we will have total internal reflection.

12.4.3 Reflection of a Plane Wave by a Plane Surface

Let us consider a plane wave AB incident at an angle i on a plane mirror as shown in Fig. 12.7. We consider the reflection of the plane wave and try to obtain the shape of the reflected wave front. Let the position of the wave front at $t = 0$ be AB. If the mirror were not present, then at a later time τ the position of the wave front would be CB', where $BB' = PP' = AC = v\tau$ and v is the speed of propagation of the wave. To determine the shape of the reflected wave front at the instant $t = \tau$, we consider an arbitrary point P on the wave front AB and let τ_1 be the time taken by a disturbance to reach point P_1 from P. From point P_1, we draw a sphere of radius $v(\tau - \tau_1)$. We draw a tangent plane on this sphere from point B'. Since $BB_1 = PP_1 = v\tau_1$, the distance B_1B' is equal to P_1P_2 $[= v(\tau - \tau_1)]$. If we consider triangles P_2P_1B' and B_1P_1B', then the side P_1B' is common to both and since $P_1P' = B'B_2$, and since both the triangles are right-angle triangles, $\angle P_2B'P_1 = \angle B_1P_1B'$. The former is the angle of reflection, and the latter is the angle of incidence. Thus, we have the law of reflection; when a plane wave front gets reflected from a plane surface, the angle of reflection is equal to the angle of incidence and the reflected wave is a plane wave.

12.4.4 Diffuse Reflection

In the above we have considered the reflection of light from a smooth surface. This is known as specular reflection. If the surface is irregular (as shown in Fig. 12.8) we have diffuse reflection. The secondary wavelets emanating from the irregular surface travel in many directions, and we do not have a well-defined reflected wave. Indeed, it can be shown that if

Fig. 12.8 Diffuse reflection of a plane wave front from a rough surface. It is evident that one does not have a well-defined reflected beam.

the irregularity in the surface is considerably greater than the wavelength, we will have diffuse reflection.

12.4.5 Reflection of Light from a Point Source near a Mirror

Let us consider spherical waves (emanating from a point source P) incident on a plane mirror MM', as shown in Fig. 12.9. Let ABC denote the shape of the wave front at time $t = 0$. In the absence of the mirror, the shape of the wave front at a later time τ would be $A_1B_1C_1$, where $AA_1 = BB_1 = CC_1 = v\tau$, Q being an arbitrary point on the wave front. If the time taken for the disturbance to traverse the distance QQ' is τ_1, then, to determine the shape of the reflected wave front, we draw a sphere of radius $v(\tau - \tau_1)$ whose center is at point Q'. In a similar manner we can draw the secondary wavelets emanating from other points on the mirror, and, in particular, from point B we have to draw a sphere of radius $v\tau$. The shape of the reflected wave front is obtained by drawing a common tangent plane to all these spheres, which is shown as $A_1B_1'C_1$ in the figure. It can be seen immediately that $A_1B_1'C_1$ will have an exactly similar shape as $A_1B_1C_1$ except that $A_1B_1'C_1$ will have its center of curvature at point P' where $PB = BP'$. Thus the reflected waves will appear to emanate from point P' which will be the virtual image of point P.

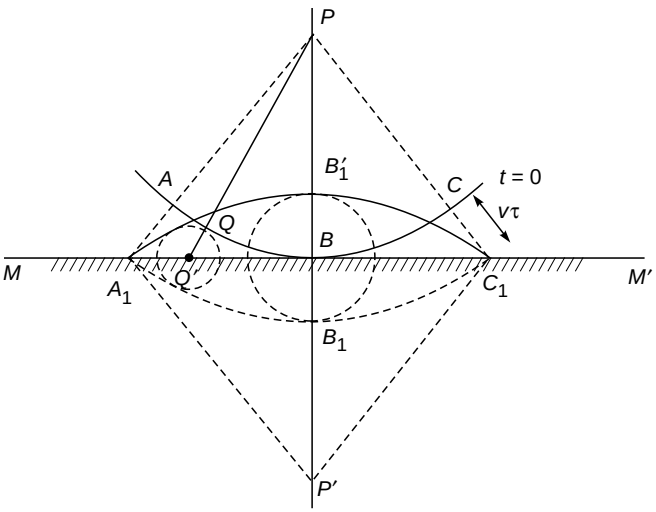


Fig. 12.9 Point P is a point source placed in front of a plane mirror MM'. ABC is the incident wave front (which is spherical and centered at P), and $A_1B_1'C_1$ is the corresponding reflected wave front (which is spherical and centered at P'). Point P' is the virtual image of P.

12.4.6 Refraction of a Spherical Wave by a Spherical Surface

Let us consider spherical waves (emanating from point P) incident on the curved spherical surface SBS'. Let the shape of the wave front at time $t = 0$ be ABC [see Fig. 12.10(a)]. Let the refractive indices on the left and on the right of the spherical surface be n_1 and n_2, respectively. In the absence of the spherical surface, the shape of the wave front at a later time τ would be $A_1B_1C_1$, where $AA_1 = BB_1 = CC_1 = v_1\tau$. We consider an arbitrary point Q on the wave front ABC and let τ_1 be the time taken for the disturbance to reach point Q' (on the surface of the spherical wave); thus $QQ' = v_1\tau_1$. To determine the shape of the refracted wave front at a later time τ, we draw a sphere of radius $v_2(\tau - \tau_1)$ from point Q'. We may draw similar spheres from other points on the spherical surface; in particular, the radius of the spherical wave front from point B, which is equal to BB_2, will be $v_2\tau$. The envelope of these spherical wavelets is shown as $A_1B_2C_1$ which, in general, will not be a sphere.[4] However, a small portion of any curved surface can be considered as a sphere, and in this approximation we may consider $A_1B_2C_1$ to be a sphere whose center of curvature is at point M. The spherical wave front will, therefore, converge toward point M, and hence point M represents the real image of point P.

We adopt a sign convention in which all distances, measured to the left of point B, are negative and all distances measured to the right of point B are positive. Thus

$$PB = -u$$

where u itself is a negative quantity. Further, since point M lies on the right of B, we have

$$BM = v$$

and similarly,

$$BO = R$$

where O represents the center of curvature of the spherical surface.

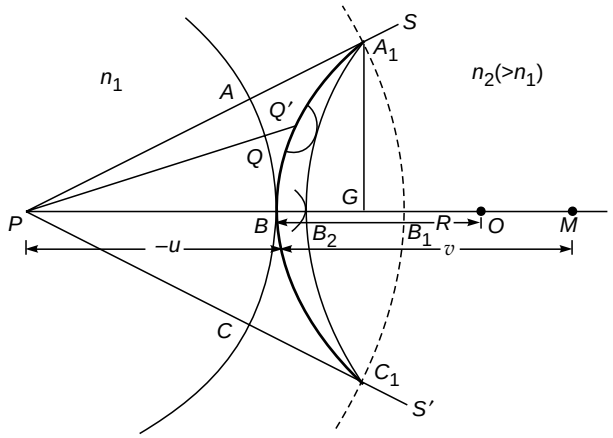
To derive a relation among u, v, and R, we use a theorem in geometry, according to which

$$(A_1G)^2 = GB \times (2R - GB) \qquad \textbf{(6)}$$

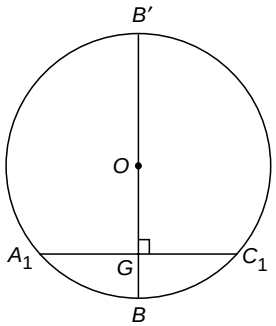
where G is the foot of the perpendicular on the axis PM [see Fig. 12.10(b)]. In Fig. 12.10(b) the diameter $B'OB$ intersects the chord A_1GC_1 normally. If $GB \ll R$, then

$$(A_1G)^2 \simeq 2R(GB)$$

[4] The fact that the refracted wave front is not, in general, a sphere leads to, what are known as aberrations.



(a)



(b)

Fig. 12.10 (a) Refraction of a spherical wave *ABC* (emanating from the point source *P*) by a convex spherical surface *SBS′* separating media of refractive indices n_1 and n_2 (> n_1). $A_1B_2C_1$ is the refracted wave front, which is approximately spherical and whose center of curvature is at *M*. Thus *M* is the real image of *P*. Point *O* is the center of curvature of *SS′*. (b) The diameter *B′OB* intersects the chord A_1GC_1 normally.

Consider the spherical surface *SBS′* [see Fig. 12.10(a)] whose radius is *R*. Clearly,

$$(A_1G)^2 = (2R - GB)GB \simeq 2R(GB) \qquad \textbf{(7)}$$

where we have assumed $GB \ll R$. Similarly by considering the spherical surface $A_1B_2C_1$ (whose center is at point *M*) we obtain

$$(A_1G)^2 \simeq 2v(GB_2) \qquad \textbf{(8)}$$

where $v = BM \simeq B_2M$. In a similar manner,

$$(A_1G)^2 \simeq 2(-u)GB_1 \qquad \textbf{(9)}$$

Since *u* is a negative quantity, $(A_1G)^2$ is positive. Now

$$BB_1 = v_1\tau \qquad \text{and} \qquad BB_2 = v_2\tau$$

Therefore

$$\frac{BB_1}{BB_2} = \frac{v_1}{v_2} = \frac{n_2}{n_1}$$

or $$n_1BB_1 = n_2BB_2$$

or $$n_1(BG + GB_1) = n_2(BG - GB_2)$$

or

$$n_1\left[\frac{(A_1G)^2}{2R} - \frac{(A_1G)^2}{2u}\right] = n_2\left[\frac{(A_1G)^2}{2R} - \frac{(A_1G)^2}{2v}\right]$$

where we used Eqs. (7), (8), and (9). Thus

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R} \qquad \textbf{(10)}$$

which may be rewritten in the form

$$\frac{n_2}{v} = \frac{n_1}{u} + \frac{n_2 - n_1}{R} \qquad \textbf{(11)}$$

Thus, if

$$\frac{n_1}{|u|} > \frac{n_2 - n_1}{R}$$

or $$|u| < \frac{Rn_1}{n_2 - n_1}$$

we will obtain a virtual image. (We are of course assuming that the second medium is a denser medium, i.e., $n_2 > n_1$; if $n_2 < n_1$, we will always have a virtual image.)

A converging spherical wave front will propagate in a manner shown in Fig. 12.11. Beyond the focal point it will start diverging as shown in the figure.[5]

In a similar manner we can consider the refraction of a spherical wave from a surface *SBS′* shown in Fig. 12.12 ($n_2 > n_1$). Here the center of curvature will also lie on the left

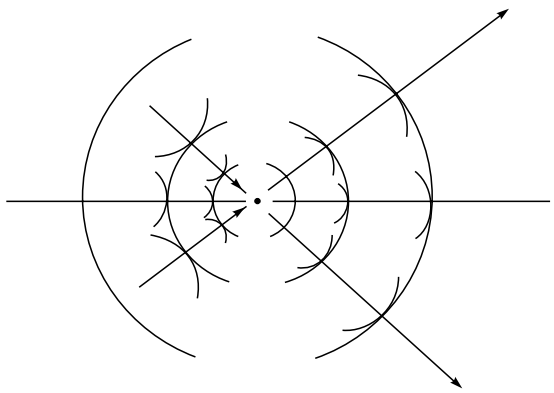


Fig. 12.11 Propagation of a converging spherical wave using Huygens' principle.

[5] Very close to the focal point, one has to use a more rigorous wave theory, and the shape of the wave front is very much different from spherical (see Ref. 8). However, much beyond the focal point the wave fronts again become spherical.

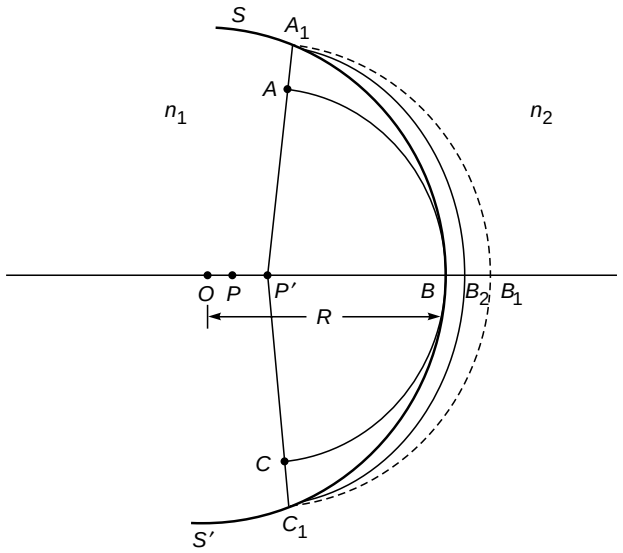


Fig. 12.12 Refraction of a spherical wave by a concave surface separating media of refractive indices n_1 and n_2 (> n_1). Point P' is the virtual image of P.

of point B, and both u and R will be negative quantities. Thus no matter what the values of u and R may be, v will be negative and we will obtain a virtual image.

Using Eq. (10), we can easily derive the thin lens formula. We assume a thin lens made of a material of refractive index n_2 to be placed in a medium of refractive index n_1 (see Fig. 12.13). Let the radii of curvature of the first and second surfaces be R_1 and R_2, respectively. Let v' be the distance of the image of the object P if the second surface were not present. Then

$$\frac{n_2}{v'} - \frac{n_1}{u} = \frac{n_2 - n_1}{R_1} \qquad \textbf{(12)}$$

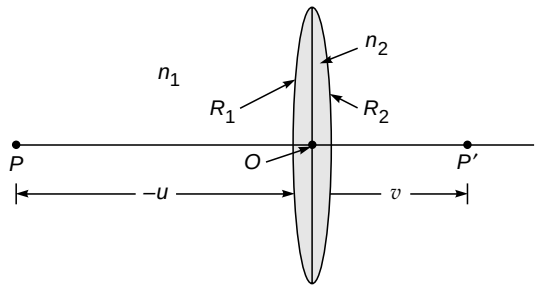


Fig. 12.13 A thin lens made of a medium of refractive index n_2 placed in a medium of refractive index n_1. The radii of curvature of the two surfaces are R_1 and R_2. Point P is the image (at a distance v from point O) of the point object P (at a distance $-u$ from point O).

(Since the lens is assumed to be thin, all the distances are measured from point O). This image now acts as an object to the spherical surface R_2 on the left of which is the medium of refractive index n_2 and on the right of which is the medium of refractive index n_1. Thus, if v is the distance of the final image point from O, then

$$\frac{n_1}{v} - \frac{n_2}{v'} = \frac{n_1 - n_2}{R_2} \qquad \textbf{(13)}$$

Adding Eqs. (12) and (13), we obtain

$$\frac{n_1}{v} - \frac{n_1}{u} = (n_2 - n_1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad \textbf{(14)}$$

or

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f} \qquad \textbf{(15)}$$

where

$$\frac{1}{f} = \frac{n_2 - n_1}{n_1}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad \textbf{(16)}$$

Notice that we do not have to worry whether v' is positive or negative; it is automatically taken care of through the sign convention. Further, the relation derived is valid for any lens; for example, for a double convex lens, R_1 is positive and R_2 is negative; and for a double concave lens, R_1 is negative and R_2 is positive. Similarly it follows for other types of lenses (see Fig. 5.6).

Example 12.1 Consider a vibrating source moving through a medium with a speed V. Let the speed of propagation of the wave in the medium be v. Show that if $V > v$, then a conical wave front is set up whose half angle is given by

$$\theta = \sin^{-1}\left(\frac{v}{V}\right) \qquad \textbf{(17)}$$

Solution: At $t = 0$ let the source be at point P_0 moving with a speed V in the x direction (see Fig. 12.14). We wish to find out the wave front at a later time τ. The disturbance emanating from point P_0 traverses a distance $v\tau$ in time τ. Thus from point P_0 we draw a

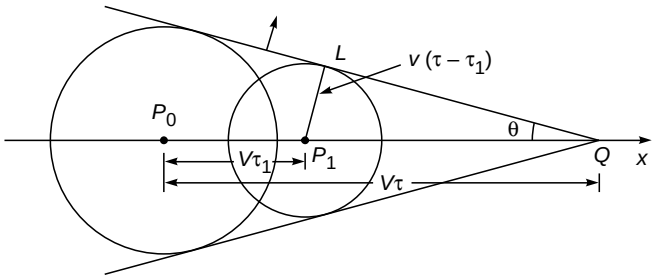


Fig. 12.14 Generation of a shock wave front by a vibrating particle P_0 moving with a speed V in a medium in which the velocity of propagation of the wave is v (< V).

sphere of radius $v\tau$. We next consider the waves emanating from the source at a time τ_1 ($<\tau$). At time τ_1 let the source be at position P_1; consequently,

$$P_0P_1 = V\tau_1$$

To determine the shape of the wave front at τ, we draw a sphere of radius $v(\tau - \tau_1)$ centered at P_1. Let the source be at position Q at the instant τ. Then

$$P_0Q = V\tau$$

We draw a tangent plane from point Q on the sphere whose origin is point P_1. Since

$$P_1L = v(\tau - \tau_1) \quad \text{and} \quad P_1Q = V(\tau - \tau_1)$$

$$\sin\theta = \frac{P_1L}{P_1Q} = \frac{v}{V} \quad \text{(independent of } \tau_1\text{)}$$

Since θ is independent of τ_1, all the spheres drawn from any point on line P_0Q will have a common tangent plane. This plane is known as the shock wave front and propagates with a speed v.

It is interesting to point out that even when the source is not vibrating, if its speed is greater than the speed of sound waves, a shock wave front is always set up. A similar phenomenon also occurs when a charged particle (such as an electron) moves in a medium with a speed greater than the speed of light in that medium.[6] The emitted light is known as Cerenkov radiation. If you ever see a swimming pool type of reactor, you will find a blue glow coming from it; this is due to the Cerenkov radiation emitted by the fast-moving electrons.

Summary

- According to Huygens' principle, each point of a wave front is a source of secondary disturbance, and the wavelets emanating from these points spread out in all directions with the speed of the wave. The envelope of these wavelets gives the shape of the new wave front.
- Huygens' principle, along with the fact that the secondary wavelets mutually interfere, is known as the Huygens–Fresnel principle.
- Laws of reflection and Snell's law of refraction can be derived using Huygens' principle.
- Using Huygens' principle, one can derive the lens formula $\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$.

Problems

12.1 Use Huygens' principle to study the reflection of a spherical wave emanating from a point on the axis at a concave mirror of radius of curvature R, and obtain the mirror equation

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{R}$$

12.2 Consider a plane wave incident obliquely on the face of a prism. Using Huygens' principle, construct the transmitted wave front and show that the deviation produced by the prism is given by

$$\delta = i + t - A$$

where A is the angle of the prism and i and t are the angles of incidence and transmittance.

REFERENCES AND SUGGESTED READINGS

1. A. B. Arons, *Development of Concepts in Physics*, Addison-Wesley Publishing Co., Reading, Mass., 1965.
2. B. B. Baker and E. J. Copson, *The Mathematical Theory of Huygens' Principle*, Oxford University Press, London, 1969.
3. H. J. J. Braddick, *Vibration, Waves and Diffraction*, McGraw-Hill Publishing Co., London, 1965.
4. A. J. DeWitte, "Equivalence of Huygens' Principle and Fermat's Principle in Ray Geometry," *American Journal of Physics*, Vol. 27, p. 293, 1959.
5. C. Huygens, *Treatise on Light*, Dover Publications, New York, 1962.
6. PSSC, *Physics*, D.C. Heath and Company, Boston, Mass., 1965.
7. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 3d ed., McGraw-Hill, 1957, p. 465.
8. M. Born and E. Wolf, *Principle of Optics*, Pergamon Press, Oxford, 1975.

[6] This does not contradict the theory of relativity according to which no particle can have a speed greater than the speed of light in free space ($= 3 \times 10^8$ m s^{-1}). The speed of light in a medium will be equal to c/n, where n represents the refractive index. For example, in water, the speed of light will be about 2.25×10^8 m s^{-1}, and the speed of the electron could be greater than this value.

PART 3
Interference

This part covers the very important and fascinating area of interference and many beautiful experiments associated with it—the underlying principle is the superposition principle, which is discussed in Chap. 13. Chapter 14 discusses interference by division of wave front including the famous Young's double-hole interference experiment. In Chap. 15, interference by division of amplitude is discussed which allows us to understand colors of thin films and applications like antireflection films, etc. The basic principle of the working of the fiber Bragg gratings (usually abbreviated as FBG) is discussed along with some of their important applications in the industry. In the same chapter, the Michelson interferometer is also discussed which is *perhaps one of the most ingenious and sensational optical instruments ever and* for which Michelson received the Nobel Prize in Physics in 1907. Chapter 16 discusses the Fabry–Perot interferometer that is based on multiple-beam interference and is characterized by a high resolving power and hence finds applications in high-resolution spectroscopy. Chapter 17 discusses the basic concept of temporal and spatial coherence. The ingenious experiment of Michelson, which used the concept of spatial coherence to determine the angular diameter of stars, has been discussed in detail. Topics like optical beats and Fourier transform spectroscopy have also been discussed.

Chapter Thirteen SUPERPOSITION OF WAVES

The experiments described appear to me, at any rate, eminently adapted to remove any doubt as to the identity of light, radiant heat, and electromagnetic wave motion. I believe that from now on we shall have greater confidence in making use of the advantages which this identity enables us to derive both in the study of optics and of electricity.

—Heinrich Hertz (1888)[1]

13.1 INTRODUCTION

In this chapter we will discuss the applications of the principle of superposition of waves according to which the resultant displacement (at a particular point) produced by a number of waves is the vector sum of the displacements produced by each one of the disturbances. As a simple example, we consider a long stretched string *AB* (see Fig. 13.1). From the end *A*, a triangular pulse is generated which propagates to the right with a certain speed v. In the absence of any other disturbance, this pulse would have propagated in the $+x$ direction without any change in shape; we are, of course, neglecting any attenuation or distortion of the pulse. We next assume that from the end *B* an identical pulse is generated which starts moving to the left with the same speed v. (As shown in Sec. 11.6, the speed of the wave is determined by the ratio of the tension in the string to its mass per unit length.) At $t = 0$, the snapshot of the string is shown in Fig. 13.1(a). At a little later time each pulse moves close to the other, as shown in Fig. 13.1(b), without any interference. Figure 13.1(c) represents a snapshot at an instant when the two pulses interfere; the dashed curves represent the profile of the string if each of the impulses was moving all by itself, whereas the solid curve shows the resultant displacement obtained by algebraic addition of each displacement. Shortly later [Fig. 13.1(d)] the two pulses exactly overlap each other, and the resultant displacement is zero everywhere (where has the energy gone?). At a much later time the impulses sort of cross each other [Fig. 13.1(e)] and move as if nothing had happened. This is a characteristic feature of superposition of waves.

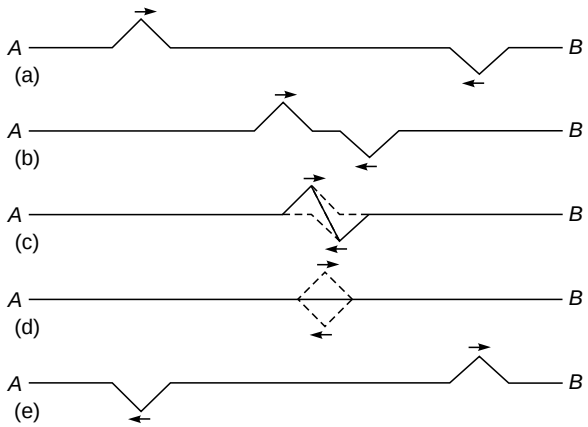


Fig. 13.1 The propagation in opposite directions of two triangular pulses in a stretched string. The solid line gives the actual shape of the string; (a), (b), (c), (d), and (e) correspond to different instants of time.

The phenomenon of interference contains no more physics than embodied in the above example. In the following sections we will consider some more examples.

13.2 STATIONARY WAVES ON A STRING

Consider a string which is fixed at point *A* (see Fig. 13.2). A transverse sinusoidal wave is sent down the string along the

[1] The author found this quotation in the book by Smith and King (Ref. 1).

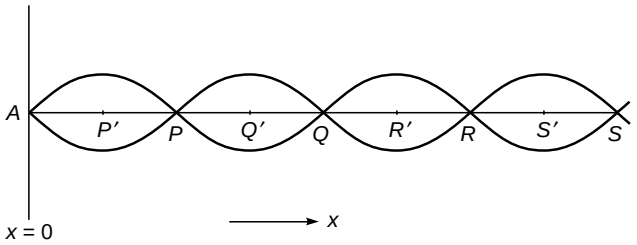


Fig. 13.2 Reflection of a wave at $x = 0$.

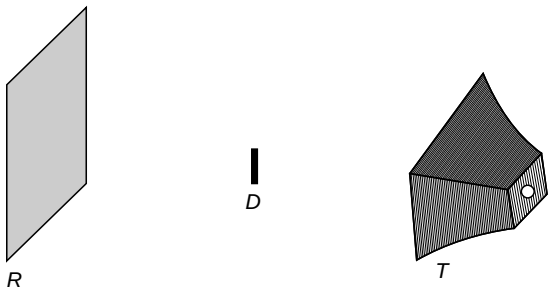


Fig. 13.3 An arrangement for studying standing electromagnetic waves.

$-x$ direction. The displacement at any point on the string due to this wave is given by

$$y_i = a \sin\left[\frac{2\pi}{\lambda}(x + vt) + \phi\right] \quad \textbf{(1)}$$

where the subscript i refers to the fact that we are considering the incident wave. Without any loss of generality we can set $\phi = 0$; thus we may write

$$y_i = a \sin\left[\frac{2\pi}{\lambda}(x + vt)\right]$$

$$= a \sin\left[2\pi\left(\frac{x}{\lambda} + vt\right)\right] \quad \textbf{(2)}$$

Thus, because of the incident wave, the displacement at point A is

$$y_i\big|_{x=0} = a\sin(2\pi\nu t) \quad \textbf{(3)}$$

where $\nu = v/\lambda$ and we have assumed point A to correspond to $x = 0$. Since point A is fixed, there must be a reflected wave such that the displacement due to this reflected wave (at point A) is equal and opposite to y_i:

$$y_r\big|_{x=0} = -a \sin(2\pi\nu t) \quad \textbf{(4)}$$

where the subscript r refers to the fact that we are considering the reflected wave. Since the reflected wave propagates in the $+x$ direction, we must have

$$y_r = +a \sin 2\pi\left(\frac{x}{\lambda} - \nu t\right) \quad \textbf{(5)}$$

The resultant displacement is given by

$$y = y_i + y_r = a\left[\sin 2\pi\left(\frac{x}{\lambda} + \nu t\right) + \sin 2\pi\left(\frac{x}{\lambda} - \nu t\right)\right]$$

$$= 2a \sin\frac{2\pi}{\lambda} x \cos 2\pi\nu t \quad \textbf{(6)}$$

Thus, for values of x such that

$$\sin\left(\frac{2\pi}{\lambda}x\right) = 0 \quad \textbf{(7)}$$

the displacement y is zero *at all times*. Such points are known as *nodes*; the x coordinates of the nodes are given by

$$x = 0, \frac{\lambda}{2}, \lambda, \frac{3\lambda}{2}, 2\lambda, \ldots \quad \textbf{(8)}$$

and are marked as points A, P, Q, and R in Fig. 13.2. The nodes are separated by a distance $\lambda/2$, and at the midpoint between two consecutive nodes, i.e., at

$$x = \frac{\lambda}{4}, \frac{3\lambda}{4}, \frac{5\lambda}{4}, \ldots$$

the amplitude of the vibration is maximum. The displacements at these points (which are knowns as *antinodes*) are given by

$$y = \pm 2a \cos 2\pi\nu t \quad \textbf{(9)}$$

At the antinodes the kinetic energy density is given by (see Sec. 7.2)

$$\text{Kinetic energy/unit length} = \frac{1}{2}\rho(2a)^2\omega^2\cos^2\omega t$$

$$= 2\rho a^2\omega^2\cos^2\omega t \quad \textbf{(10)}$$

where $\omega = 2\pi\nu$ is the angular frequency and ρ the mass per unit length of the string.

We can also carry out a similar experiment for electromagnetic waves. In Fig. 13.3, T represents a transmitter of electromagnetic waves (the wavelength of which may be of the order of few centimeters); R represents a reflector which may be a highly polished metal surface, and D represents the detector which can measure the variation of the intensity of the electromagnetic waves at different points. One may approximately assume plane waves to be incident on the reflector; the incident and reflected waves interfere and produce nodes and antinodes. The result of a typical experiment is shown in Fig. 13.4. One can see the periodic variation of intensity. Two consecutive maxima are separated by about 5.8 cm; thus $\lambda \approx 13.6$ cm. The corresponding frequency ($\approx 2.6 \times 10^9$ s^{-1}) can be easily generated in the laboratory. If the frequency is changed, one can observe the change in the

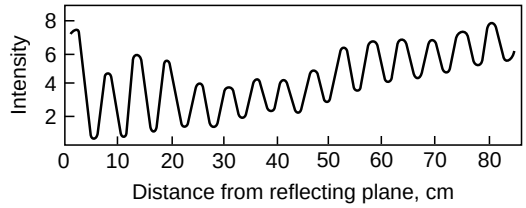


Fig. 13.4 A typical variation of the intensity between the reflector and the transmitter [Adapted from Ref. 2].

distance between the antinodes. One should notice that the minima do not really correspond to zero intensity and that the intensities at the maxima are not constant. This is so because the incident wave is really not a plane wave[2] and the reflection is not really perfect. In fact, one can introduce a coefficient of reflection r which is defined as the ratio of the energy of the reflected beam to the energy of the incident beam. Thus the ratio of the amplitudes is $\sqrt{r}$, and if the incident wave is given by

$$E_{\text{incident}} = a \sin\left[2\pi\left(\frac{x}{\lambda} + \nu t\right)\right] \tag{11}$$

then the reflected wave is given by

$$E_{\text{reflected}} = a\sqrt{r} \sin\left[2\pi\left(\frac{x}{\lambda} - \nu t\right)\right] \tag{12}$$

where the plane $x = 0$ corresponds to the plane of the reflector. Here E represents the electric field associated with the electromagnetic wave. Thus the resultant field is given by

$$\begin{aligned} E_{\text{resultant}} &= E_{\text{incident}} + E_{\text{reflected}} \\ &= a \sin\left[2\pi\left(\frac{x}{\lambda} + \nu t\right)\right] + a\sqrt{r} \sin\left[2\pi\left(\frac{x}{\lambda} - \nu t\right)\right] \\ &= a\sqrt{r}\left\{\sin\left[2\pi\left(\frac{x}{\lambda} + \nu t\right)\right] + \sin\left[2\pi\left(\frac{x}{\lambda} - \nu t\right)\right]\right\} \\ &\quad + a(1 - \sqrt{r}) \sin\left[2\pi\left(\frac{x}{\lambda} + \nu t\right)\right] \\ &= 2a\sqrt{r} \sin\left(\frac{2\pi}{\lambda} x\right) \cos 2\pi\nu t \\ &\quad + a(1 - \sqrt{r}) \sin\left[2\pi\left(\frac{x}{\lambda} + \nu t\right)\right] \end{aligned} \tag{13}$$

The first term represents the stationary component of the wave, and the second term (which is small if r is close to unity) represents the progressive part of the beam.

13.3 STATIONARY WAVES ON A STRING WHOSE ENDS ARE FIXED

In Sec. 13.2, while discussing the stationary waves on a string we assumed only one end of the string ($x = 0$) to be fixed; and the resultant displacement was given by [see Eq. (6)]

$$y = 2a \sin\left(\frac{2\pi}{\lambda} x\right) \cos 2\pi\nu t \tag{14}$$

If the other end of the string (say, at $x = L$) is also fixed, then we must have

$$2a \sin\left(\frac{2\pi}{\lambda} L\right) \cos 2\pi\nu t = 0 \tag{15}$$

Equation (15) is to be valid at all times; therefore,

$$\sin\left(\frac{2\pi}{\lambda} L\right) = 0 = \sin n\pi \tag{16}$$

or

$$\lambda = \lambda_n = \frac{2L}{n} \qquad n = 1, 2, 3, \ldots \tag{17}$$

The corresponding frequencies are

$$\nu_n = \frac{v}{\lambda_n} = \frac{nv}{2L} \qquad n = 1, 2, 3, \ldots \tag{18}$$

Thus, if a string of length L is clamped at both ends (as in a sonometer wire), then it can only vibrate with certain well-defined wavelengths. When $\lambda = 2L$ (i.e., $n = 1$), the string is said to vibrate in its fundamental mode [Fig. 13.5(a)]. Similarly when $\lambda = 2L/2$ and $2L/3$, the string is said to vibrate in its first and second harmonic. In general, if the string is plucked and then made to vibrate, the displacement is given by

$$y(x, t) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{2\pi}{\lambda_n} x\right) \cos\left(2\pi\, \nu_n t + \phi_n\right) \tag{19}$$

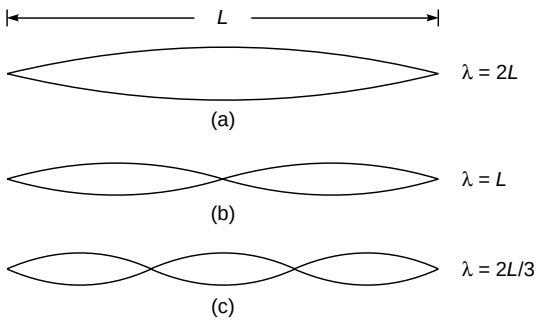


Fig. 13.5 Standing waves on a stretched string clamped at both ends.

[2] A plane wave is obtained by a point source at a very large distance from the point of observation (see Chap. 11).

where the constants a_n and ϕ_n are determined by the values of $y(x, t = 0)$ and $\partial y/\partial t\big|_{t=0}$; these are known as the initial conditions. A more detailed discussion of the vibration of stretched strings is given in Sec. 8.2.

When a string is vibrating in a particular mode, there is no net transfer of energy although each element of the string is associated with a certain energy density [see Eq. (10)]. The energy density is maximum at the antinodes and minimum at nodes. The distances between two successive antinodes and successive nodes are $\lambda/2$.

13.4 STATIONARY LIGHT WAVES: IVES' AND WIENER'S EXPERIMENTS

It is difficult to carry out experiments in which one obtains stationary light waves. This is so because light wavelengths are extremely small ($\approx 5 \times 10^{-5}$ cm). In the experimental arrangement of Ives, the emulsion side of a photographic plate was placed in contact with a film of mercury as shown in Fig. 13.6. A parallel beam of monochromatic light was allowed to fall normally on the glass plate. The beam was reflected on the mercury surface, and the incident wave interfered with the reflected wave, forming standing waves. A section of the photographic film was cut along a plane normal to the surface. The cut section was viewed under a microscope, and bright and dark bands (separated by regular intervals)) were observed. By measuring the distance between two consecutive dark bands (which is equal to $\lambda/2$) one can calculate the wavelength.

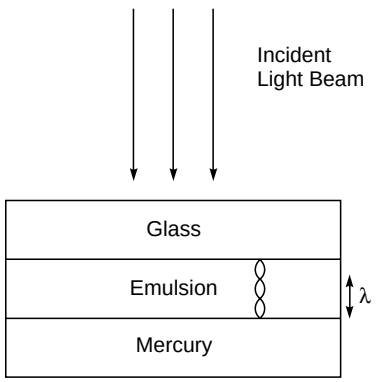


Fig. 13.6 The experimental arrangement of Ives for studying stationary light waves.

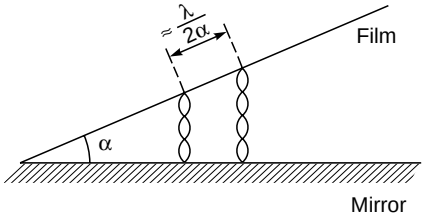


Fig. 13.7 The experimental arrangement of Wiener for studying stationary light waves.

Because of the small wavelength of light, the distance between two consecutive dark (or bright) bands was extremely small and was, therefore, difficult to measure. Wiener overcame this difficulty by placing the photographic film at a small angle and thereby increasing considerably the distance between the dark (or bright) bands (Fig. 13.7).

Example 13.1 In a typical experimental arrangement of Wiener, the angle between the film and the mirror was about 10^{-3} rad. For $\lambda = 5 \times 10^{-5}$ cm what is the distance between two consecutive dark bands?

Solution: The required distance is

$$\frac{\lambda}{2\alpha} = \frac{5 \times 10^{-5}}{2 \times 10^{-3}} \text{ cm} = 0.25 \text{ mm}$$

On the other hand, in the setup of Ives the distance is 2.5×10^{-4} mm.

13.5 SUPERPOSITION OF TWO SINUSOIDAL WAVES

Let us consider the superposition of two sinusoidal waves (having the same frequency) at a particular point. Let

$$\text{and} \qquad \begin{aligned} x_1(t) &= a_1 \cos(\omega t + \theta_1) \\ x_2(t) &= a_2 \cos(\omega t + \theta_2) \end{aligned} \qquad \textbf{(20)}$$

represent the displacements produced by each of the disturbances: we are assuming that the displacements are in the same direction.[3] However, they may have different amplitudes and different initial phases. In Sec. 17.5 we will consider the superposition of waves having nearly equal frequencies which leads to the phenomenon of beats. Now, according to the superposition principle, the resultant displacement $x(t)$ is given by

$$\begin{aligned} x(t) &= x_1(t) + x_2(t) \\ &= a_1 \cos(\omega t + \theta_1) + a_2 \cos(\omega t + \theta_2) \end{aligned} \qquad \textbf{(21)}$$

[3] Indeed in Sec. 13.2, while discussing stationary waves on a string, we had, at a particular value of x, two sinusoidal waves of the same frequency (but having different initial phases) superposing on each other. However, in general, one could have superposition of displacements which are in different directions, for example, the superposition of two linearly polarized waves to produce a circularly polarized wave (see Chap. 22).

which can be written in the form

$$x(t) = a \cos(\omega t + \theta) \quad \textbf{(22)}$$

where

$$a \cos\theta = a_1 \cos\theta_1 + a_2 \cos\theta_2 \quad \textbf{(23)}$$

and

$$a \sin\theta = a_1 \sin\theta_1 + a_2 \sin\theta_2 \quad \textbf{(24)}$$

Thus the resultant disturbance is also simple harmonic in character having the same frequency but different amplitude and different initial phase. If we square and add Eqs. (23) and (24), we obtain

$$a = [a_1^2 + a_2^2 + 2a_1a_2 \cos(\theta_1 - \theta_2)]^{1/2} \quad \textbf{(25)}$$

Further

$$\tan\theta = \frac{a_1 \sin\theta_1 + a_2 \sin\theta_2}{a_1 \cos\theta_1 + a_2 \cos\theta_2} \quad \textbf{(26)}$$

The angle θ is not uniquely determined from Eq. (26); however, if we assume a to be always positive, then $\cos\theta$ and $\sin\theta$ can be determined from Eqs. (23) and (24) which will uniquely determine θ.

From Eq. (25) we find that if

$$\theta_1 \sim \theta_2 = 0, 2\pi, 4\pi, \ldots \quad \textbf{(27)}$$

then $$a = a_1 + a_2 \quad \textbf{(28)}$$

Thus, if the two displacements are in phase, then the resultant amplitude will be the sum of the two amplitudes; this is known as *constructive interference*. Similarly, if

$$\theta_1 \sim \theta_2 = \pi, 3\pi, 5\pi, \ldots \quad \textbf{(29)}$$

then

$$a = a_1 \sim a_2 \quad \textbf{(30)}$$

and the resultant amplitude is the difference of the two amplitudes. This is known as *destructive interference*. If we refer to Fig. 13.2, then we can see that constructive interference occurs at $x = \lambda/4, 3\lambda/4, 5\lambda/4, \ldots$ (i.e., at the points P', Q', R', ...) and destructive interference occurs at $x = 0, \lambda/2, \lambda, 3\lambda/2, \ldots$ (i.e., at points $A, P, Q, R, \ldots$). When constructive and destructive interferences occur, there is no violation of the principle of conservation of energy; the energy is just redistributed.

In general, if we have n displacements

$$\begin{aligned} x_1 &= a_1 \cos(\omega t + \theta_1) \\ x_2 &= a_2 \cos(\omega t + \theta_2) \\ &\cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots \\ x_n &= a_n \cos(\omega t + \theta_n) \end{aligned} \quad \textbf{(31)}$$

then

$$x = x_1 + x_2 + \cdots + x_n = a \cos(\omega t + \theta) \quad \textbf{(32)}$$

where

$$a \cos\theta = a_1 \cos\theta_1 + \cdots + a_n \cos\theta_n \quad \textbf{(33)}$$

and

$$a \sin\theta = a_1 \sin\theta_1 + \cdots + a_n \sin\theta_n \quad \textbf{(34)}$$

13.6 THE GRAPHICAL METHOD FOR STUDYING SUPERPOSITION OF SINUSOIDAL WAVES

In this section we will discuss the graphical method for adding displacements of the same frequency. This method is particularly useful when we have a large number of superposing waves, as indeed happens when we consider the phenomenon of diffraction.

Let us first try to obtain the resultant of the two displacements given by Eq. (20), using the graphical method. We draw a circle of radius a_1 and let point P on the circle be such that OP makes an angle θ_1 with the x axis[4] (see Fig. 13.8). We next draw a circle of radius a_2 and let point Q on the circle be such that OQ makes an angle θ_2 with the x axis. We use the law of parallelograms to find the resultant $\overrightarrow{OR}$ of the vectors $\overrightarrow{OP}$ and $\overrightarrow{OQ}$. The length of the vector $\overrightarrow{OR}$ will represent the amplitude of the resultant displacement, and if θ is the angle that OR makes with the x axis, then the initial phase of the resultant will be θ. This can be easily seen by noting that

$$\begin{aligned} OR \cos\theta &= OP \cos\theta_1 + PR \cos\theta_2 \\ &= a_1 \cos\theta_1 + a_2 \cos\theta_2 \end{aligned} \quad \textbf{(35)}$$

Similarly,

$$OR \sin\theta = a_1 \sin\theta_1 + a_2 \sin\theta_2 \quad \textbf{(36)}$$

consistent with Eqs. (23) and (24). Further, as vectors $\overrightarrow{OP}$ and $\overrightarrow{OQ}$ rotate on the circumference of the circles of radii a_1

[4] Clearly, if we assume vector $\overrightarrow{OP}$ to rotate (in the counterclockwise direction) with angular velocity ω, then the x coordinate of vector $\overrightarrow{OP}$ will be $a_1 \cos(\omega t + \theta_1)$, where $t = 0$ corresponds to the instant when the rotating vector is at point P.

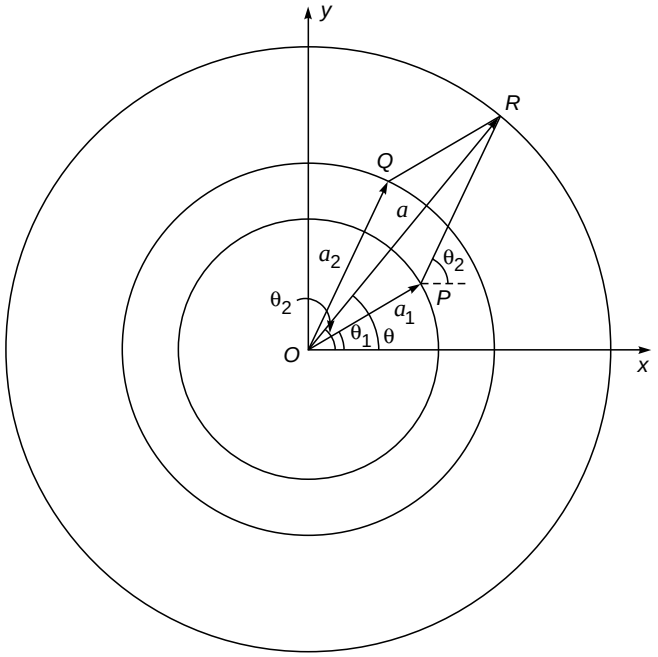


Fig. 13.8 The graphical method for determining the resultant of two simple harmonic motions along the same direction and having the same frequency.

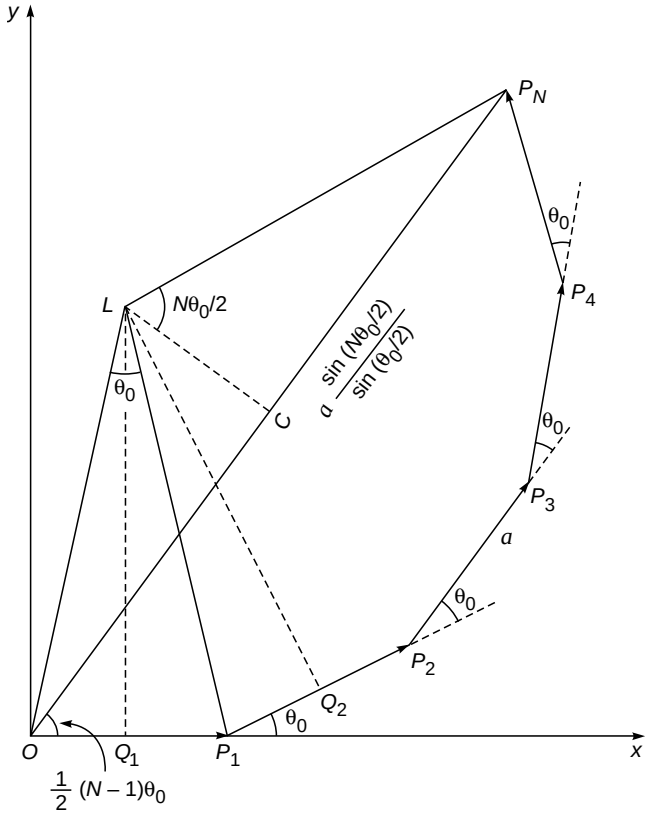


Fig. 13.9 The graphical method for determining the resultant of N simple harmonic motions along the same direction and having the same frequency.

and a_2, vector $\overrightarrow{OR}$ rotates on the circumference of the circle of radius OR with the same frequency.

Thus, if we wish to find the resultant of the two displacements given by Eq. (20), then we must first draw a vector ($\overrightarrow{OP}$) of length a_1 making an angle θ_1 with the axis; from the tip of this vector we must draw another vector ($\overrightarrow{PR}$) of length a_2 making an angle θ_2 with the axis. The length of vector $\overrightarrow{OR}$ will represent the resultant amplitude, and the angle that it makes with the axis will represent the initial phase of the resultant displacement. It can be easily seen that if we have a third displacement

$$x_3 = a_3 \cos(\omega t + \theta_3) \qquad \textbf{(37)}$$

then from point R we must draw a vector $\overrightarrow{RR'}$ of length a_3 which makes an angle θ_3 with the axis; vector $\overrightarrow{OR'}$ will represent the resultant of x_1, x_2, and x_3.

As an illustration of the above procedure, we consider the resultant of N simple harmonic motions all having the same amplitude and with their phases increasing in arithmetic progression. Thus

$$\begin{aligned} x_1 &= a\cos\omega t \\ x_2 &= a\cos(\omega t + \theta_0) \qquad \textbf{(38)} \\ &\cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots \\ x_N &= a\cos[\omega t + (N-1)\theta_0] \end{aligned}$$

In Fig. 13.9 vectors $\overrightarrow{OP_1}$, $\overrightarrow{P_1P_2}$, $\overrightarrow{P_2P_3}$, ... correspond to x_1, x_2, x_3, ... , respectively. The resultant is denoted by vector $\overrightarrow{OP_N}$. Let Q_1L and Q_2L be the perpendicular bisectors of OP_1 and P_1P_2. It is easy to prove that

$$\Delta LQ_1P_1 \equiv \Delta LQ_2P_1$$

Thus $LO = LP_1 = LP_2$. Therefore, points $O, P_1, P_2, P_3, \ldots, P_N$ will lie on the circumference of a circle whose center is L and whose radius is LO. Further, $\angle LP_1O = \pi - \theta_0/2$, and therefore $\angle OLP_1 = \theta_0$. Thus,

$$LO = \frac{a/2}{\sin(\theta_0/2)}$$

and

$$\begin{aligned} OP_N &= 2OC = 2LO\sin\frac{N\theta_0}{2} \\ &= a\frac{\sin N\theta_0/2}{\sin\theta_0/2} \end{aligned} \qquad \textbf{(39)}$$

Further, the phase of the resultant displacement is

$$\angle P_N OX = \frac{1}{2}(N-1)\theta_0$$

Thus

$$a\cos\omega t + a\cos(\omega t+\theta_0) + \cdots + a\cos[\omega t + (N-1)\theta_0] = A\cos(\omega t+\theta) \quad \textbf{(40)}$$

where

$$A = \frac{a\sin N\theta_0/2}{\sin\theta_0/2} \quad \textbf{(41)}$$

and

$$\theta = \frac{1}{2}(N-1)\theta_0 \quad \textbf{(42)}$$

We will use this result in Chap. 18.

13.7 THE COMPLEX REPRESENTATION

Often it is more convenient to use the complex representation in which the displacement

$$x_1 = a_1\cos(\omega t+\theta_1) \quad \textbf{(43)}$$

is written as

$$x_1 = a_1 e^{i(\omega t+\theta_1)} \quad \textbf{(44)}$$

where it is implied that the actual displacement is the real part of x_1. Further, if

$$x_2 = a_2 e^{i(\omega t+\theta_2)}$$

then

$$x_1 + x_2 = (a_1 e^{i\theta_1} + a_2 e^{i\theta_2})e^{i\omega t} = ae^{i(\omega t+\theta)} \quad \textbf{(45)}$$

where

$$ae^{i\theta} = a_1 e^{i\theta_1} + a_2 e^{i\theta_2} \quad \textbf{(46)}$$

If we equate the real and imaginary parts of Eq. (46), we obtain Eqs. (23) and (24).

An interesting illustration of the usefulness of this method is to consider the resultant of the N displacements described by Eq. (38). Thus we write

$$x_1 = ae^{i\omega t}, \quad x_2 = ae^{i(\omega t+\theta_0)}, \ldots$$

Hence

$$\begin{aligned} x &= x_1 + x_2 + \cdots \\ &= ae^{i\omega t}(1 + e^{i\theta_0} + e^{2i\theta_0} + \cdots + e^{i(N-1)\theta_0}) \\ &= ae^{i\omega t}\frac{1-e^{Ni\theta_0}}{1-e^{i\theta_0}} \\ &= ae^{i\omega t}\frac{e^{iN\theta_0/2}}{e^{i\theta_0/2}}\cdot\frac{e^{iN\theta_0/2}-e^{-iN\theta_0/2}}{e^{i\theta_0/2}-e^{-i\theta_0/2}} \\ &= \frac{a\sin(N\theta_0/2)}{\sin(\theta_0/2)}\exp\left\{i\left[\omega t + (N-1)\frac{\theta_0}{2}\right]\right\} \quad \textbf{(47)} \end{aligned}$$

which is consistent with Eq. (40). The complex representation is also very useful in considering the spreading of a wave packet (see Sec. 10.3).

Note that whereas

$$\mathrm{Re}(x_1) + \mathrm{Re}(x_2) = \mathrm{Re}(x_1 + x_2)$$
$$(\mathrm{Re}\,x_1)(\mathrm{Re}\,x_2) \neq \mathrm{Re}(x_1 x_2)$$

where Re (. . .) denotes the "real part" of the quantity inside the parentheses. Thus, one must be careful in calculating the intensity of a wave which is proportional to the square of the amplitude. While using the complex representation, one must calculate first the amplitude and then the intensity.

Summary

- According to the principle of superposition of waves, the resultant displacement (at a particular point) produced by a number of waves is the vector sum of the displacements produced by each one of the disturbances.
- The stationary waves on a string and the formation of standing electromagnetic waves are formed by the superposition of waves traveling in opposite directions.
- If the two displacements (produced by two sinusoidal waves) are in phase, then the resultant amplitude is the sum of the two amplitudes; this is known as constructive interference. On the other hand, if the two displacements are π out of phase, then the resultant amplitude is the difference of the two amplitudes; this is known as destructive interference.

Problems

13.1 Standing waves are formed on a stretched string under tension of 1 N. The length of the string is 30 cm, and it vibrates in three loops. If the mass per unit length of the wire is 10 mg cm^{-1}, calculate the frequency of the vibrations.

13.2 In Prob. 13.1, if the string is made to vibrate in its fundamental mode, what will be the frequency of vibration?

13.3 In the experimental arrangement of Wiener, what should be the angle between the film and the mirror if the distance between two consecutive dark bands is 7×10^{-3} cm^2. Assume $\lambda = 6 \times 10^{-5}$ cm.

[**Ans.:** ~1/4°]

13.4 Standing waves with five loops are produced on a stretched string under tension. The length of the string is 50 cm, and

the frequency of vibrations is 250 s^{-1}. Calculate the time variation of the displacement of the points which are at distances of 2, 5, 15, 18, 20, 35, and 45 cm from one end of the string.

13.5 The displacements associated with two waves (propagating in the same direction) having same amplitude but slightly different frequencies can be written in the form

$$a \cos 2\pi\left(\nu t - \frac{x}{\lambda}\right)$$

and

$$a \cos 2\pi\left[(\nu + \Delta\nu)t - \frac{x}{\lambda - \Delta\lambda}\right]$$

(Such displacements are indeed obtained when we have two tuning forks with slightly different frequencies.) Discuss the superposition of the displacements, and show that at a particular value of x, the intensity will vary with time.

13.6 In Prob. 13.5 assume v = 330 m s^{-1}, ν = 256 s^{-1}, $\Delta\nu$ = 2 s^{-1}, and a = 0.1 cm. Plot the time variation of the intensity at $x = 0$, $\lambda/4$, and $\lambda/2$.

13.7 Use the complex representation to study the time variation of the resultant displacement at $x = 0$ in Prob.13.5 and 13.6.

13.8 Discuss the superposition of two plane waves (of the same frequency and propagating in the same direction) as a function of the phase difference between them. (Such a situation indeed arises when a plane wave gets reflected at the upper and lower surfaces of a glass slab; see Sec. 15.2.)

13.9 In Example 11.1 we discussed the propagation of a semicircular pulse on a string. Consider two semicircular pulses propagating in opposite directions. At $t = 0$, the displacements associated with the pulses propagating in the $+x$ and in the $-x$ directions are given by

$$(R^2 - x^2)^{1/2} \qquad \text{and} \qquad -[R^2 - (x - 10R)^2]^{1/2}$$

respectively. Plot the resultant disturbance at $t = R/v$, $2.5R/v$, $5R/v$, $7.5R/v$, and $10R/v$, where v denotes the speed of propagation of the wave.

REFERENCES AND SUGGESTED READINGS

See those at the end of Chap. 14.

Chapter Fourteen

TWO BEAM INTERFERENCE BY DIVISION OF WAVE FRONT

'The wave nature of light was demonstrated convincingly for the first time in 1801 by Thomas Young by a wonderfully simple experiment . . . He let a ray of sunlight into a dark room, placed a dark screen in front of it, pierced with two small pinholes, and beyond this, at some distance a white screen. He then saw two darkish lines at both sides of a bright line, which gave him sufficient encouragement to repeat the experiment, this time with spirit flame as light source, with a little salt in it, to produce the bright yellow sodium light. This time he saw a number of dark lines, regularly spaced; the first clear proof that light added to light can produce darkness. This phenomenon is called interference. Thomas Young had expected it because he believed in the wave theory of light.

—Dennis Gabor in his Nobel Lecture, December 11, 1971

Thomas Young had amazing broad interests and talents . . . From his discoveries in medicine and science, Helmholtz concluded: 'His was one of the most profound minds that the world has ever seen.'

—From the Internet

14.1 INTRODUCTION

In Chap.13, we had considered the superposition of one-dimensional waves propagating on a string and showed that there is a variation of energy density along the length of the string due to the interference of two waves (see Fig. 13.5). In general, whenever two waves superpose, one obtains an intensity distribution which is known as the interference pattern. In this chapter, we will consider the interference pattern produced by waves emanating from two point sources. We may note that with sound waves the interference pattern can be observed without much difficulty because the two interfering waves maintain a constant phase relationship; this is also the case for microwaves. However, for light waves, due to the very process of emission, one cannot observe interference between the waves from two independent sources,[1] although the interference does take place (see Sec. 14.4). Thus, one tries to derive interfering waves from a single wave so that the phase relationship is maintained. The methods to achieve this can be classified under two broad categories. Under the first category, in a typical arrangement, a beam is allowed to fall on two closely spaced holes, and the two beams emanating from the holes, interfere. This method is known as division of wave front and will be discussed in detail in this chapter. In the other method, known as division of amplitude, a beam is divided at two or more reflecting surfaces and the reflected beams interfere. This will be discussed in Chap.15. We must, however, emphasize that the present and the following chapters are based on one underlying principle, namely, the superposition principle.

It is also possible to observe interference by using multiple-beams; this is known as multiple-beam interferometry and will be discussed in Chap. 16. It will be shown that multiple beam interferometry offers some unique advantages over two-beam interferometry.

[1] It is difficult to observe the interference pattern even with two laser beams unless they are phase-locked.

14.2 INTERFERENCE PATTERN PRODUCED ON THE SURFACE OF WATER

We consider surface waves emanating from two point sources in a water tank. We may have, for example, two sharp needles vibrating up and down at points S_1 and S_2 (see Fig. 14.1). Although water waves are not really transverse, we will, for the sake of simplicity, assume water waves to produce displacements which are transverse to the direction of propagation.

If there were only one needle (say, at S_1) vibrating with a certain frequency ν, then circular ripples would have spread out from point S_1. The wavelength would have been v/ν, and the crests and troughs would have moved outward. Similarly for the vibrating needle at S_2. However, if both needles are vibrating, then waves emanating from S_1 will interfere with the waves emanating from S_2. We assume that the needle at S_2 vibrates in phase with the needle at S_1; i.e., S_1 and S_2 go up simultaneously, and they also reach the lowest position at the same time. Thus, if at a certain instant, the disturbance emanating from the source S_1 produced a crest at a distance ρ from S_1 then the disturbance from S_2 would also produce a crest at a distance ρ from S_2, etc. This is explicitly shown in Fig. 14.1, where the solid curves represent (at a particular instant) the positions of the crests due to disturbances emanating from S_1 and S_2. Similarly, the dashed curves represent (at the same instant) the positions of the troughs. Notice that at all points on the perpendicular bisector OY, the disturbances reaching from S_1 and from S_2 will always be in phase. Consequently, at an arbitrary point A (on the perpendicular bisector) we may write the resultant disturbance as

$$y = y_1 + y_2 = 2a\cos\omega t \qquad \textbf{(1)}$$

where y_1 (= a cos ωt) and y_2 (= a cos ωt) represent the displacements at point A due to S_1 and S_2, respectively. We see that the amplitude at A is twice the amplitude produced by each one of the sources. At $t = T/4$ $(= 1/4\nu = \pi/2\omega)$ the displacements produced at point A by each of the sources will be zero, and the resultant will also be zero. This is also obvious from Eq. (1).

Next, let us consider a point B such that

$$S_2B - S_1B = \lambda/2 \qquad \textbf{(2)}$$

At such a point the disturbance reaching from source S_1 will always be out of phase with the disturbance reaching from S_2. This follows from the fact that the disturbance reaching point B from source S_2 must have started one-half of a period (= $T/2$) earlier than the disturbance reaching B from S_1. Consequently, if the displacement at B due to S_1 is given by

$$y_1 = a\cos\omega t$$

then the displacement at B due to S_2 is given by

$$y_2 = a\cos(\omega t - \pi) = -a\cos\omega t$$

and the resultant $y = y_1 + y_2$ is zero at all times. Such a point corresponds to destructive interference and is known as a node and corresponds to minimum intensity. The amplitudes of the two vibrations reaching the point B will not really be equal, as it is at different distances from S_1 and S_2. However, if the distances involved are large (in comparison to the wavelength), the two amplitudes will be very nearly equal and the resultant intensity will be very nearly zero.

In a similar manner we may consider a point C such that

$$S_2C - S_1C = \lambda$$

where the phases of the vibration (reaching from S_1 and S_2) are exactly the same as at point A. Consequently we will again have constructive interference. In general, if a point P is such that

$$S_2P \sim S_1P = n\lambda \qquad \text{(maxima)} \qquad \textbf{(3)}$$

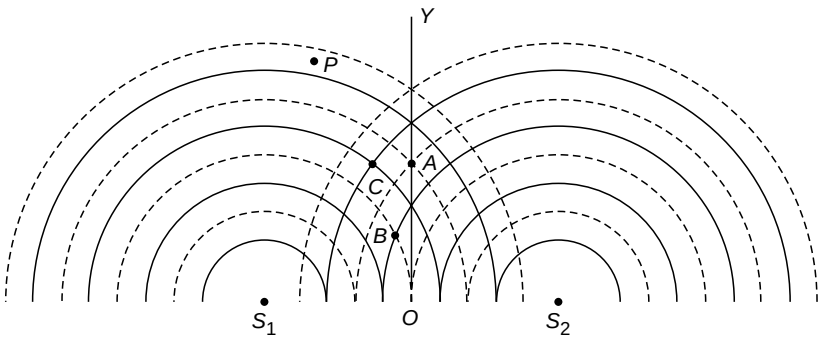


Fig. 14.1 Waves emanating from two point sources S_1 and S_2 vibrating in phase. The solid and the dashed curves represent the positions of the crests and troughs, respectively.

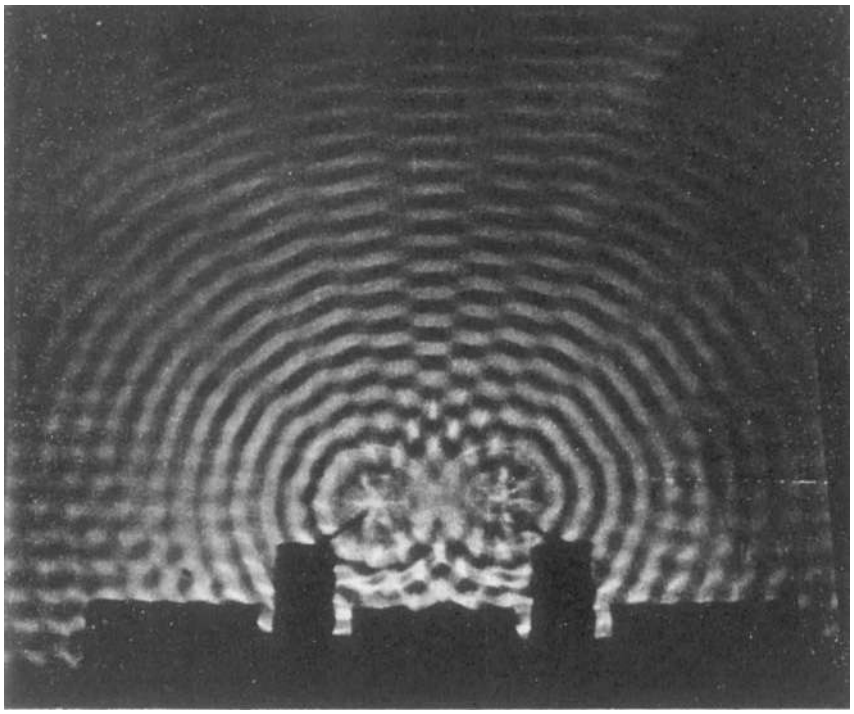


Fig. 14.2 The actual interference pattern produced from two point sources vibrating in phase in a ripple tank (After Ref. 9, used with permission).

$n = 0, 1, 2, \ldots$, then the disturbances reaching point P from the two sources will be in phase, the interference will be constructive, and the intensity will be maximum. On the other hand, if point P is such that

$$S_2P \sim S_1P = \left(n + \frac{1}{2}\right)\lambda \qquad \text{(minima)} \qquad \textbf{(4)}$$

then the disturbances reaching point P from the two sources will be out of phase, the interference will be destructive, and the intensity will be minimum. The actual interference pattern produced from two point sources vibrating in phase in a ripple tank is shown in Fig. 14.2.

Example 14.1 The intensity at the point which statisfies neither Eq. (3) nor Eq. (4) will not be a maximum or zero. Consider a point P such that $S_2P - S_1P = \lambda/3$. Find the ratio of the intensity at point P to that at a maximum.

Solution: If the disturbance reaching point P from S_1 is given by

$$y_1 = a \cos \omega t$$

then the disturbance from S_2 is given by

$$y_2 = a \cos \left(\omega t - \frac{2\pi}{3}\right)$$

because a path difference of $\lambda/3$ corresponds to a phase difference of $2\pi/3$.

Thus the resultant displacement is

$$y = y_1 + y_2$$

$$= a \left[\cos \omega t + \cos\left(\omega t - \frac{2\pi}{3}\right)\right]$$

$$= 2a \cos \left(\omega t - \frac{\pi}{3}\right) \cos \frac{\pi}{3}$$

$$= a \cos \left(\omega t - \frac{\pi}{3}\right)$$

The intensity is therefore one-fourth of the intensity at the maxima. In a similar manner one can calculate the intensity at any other point.

Example 14.2 The locus of points which correspond to minima is known as nodal lines. Show that the equation of a nodal line is a hyperbola. Also obtain the locus of points which correspond to maxima.

Solution: For the sake of generality we find the locus of point P which satisfies the following equation:

$$S_1P - S_2P = \Delta \qquad \textbf{(5)}$$

Thus, if $\Delta = n\lambda$, we have a maximum; and if $\Delta = \left(n + \frac{1}{2}\right)\lambda$, we have a minimum. We choose the midpoint of S_1S_2 as the origin,

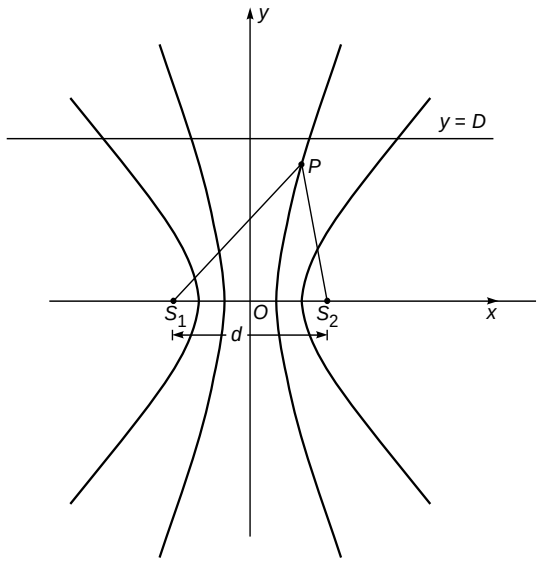


Fig. 14.3 The nodal curves.

with the x axis along S_1S_2 and the y axis perpendicular to it (see Fig. 14.3). If the distance between S_1 and S_2 is d, then the coordinates of points S_1 and S_2 are $(-d/2, 0)$ and $(+d/2, 0)$ respectively. Let the coordinates of the point P be (x, y). Then

$$S_1P = \left[\left(x+\frac{d}{2}\right)^2 + y^2\right]^{1/2}$$

and

$$S_2P = \left[\left(x-\frac{d}{2}\right)^2 + y^2\right]^{1/2}$$

Therefore,

$$S_1P - S_2P = \left[\left(x+\frac{d}{2}\right)^2 + y^2\right]^{1/2} - \left[\left(x-\frac{d}{2}\right)^2 + y^2\right]^{1/2} = \Delta$$

or

$$\left(x+\frac{d}{2}\right)^2 + y^2 = \Delta^2 + \left(x-\frac{d}{2}\right)^2 + y^2 + 2\Delta\left[\left(x-\frac{d}{2}\right)^2 + y^2\right]^{1/2}$$

or

$$2xd - \Delta^2 = 2\Delta\left[\left(x-\frac{d}{2}\right)^2 + y^2\right]^{1/2}$$

On squaring, we obtain

$$4x^2d^2 - 4xd\Delta^2 + \Delta^4 = 4\Delta^2\left(x^2 - xd + \frac{d^2}{4} + y^2\right)$$

Thus we obtain

$$\frac{x^2}{\frac{1}{4}\Delta^2} - \frac{y^2}{\frac{1}{4}(d^2-\Delta^2)} = 1 \qquad \textbf{(6)}$$

which is the equation of a hyperbola. When $\Delta = \left(n+\frac{1}{2}\right)\lambda$, the curves correspond to minima; and when $\Delta = n\lambda$, the curves correspond to maxima. For large values of x and y, the curves asymptotically tend to the straight lines

$$y = \pm\left(\frac{d^2-\Delta^2}{\Delta^2}\right)^{1/2} x \qquad \textbf{(7)}$$

There is no point P for which $S_1P \sim S_2P > d$ ($S_1P \sim S_2P$ equals d on the x axis only). Now, it appears from Eq. (6) that when $\Delta > d$, the resulting equation is an ellipse, which we know is impossible. The fallacy is a result of the fact that because of a few squaring operations, Eq. (6) also represents the locus of all those points for which $S_1P + S_2P = \Delta$, and obviously in this case Δ can exceed d.

Example 14.3 Consider a line parallel to the x axis at a distance D from the origin (see Fig. 14.3). Assume $D >>> \lambda$. Find the points on this line where minimum intensity will occur.

Solution: The equation of this line is

$$y = D \qquad \textbf{(8)}$$

Further, at large distances from the origin the equation of the nodal lines is

$$y = \pm\left(\frac{d^2-\Delta_n^2}{\Delta_n^2}\right)^{1/2} x \qquad \textbf{(9)}$$

where $\Delta_n = \left(n+\frac{1}{2}\right)\lambda$; $n = 0, 1, 2, \ldots$ Clearly the points at which minima will occur (on the line $y = D$) are given by

$$x_n = \pm\left(\frac{\Delta_n^2}{d^2-\Delta_n^2}\right)^{1/2} D = \pm\frac{\Delta_n}{d}\left(1-\frac{\Delta_n^2}{d^2}\right)^{-1/2} D \approx \pm\left(n+\frac{1}{2}\right)\frac{\lambda D}{d} \qquad \textbf{(10)}$$

where we have assumed $\Delta_n << d$. Thus the points corresponding to minima will be equally spaced with a spacing of $\lambda D/d$.

Example 14.4 Until now we have assumed the needles at S_1 and S_2 (see Fig. 14.1) to vibrate in phase. Assume now that the needles vibrate with a phase difference of π, and obtain the nodal lines. Generalize the result for an arbitrary phase difference between the vibrations of the two needles.

Solution: The two needles S_1 and S_2 vibrate out of phase. Thus if, at any instant, the needle at S_1 produces a crest at a distance R

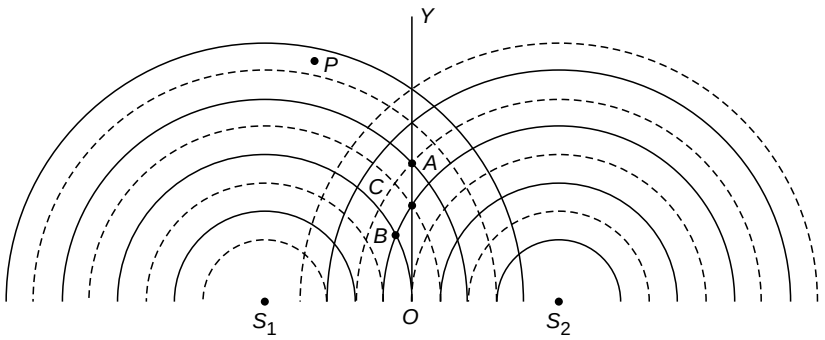


Fig. 14.4 Waves emanating from two point sources S_1 and S_2 vibrating out of phase.

from it, then the needle at S_2 produces a trough at a distance R from S_2. Therefore, at all points on the perpendicular bisector OY (see Fig. 14.4), the two vibrations will always be out of phase and we will have a minimum. On the other hand, at point B which satisfies the equation

$$S_2B - S_1B = \lambda/2$$

the two vibrations will be in phase, and we will have a maximum. Thus, because of the initial phase difference of π, the conditions for maxima and minima are reversed; i.e., when

$$S_2P \sim S_1P = \left(n + \frac{1}{2}\right)\lambda \qquad \text{(maxima)}$$

the interference will be constructive and we will have maxima, and when

$$S_2P \sim S_1P = n\lambda \qquad \text{(minima)}$$

the interference will be destructive and we will have minima. Notice that one again obtains a stationary interference pattern with nodal lines as hyperbolas.

The above analysis can be easily generalized for an arbitrary phase difference between the two needles. Assume, for example, that there is a phase difference of $\pi/3$; i.e., if there is a crest at a distance R from S_1, then there is a crest at a distance $R - \lambda/6$ from S_2. Consequently, the condition

$$S_1P - S_2P = n\lambda + \frac{\lambda}{6} \qquad n = 0, \pm1, \pm2, \ldots$$

will correspond to maxima.

14.3 COHERENCE

From the above examples we find that whenever the two needles vibrate with a constant phase difference, a stationary interference pattern is produced. The positions of the maxima and minima will, however, depend on the phase difference in the vibration of the two needles. Two sources which vibrate with a fixed phase difference between them are said to be *coherent.*

We next assume that the two needles are sometimes vibrating in phase, sometimes vibrating out of phase, sometimes vibrating with a phase difference of $\pi/3$, etc.; then the interference pattern will keep on changing. If the phase difference changes with such great rapidity that a stationary interference cannot be observed, then the sources are said to be *incoherent*.

Let the displacement produced by the sources at S_1 and S_2 be given by

$$\begin{aligned} y_1 &= a\cos\omega t \\ y_2 &= a\cos(\omega t + \phi) \end{aligned} \qquad \textbf{(11)}$$

Then the resultant displacement is

$$y = y_1 + y_2 = 2a\cos\frac{\phi}{2}\cos\left(\omega t + \frac{\phi}{2}\right) \qquad \textbf{(12)}$$

The intensity I which is proportional to the square of the amplitude can be written in the form

$$I = 4I_0\cos^2\frac{\phi}{2} \qquad \textbf{(13)}$$

where I_0 is the intensity produced by each one of the sources individually. Clearly if $\phi = \pm\pi, \pm3\pi, \ldots$, the resultant intensity will be zero and we will have minima. On the other hand, when $\phi = 0, \pm2\pi, \pm4\pi, \ldots$, the intensity will be maximum ($= 4I_0$). However, if the phase difference between sources S_1 and S_2 (i.e., ϕ) is changing with time, the observed intensity is given by

$$I = 4I_0\left\langle\cos^2\frac{\phi}{2}\right\rangle \qquad \textbf{(14)}$$

where $\langle\ldots\rangle$ denotes the time average of the quantity inside the angular brackets; the time average of a time-dependent function is defined by the relation

$$\langle f(t)\rangle = \frac{1}{\tau}\int_{-\tau/2}^{+\tau/2} f(t)\,dt \qquad \textbf{(15)}$$

where τ represents the time over which the averaging is carried out. For example, if the interference pattern is viewed by a normal eye, this averaging will be over about 0.1; for a camera with exposure time 0.001 s, $\tau = 0.001$ s, etc. Clearly, if ϕ varies in a random manner in times which are small compared to τ, then $\cos^2(\phi/2)$ will randomly vary between 0 and 1 and $\left\langle \cos^2(\phi/2) \right\rangle$ will be $\frac{1}{2}$ (see also Sec. 14.6). For such a case

$$I = 2I_0 \qquad \textbf{(16)}$$

which implies that if the sources are incoherent, then the resultant intensity is the sum of the two intensities and there is no variation of intensity! Thus, if one (or both) of the two vibrating sources is such that the phase difference between the vibrations of the two sources varies rapidly, then the interference phenomenon will not be observed. We will discuss this point again in Sec. 14.6 and in Chap. 17.

14.4 INTERFERENCE OF LIGHT WAVES

Until now we have considered interference of waves produced on the surface of water. We will now discuss the interference pattern produced by light waves; however, for light waves it is difficult to observe a stationary interference pattern. For example, if we use two conventional light sources (such as two sodium lamps) illuminating two pinholes (see Fig. 14.5), we will not observe any interference pattern on the screen. This can be understood from the following reasoning: In a conventional light source, light comes from a large number of independent atoms, each atom emitting light for about 10^{-10} s, i.e., light emitted by an atom is essentially a pulse lasting for only 10^{-10} s. However, since the optical frequencies are of the order of 10^{15} s^{-1}, such a short pulse consists of about 1 million oscillations; thus it is almost monochromatic (see Chap. 17). Even if the atoms were emitting under similar conditions, waves from different atoms would differ in their initial phases.

Consequently, light coming out from holes S_1 and S_2 will have a fixed phase relationship for about 10^{-10} s, hence the interference pattern will keep on changing every billionth of a second. The eye can notice intensity changes which last at least for 0.1s, and hence we will observe a uniform intensity over the screen. However, if we have a camera whose time of shutter opening can be made less than 10^{-10} s, then the film will record an interference pattern.[2] We summarize the above results by noting that light beams from two independent sources do not have any fixed relationship, as such, they do not produce any stationary interference pattern.

Thomas Young in 1801 devised an ingenious but simple method to lock the phase relationship between the two sources. The trick lies in the division of a single wave front into two; these two split wave fronts act as if they emanated from two sources having a fixed phase relationship, and therefore when these two waves were allowed to interfere, a stationary interference pattern was obtained. In the actual experiment a light source illuminates pinhole S (see Fig. 14.6). Light diverging from this pinhole fell on a barrier which contained two pinholes S_1 and S_2 that were very close to each other and were located equidistant from S. Spherical

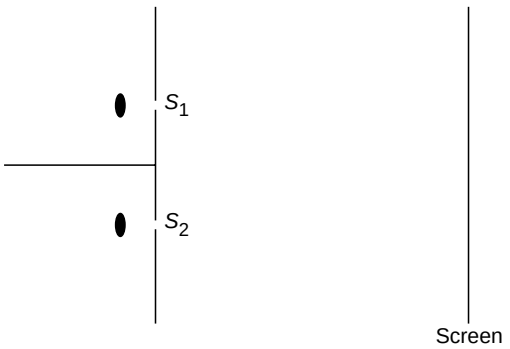


Fig. 14.5 If two sodium lamps illuminate two pinholes S_1 and S_2, no interference pattern will be observed on the screen.

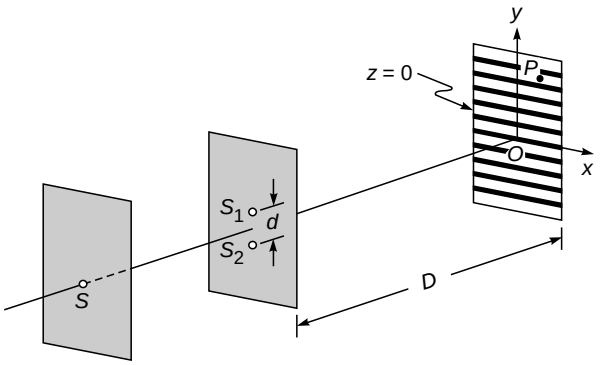


Fig. 14.6 Young's arrangement to produce interference pattern.

[2] This interference pattern will be a set of dark and bright bands only if the light waves have the same state of polarization. This can, however, be easily done by putting two Polaroids in front of S_1 and S_2 (see Fig. 14.5).

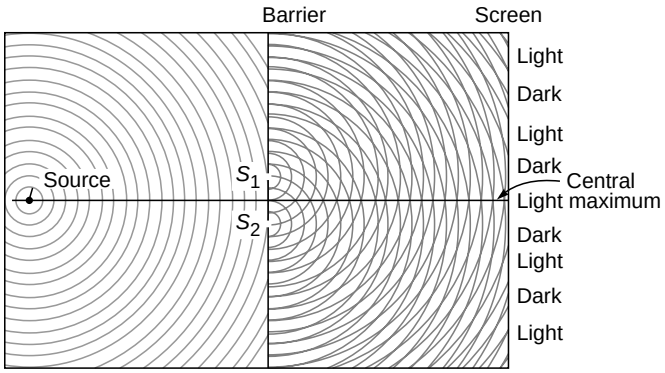


Fig. 14.7 Sections of the spherical wave fronts emanating from S, S_1, and S_2 (Adapted from Ref. 7; used with permission).

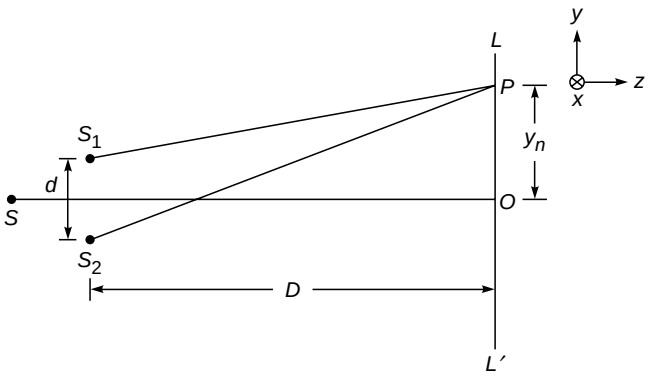


Fig. 14.8 Arrangement for producing Young's interference pattern.

waves emanating from S_1 and S_2 (see Fig. 14.7) were coherent, and on the screen beautiful interference fringes were obtained. To show that this was indeed an interference effect, Young showed that the fringes on the screen disappear when S_1 (or S_2) is covered up. Young explained the interference pattern by considering the principle of superposition, and by measuring the distance between the fringes he calculated the wavelength. Figure 14.7 shows the section of the wave front on the plane containing S, S_1, and S_2.

14.5 THE INTERFERENCE PATTERN

Let S_1 and S_2 represent the two pinholes of Young's interference experiment. We want to determine the positions of maxima and of minima on line LL' which is parallel to the y axis and lies in the plane containing points S, S_1, and S_2 (see Fig. 14.8). We will show that the interference pattern (around point O) consists of a series of dark and bright lines perpendicular to the plane of Fig. 14.8; point O is the foot of the perpendicular from point S on the screen.

For an arbitrary point P (on line LL') to correspond to a maximum, we must have

$$S_2P - S_1P = n\lambda \qquad n = 0, 1, 2, \ldots \tag{17}$$

Now,

$$(S_2P)^2 - (S_1P)^2 = \left[D^2 + \left(y_n + \frac{d}{2}\right)^2\right] - \left[D^2 + \left(y_n - \frac{d}{2}\right)^2\right] = 2y_nd$$

where

$$S_1S_2 = d \qquad \text{and} \qquad OP = y_n$$

Thus

$$S_2P - S_1P = \frac{2y_nd}{S_2P + S_1P} \tag{18}$$

If y_n, $d << D$, then negligible error will be introduced if $S_2P + S_1P$ is replaced by $2D$. For example, for $d = 0.02$ cm, D=50 cm, and $OP = 0.5$ cm (which corresponds to typical values for a light interference experiment)

$$S_2P + S_1P = [(50)^2 + (0.51)^2]^{1/2} + [(50)^2 + (0.49)^2]^{1/2} \approx 100.005 \text{ cm}$$

Thus if we replace $S_2P + S_1P$ by $2D$, the error involved is about 0.005%. In this approximation, Eq. (18) becomes

$$S_2P - S_1P \approx \frac{y_nd}{D} \tag{19}$$

Using Eq. (17), we obtain

$$y_n = \frac{n\lambda D}{d} \tag{20}$$

Thus the dark and bright fringes are equally spaced, and the distance between two consecutive dark (or bright) fringes is given by

$$\beta = y_{n+1} - y_n = \frac{(n+1)\lambda D}{d} - \frac{n\lambda D}{d}$$

or

$$\beta = \frac{\lambda D}{d} \tag{21}$$

which is the expression for the fringe width.

To determine the shape of the interference pattern, we first note that the locus of point P such that

$$S_2P - S_1P = \Delta \tag{22}$$

is a hyperbola in any plane containing points S_1 and S_2 (see Example 14.2). Consequently, the locus is a hyperbola of

revolution obtained by rotating the hyperbola about the axis S_1S_2. To find the shape of the fringe on the screen, we assume the origin to be at point O and the z axis to be perpendicular to the plane of the screen as shown in Fig. 14.6. The y axis is assumed to be parallel to S_2S_1. We consider an arbitrary point P on the plane of the screen (i.e., $z = 0$) (see Fig. 14.6). Let its coordinates be $(x, y, 0)$. The coordinates of points S_1 and S_2 are $(0, d/2, D)$ and $(0, -d/2, D)$ respectively. Thus

$$S_2P - S_1P = \left[x^2 + \left(y + \frac{d}{2}\right)^2 + D^2\right]^{1/2} - \left[x^2 + \left(y - \frac{d}{2}\right)^2 + D^2\right]^{1/2} = \Delta \quad \text{(say)}$$

or

$$\left[x^2 + \left(y + \frac{d}{2}\right)^2 + D^2\right] = \left\{\Delta + \left[x^2 + \left(y - \frac{d}{2}\right)^2 + D^2\right]^{1/2}\right\}^2$$

or $$(2yd - \Delta^2)^2 = (2\Delta)^2\left[x^2 + \left(y - \frac{d}{2}\right)^2 + D^2\right]$$

Hence,

$$(d^2 - \Delta^2)y^2 - \Delta^2x^2 = \Delta^2\left[D^2 + \frac{1}{4}(d^2 - \Delta^2)\right]$$

which is the equation of a hyperbola. Thus the shape of the fringes is hyperbolic. On rearranging, we get

$$y = \pm\left(\frac{\Delta^2}{d^2 - \Delta^2}\right)^{1/2}\left[x^2 + D^2 + \frac{1}{4}(d^2 - \Delta^2)\right]^{1/2} \quad \textbf{(23)}$$

For values of x such that

$$x^2 << D^2 \quad \textbf{(24)}$$

the loci are straight lines parallel to the x axis. Thus we obtain approximately straight-line fringes on the screen. It should be emphasized that the fringes are straight lines although sources S_1 and S_2 are point sources. It is easy to see that if we had slits instead of the point sources, we would have obtained again straight-line fringes with increased intensities.

The fringes so produced are said to be nonlocalized; they can be photographed by just placing a film on the screen; they can also be seen through an eyepiece.

14.6 THE INTENSITY DISTRIBUTION

Let $\mathbf{E}_1$ and $\mathbf{E}_2$ be the electric fields produced at point P by S_1 and S_2, respectively (see Fig. 14.8). The electric fields $\mathbf{E}_1$ and $\mathbf{E}_2$ will, in general, have different directions and different magnitudes. However, if the distances S_1P and S_2P are very large in comparison to the distance S_1S_2, the two fields will almost be in the same direction. Thus, we may write

$$\mathbf{E_1} = \hat{\mathbf{i}}E_{01}\cos\left(\frac{2\pi}{\lambda}S_1P - \omega t\right)$$

and

$$\mathbf{E_2} = \hat{\mathbf{i}}E_{02}\cos\left(\frac{2\pi}{\lambda}S_2P - \omega t\right) \quad \textbf{(25)}$$

where $\hat{\mathbf{i}}$ represents the unit vector along the direction of either of the electric fields. The resultant field is given by

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 = \hat{\mathbf{i}}\left[E_{01}\cos\left(\frac{2\pi}{\lambda}S_1P - \omega t\right) + E_{02}\cos\left(\frac{2\pi}{\lambda}S_2P - \omega t\right)\right] \quad \textbf{(26)}$$

The intensity I is proportional to the square of the electric field and is given by

$$I = KE^2 \quad \textbf{(27)}$$

or

$$I = K\left[E_{01}^2\cos^2\left(\frac{2\pi}{\lambda}S_1P - \omega t\right) + E_{02}^2\cos^2\left(\frac{2\pi}{\lambda}S_2P - \omega t\right) + E_{01}E_{02}\left\{\cos\left[\frac{2\pi}{\lambda}(S_2P - S_1P)\right] + \cos\left[2\omega t - \frac{2\pi}{\lambda}(S_2P + S_1P)\right]\right\}\right] \quad \textbf{(28)}$$

where K is a proportionality constant.[3] For an optical beam the frequency is very large ($\omega \approx 10^{15}$ s^{-1}), and all the terms

[3] Equation (27) will be derived in Sec. 23.5. In free space the constant K will be shown to be equal to $\varepsilon_0 c^2$, where ε_0 (=8.854 × 10^{-12} C^2 N^{-1} m^{-2}) represents the permittivity of free space and c is the speed of light in free space.

depending on ωt will vary with extreme rapidity (10^{15} times per second); consequently, any detector would record an average value of various quantities. Now

$$\left\langle \cos^2(\omega t - \theta) \right\rangle = \frac{1}{2\tau}\int_{-\tau}^{+\tau} \frac{1+\cos[2(\omega t-\theta)]}{2} dt$$

$$= \frac{1}{2} + \frac{1}{16\pi}\frac{T}{\tau}\left\{\left[\sin 2(\omega t - \theta)\right]_{-\tau}^{+\tau}\right\}$$

where $T = 2\pi/\omega$ ($\approx 2\pi \times 10^{-15}$ s for an optical beam). For any practical detector T/τ <<< 1, and since the quantity within the curly braces will always be between –2 and +2, we may write

$$\langle \cos^2(\omega t - \theta) \rangle \approx \tfrac{1}{2} \qquad \textbf{(29)}$$

For the normal eye, $\tau \approx 0.1$ s; thus $T/\tau \approx 6 \times 10^{-14}$; even for a detector having 1 ns as the resolution time, $T/\tau \approx 6 \times 10^{-5}$.

The factor $\cos(2\omega t - \phi)$ will oscillate between +1 and –1, and its average will be zero as can indeed be shown mathematically. Thus the intensity, that a detector will record, will be given by

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\delta \qquad \textbf{(30)}$$

where

$$\delta = \frac{2\pi}{\lambda}(S_2P - S_1P) \qquad \textbf{(31)}$$

represents the phase difference between the displacements reaching point P from S_1 and S_2. Further

$$I_1 = \tfrac{1}{2} K E_{01}^2$$

represents the intensity produced by source S_1 if no light from S_2 is allowed to fall on the screen; similarly, $I_2 = \frac{1}{2} K E_{02}^2$ represents the intensity produced by source S_2 if no light from S_1 is allowed to fall on the screen. From Eq. (30) we may deduce the following:

1. The maximum and minimum values of $\cos\delta$ are +1 and –1, respectively; as such, the maximum and minimum values of I are given by

$$\text{and} \qquad I_{\max} = (\sqrt{I_1} + \sqrt{I_2})^2, \quad I_{\min} = (\sqrt{I_1} - \sqrt{I_2})^2 \qquad \textbf{(32)}$$

The maximum intensity occurs when

$$\delta = 2n\pi \qquad n = 0, 1, 2, \ldots$$

or

$$S_2P \sim S_1P = n\lambda$$

and the minimum intensity occurs when

$$\delta = (2n+1)\pi \qquad n = 0, 1, 2, \ldots$$

or

$$S_2P \sim S_1P = \left(n + \frac{1}{2}\right)\lambda$$

When $I_1 = I_2$, the intensity minimum is zero. In general, $I_1 \neq I_2$ and the minimum intensity is not zero.

2. If holes S_1 and S_2 are illuminated by different light sources (see Fig. 14.4), then the phase difference δ will remain constant for about 10^{-10} s (see discussion in Sec. 14.3) and thus δ would also vary with time[4] in a random way. If we now carry out the averaging over time scales which are of the order of 10^{-8} s, then

$$\langle \cos\delta \rangle = 0$$

and we obtain

$$I = I_1 + I_2$$

Thus, for two incoherent sources, the resultant intensity is the sum of the intensities produced by each one of the sources independently, and no interference pattern is observed.

3. In the arrangement shown in Fig. 14.6, if the distances S_1P and S_2P are extremely large in comparison to d, then

$$I_1 \approx I_2 = I_0 \qquad \text{(say)}$$

and

$$I = 2I_0 + 2I_0\cos\delta = 4I_0\cos^2\frac{\delta}{2} \qquad \textbf{(33)}$$

The intensity distribution (which is often termed the $\cos^2$ pattern) is shown in Fig. 14.9. The actual fringe pattern (as it will appear on the screen) is shown in Fig. 14.10. Figure 14.10(a) and (b) corresponds to $d = 0.005$ mm ($\beta \approx 5$ mm) and $d = 0.025$ mm ($\beta \approx 1$ mm), respectively. Both figures correspond to $D = 5$ cm and $\lambda = 5 \times 10^{-5}$ cm. The values of the parameters are such that one can see the hyperbolic nature of the fringe pattern in Fig. 14.10(a).

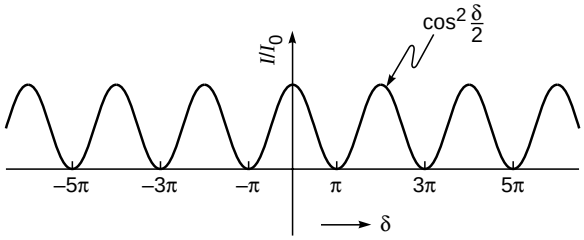


Fig. 14.9 The variation of intensity with δ.

[4] Notice that this variation occurs in times of the order of 10^{-10} s which is about 1 million times longer than the times for variation of the intensity due to the terms depending on ωt. Thus we are justified in first carrying out the averaging which leads to Eq. (30).

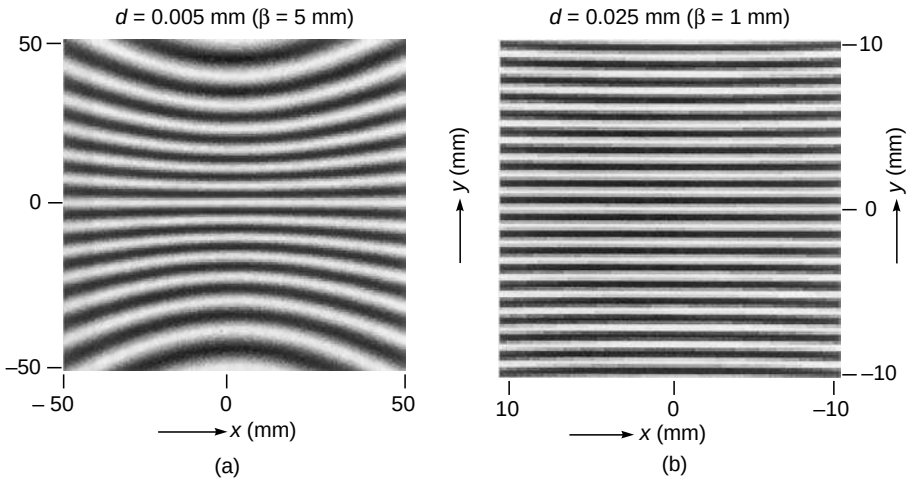


Fig. 14.10 Computer-generated fringe pattern produced by two point sources S_1 and S_2 on the screen LL' (see Fig. 14.8); (a) and (b) correspond to d = 0.005 and 0.025 mm, respectively (both figures correspond to D = 5 cm and $\lambda = 5 \times 10^{-5}$ cm).

Example 14.5 Instead of considering two point sources, we consider the superposition of two plane waves as shown in Fig. 14.11(a). The wave vectors for the two waves are given by

$$\mathbf{k}_1 = -\hat{\mathbf{y}}k \sin\theta_1 + \hat{\mathbf{z}}k \cos\theta_1$$

and

$$\mathbf{k}_2 = +\hat{\mathbf{y}}k \sin\theta_2 + \hat{\mathbf{z}}k \cos\theta_2$$

where $k = 2\pi/\lambda$ and θ_1 and θ_2 are defined in Fig. 14.11(a). Thus the electric fields of the two waves are described by the equations

$$E_1 = E_{01} \cos(\mathbf{k_1} \cdot \mathbf{r} - \omega t)$$
$$= E_{01} \cos(-ky \sin\theta_1 + kz \cos\theta_1 - \omega t)$$
$$E_2 = E_{02} \cos(\mathbf{k_2} \cdot \mathbf{r} - \omega t)$$
$$= E_{02} \cos(ky \sin\theta_2 + kz \cos\theta_2 - \omega t)$$

where we have assumed both electric fields along the same direction (say, along the x axis); if we further assume $E_{01} = E_{02} = E_0$ and $\theta_1 = \theta_2 = \theta$, then the resultant field is given by

$$E = 2E_0 \cos(ky \sin\theta) \cos(kz \cos\theta - \omega t)$$

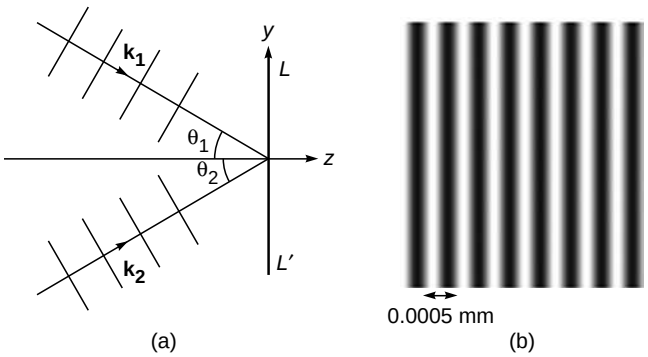


Fig. 14.11 (a) The superposition of two plane waves on LL'. (b) Computer-generated interference pattern on the screen LL' for $\theta_1 = \theta_2 = \pi/6$ and λ = 5000 Å. The fringes are parallel to the x axis.

Thus the intensity distribution on the photograph plate LL' is given by

$$I = 4I_0 \cos^2(ky \sin\theta)$$

and the fringe pattern will be strictly straight lines (parallel to the x-axis) with fringe width given by

$$\beta = \frac{\lambda}{2 \sin\theta}$$

Figure 14.11(b) shows the computer-generated interference pattern on the screen LL' for $\theta = \pi/6$ and λ = 5000 Å. Thus $\beta = \lambda$ = 0.0005 mm.

Example 14.6 In this example, we consider the interference pattern produced by two point sources S_1 and S_2 on a plane PP' which is perpendicular to the line joining S_1 and S_2 [see Fig. 14.12(a)]. Obviously, on plane PP', the locus of point P for which

$$S_1P - S_2P = \text{constant}$$

will be a circle. Figure 14.12(b) and (c) shows the fringe patterns for D = 20 and 10 cm; for both figures $S_1S_2 = d$ = 0.05 mm and λ = 5000 Å. Obviously, if O represents the center of the fringe pattern, then

$$S_1O - S_2O = d = 100\lambda$$

Thus (for this value of d) the central spot will be bright for all values of D and will correspond to n = 100. The first and second bright circles will correspond to a path difference of 99λ and 98λ, respectively. Similarly, the first and second dark rings in the interference pattern will correspond to a path difference of 99.5λ and 98.5λ, respectively. The radii of the fringes can be calculated by using the formula given in Prob. 14.10.

Example 14.7 We finally consider the interference pattern produced on PP' by the superposition of a plane wave incident

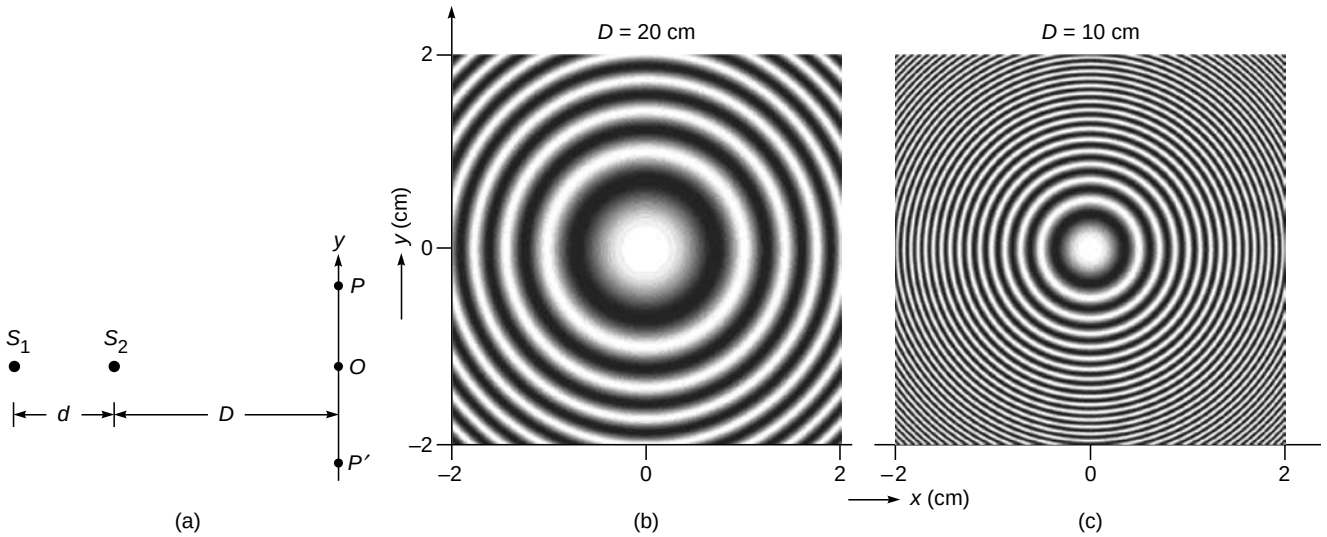


Fig. 14.12 (a) S_1 and S_2 represent two coherent sources. (b) and (c) Interference fringes observed on the screen PP' when D = 20 cm and D = 10 cm, respectively.

normally and a spherical wave emanating from point O (see Fig.14.13). The plane wave is given by

$$E_1 = E_0 \cos(kz - \omega t + \phi)$$

and the spherical wave is given by

$$E_2 = \frac{A_0}{r} \cos(kr - \omega t)$$

where r is the distance measured from point O which is assumed to be the origin. Now, on the plane PP' ($z = D$)

$$r = (x^2 + y^2 + D^2)^{1/2} \approx D\left(1 + \frac{x^2 + y^2}{2D^2}\right)$$

$$\approx D + \frac{x^2 + y^2}{2D}$$

where we have assumed $x, y << D$. On the plane $z = D$, the resultant field is given by

$$E = E_1 + E_2$$

$$\approx E_0 \cos(kD - \omega t + \phi) + \frac{A_0}{D} \cos\left[kD + \frac{k}{2D}(x^2 + y^2) - \omega t\right]$$

Thus

$$\langle E^2 \rangle = \frac{1}{2}E_0^2 + \frac{1}{2}\left(\frac{A_0}{D}\right)^2 + E_0 \frac{A_0}{D} \cos\left[\frac{k}{2D}(x^2 + y^2) - \phi\right]$$

If we assume that

$$\frac{A_0}{D} \approx E_0$$

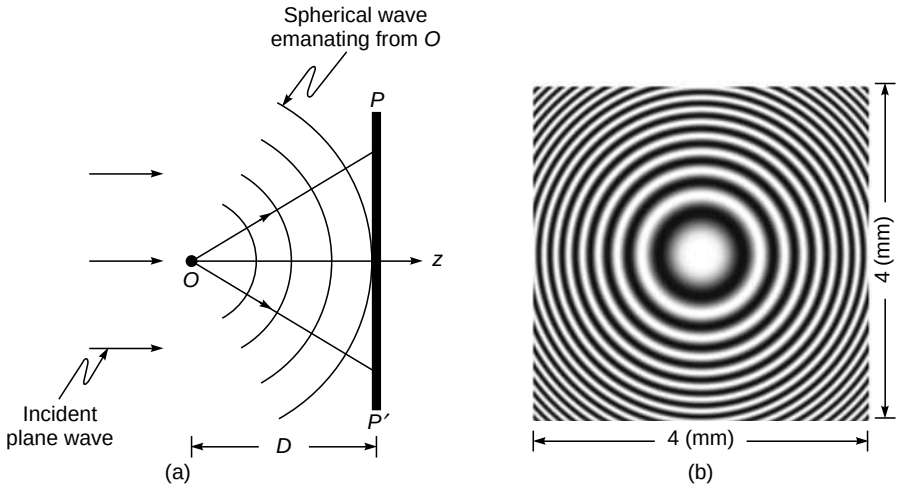


Fig. 14.13 (a) Superposition of a plane wave and a spherical wave emanating from point O; (b) interference fringes observed on the screen PP'.

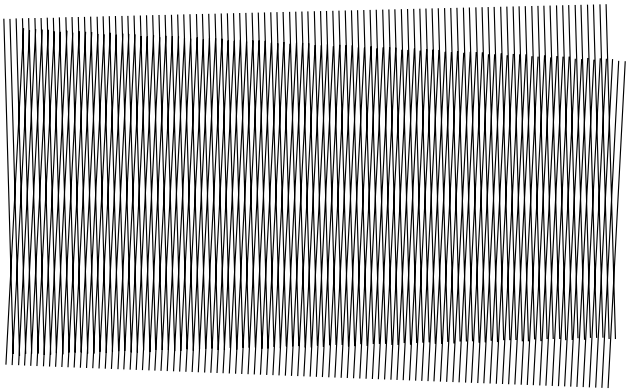


Fig. 14.14 The moiré pattern produced by two overlapping straight-line patterns.

i.e., the amplitude of the spherical wave (on plane PP') is the same as the amplitude of the plane wave, then

$$\langle E^2 \rangle \approx 2E_0^2 \cos^2\left[\frac{k}{4D}(x^2+y^2) - \frac{1}{2}\phi\right]$$

and we obtain circular interference fringes as shown in Fig. 14.13(b). If r_m and r_{m+p} denote the radii of the mth and $(m+p)$th bright rings, then

$$r_{m+p}^2 - r_m^2 = 2p\lambda D$$

14.6.1 Moiré Fringes

Moiré fringes can be very effectively used to study the formation of fringe patterns. In Fig. 14.14 we have shown the overlapping of two simple patterns from which one can understand the formation of bright and dark fringes when two plane waves propagate in slightly different directions. In a classroom, it can be easily demonstrated by having a periodic pattern on a transparency and overlapping it with its own photocopy at different angles. Similarly, if one overlaps a circular pattern (on a transparency) with its own copy, one obtains the hyperbolic fringes as shown in Fig. 14.15. (To get a clearer fringe pattern, you may have to view the patterns from a greater distance.) In Sec. 17.5 we have shown how the beat phenomenon can be understood by observing the Moiré fringes obtained by the overlapping of two patterns of slightly different periods (see Fig. 17.13).

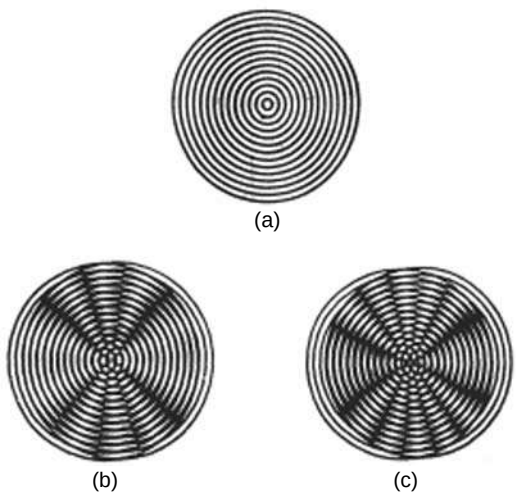


Fig. 14.15 The moiré pattern produced by two overlapping circular patterns. You will see clear hyperbolic fringes if you put the pattern at a greater distance from the eye. The circular pattern was provided by Dr. R. E. Bailey.

Example 14.8 Consider a parallel beam of light (from a distant source S' such as a star) incident (at an angle θ) on two slits S_1 and S_2 as shown in Fig. 14.16. Obviously the path difference between the waves emanating from slits S_1 and S_2 is given by

$$XS_2 = d \sin\theta$$

Therefore the intensity distribution on the screen due to S' is given by

$$I = I_0 \cos^2\frac{\delta}{2}$$

where

$$\delta = \frac{2\pi}{\lambda}\left(XS_2 + S_2P - S_1P\right)$$
$$= \frac{2\pi}{\lambda}\left[(S_2P - S_1P) + d\sin\theta\right]$$
$$= \frac{2\pi}{\lambda}\left(\frac{xd}{D} + d\sin\theta\right)$$

Thus the intensity distribution (due to light coming from the distant source S') is given by

$$I' = I_0 \cos^2\left[\frac{\pi}{\lambda}\left(\frac{xd}{D} + d\sin\theta\right)\right]$$

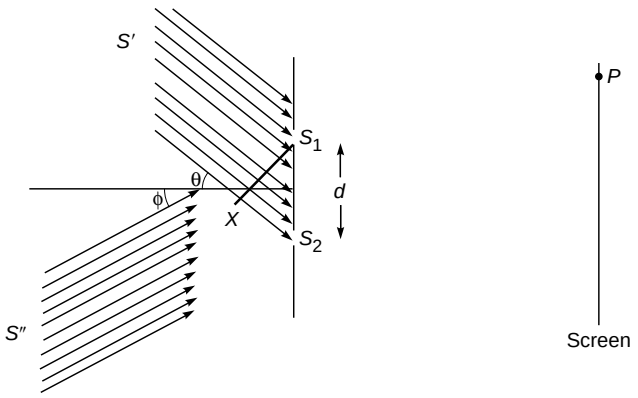


Fig. 14.16 Two distant sources illuminating the slits S_1 and S_2.

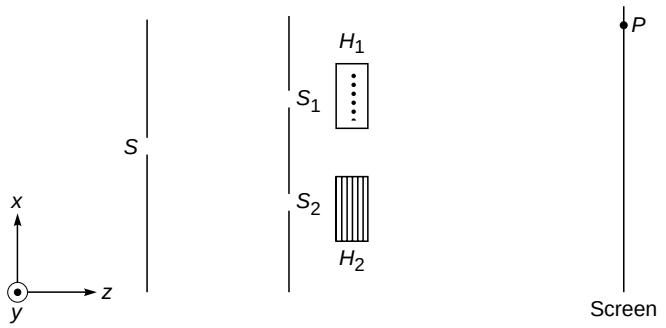


Fig. 14.17 H_1 and H_2 are half wave plates placed in front of S_1 and S_2. The optic axis of H_1 and H_2 are along y and x directions respectively.

Similarly, if there is light incident from another distant source S'' (at an angle ϕ), then the corresponding intensity distribution on the screen is given by

$$I'' = I_0 \cos^2\left[\frac{\pi}{\lambda}\left(\frac{\text{xd}}{\text{D}} - d\sin\phi\right)\right]$$

The resultant intensity distribution is given by

$$I = I' + I''$$

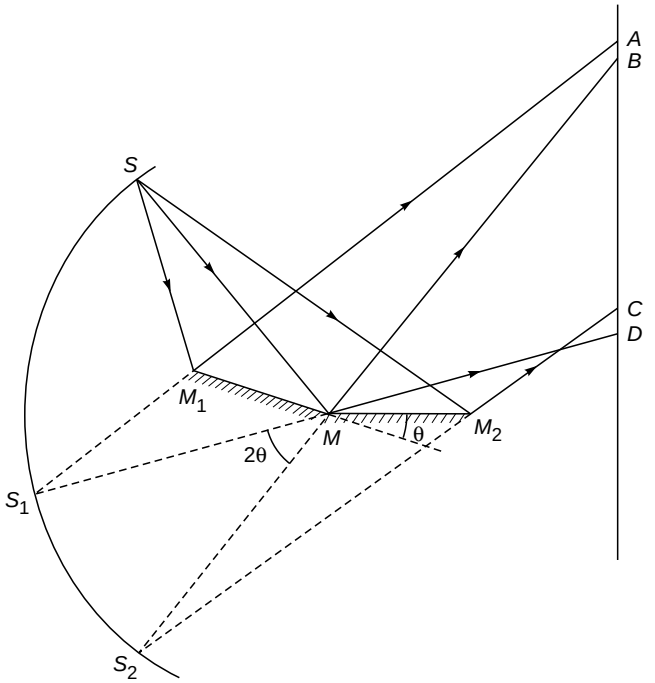


Fig. 14.18 Fresnel's two-mirror arrangement.

Example 14.9 This example presupposes the knowledge of half wave plates (see Sec. 22.6), and therefore readers may skip this example until they have gone through Chap. 22.

Consider a *y*-polarized light beam incident on a double-hole system as shown in Fig. 14.17. Behind the hole S_1 we have put a half wave plate H_1 whose optic axis is along the *y* direction, and behind the hole S_2 we have put a half wave plate H_2 whose optic axis is along the *x* direction. Thus as discussed in Sec. 22.6, in H_1 a *y*-polarized beam will propagate with velocity c/n_e; and in H_2 a *y*-polarized beam will propagate with velocity c/n_o. In calcite. $n_e < n_o$; and in a half wave plate, a phase change of π is introduced between the *o* wave and the *e* wave. Thus the whole fringe pattern will shift by $\beta/2$, where β is the fringe width. What will happen if the incident light beam is *x*-polarized?

14.7 FRESNEL'S TWO-MIRROR ARRANGEMENT

After Young's double-hole interference experiment, Fresnel devised a series of arrangements to produce the interference pattern. One of the experimental arrangements, known as the Fresnel two-mirror arrangement, is shown in Fig. 14.18; it consists of two plane mirrors which are inclined to each other at a small angle θ and touching at the point *M*. Point *S* represents a narrow slit placed perpendicular to the plane of the paper.

A portion of the wave front from *S* gets reflected from M_1M and illuminates the region *AD* of the screen. Another portion of the wave front gets reflected from the mirror MM_2 and illuminates the region *BC* of the screen. Since these two wave fronts are derived from the same source, they are coherent. Thus in the region *BC*, one observes interference fringes. The formation of the fringes can also be understood as being due to the interference of the wave fronts from the virtual sources S_1 and S_2 of *S* formed by mirrors M_1 and M_2, respectively. From simple geometric considerations, it can be shown that points *S*, S_1, and S_2 lie on a circle whose center is at point *M*. Further, if the angle between the mirrors is θ, then the angle S_1SS_2 is also θ and the angle S_1MS_2 is 2θ. Thus S_1S_2 is $2R\theta$, where *R* is the radius of the circle.

14.8 FRESNEL BIPRISM

Fresnel devised yet another simple arrangement for the production of interference pattern. He used a biprism, which was actually a simple prism, the base angles of which are extremely small (~20′). The base of the prism is shown in Fig. 14.19, and the prism is assumed to stand perpendicular to the plane of the paper. Point *S* represents the slit which is also placed perpendicular to the plane of the paper. Light from slit *S* gets refracted by the prism and produces two virtual images S_1 and S_2. These images act as coherent sources and produce interference fringes on the right of the biprism. The fringes can be viewed through an eyepiece. If *n* represents the refractive

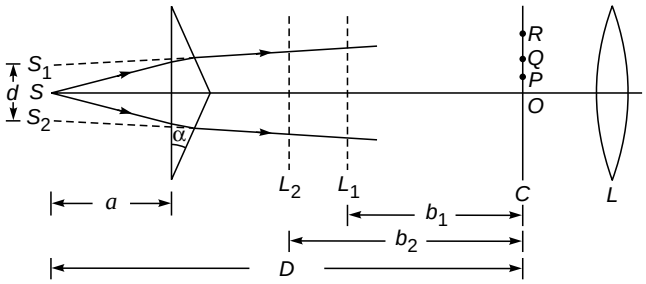


Fig. 14.19 Fresnel's biprism arrangement. Points *C* and *L* represent the positions of the crosswires and the eyepiece, respectively. To determine *d,* one introduces a lens between the biprism and the crosswires; L_1 and L_2 represent the two positions of the lens where the slits are clearly seen.

index of the material of the biprism and α the base angle, then $(n-1)\alpha$ is approximately the angular deviation produced by the prism, and therefore the distance S_1S_2 is $2a(n-1)\alpha$, where a represents the distance from S to the base of the prism. Thus, for $n = 1.5$, $\alpha \approx 20' \approx 5.8 \times 10^{-3}$ radians, $a \approx 2$ cm, one gets $d = 0.012$ cm.

The biprism arrangement can be used for the determination of wavelength of an almost monochromatic light such as the one coming from a sodium lamp. Light from the sodium lamp illuminates slit *S*, and interference fringes can be easily viewed through the eyepiece. The fringe width β can be determined by means of a micrometer attached to the eyepiece. Once β is known, λ can be determined by using the following relation:

$$\lambda = \frac{d\beta}{D} \qquad \textbf{(34)}$$

To determine *d*, one need not measure the value of α. In fact the distances *d* and *D* can be easily determined by placing a convex lens between the biprism and the eyepiece. For a fixed position of the eyepiece there will be two positions of the lens (shown as L_1 and L_2 in Fig. 14.19) where the images of S_1 and S_2 can be seen at the eyepiece. Let d_1 be the distance between the two images when the lens is at position L_1 (at a distance b_1 from the eyepiece). Let d_2 and b_2 be the corresponding distances when the lens is at L_2. Then it can be easily shown that

$$d = \sqrt{d_1 d_2}$$

and

$$D = b_1 + b_2$$

Typically for $d \approx 0.01$ cm, $\lambda \approx 6 \times 10^{-5}$ cm, $D \approx 50$ cm, and $\beta \approx 0.3$ cm.

In the above we considered here a slit instead of a point source. Since each pair of points S_1 and S_2 produces (approximately) straight-line fringes, the slit will also produce straight-line fringes of increased intensity.

14.9 INTERFERENCE WITH WHITE LIGHT

We will now discuss the interference pattern when the slit is illuminated by white light. The wavelengths corresponding to the violet and red ends of the spectrum are about 4×10^{-5} cm and 7×10^{-5} cm, respectively. Clearly, the central fringe produced at point *O* (Fig. 14.19) will be white because all wavelengths will constructively interfere here. Now, slightly below (or above) point *O* the fringes will become colored. For example, if point *P* is such that

$$S_2P \sim S_1P = 2 \times 10^{-5} \text{ cm} \left(= \frac{\lambda_{\text{violet}}}{2} \right)$$

then complete destructive interference will occur only for the violet color. Partial destructive interference will occur for other wavelengths. Consequently we will have a line devoid of the violet color that will appear reddish. The point *Q* which satisfies

$$S_2Q \sim S_1Q = 3.5 \times 10^{-5} \text{ cm} \left(= \frac{\lambda_{\text{red}}}{2} \right)$$

will be devoid of the red color. It will correspond to almost constructive interference for the violet color. No other wavelength (in the visible region) will either constructively or destructively interfere. Thus following the white central fringe we will have colored fringes; when the path difference is about 2×10^{-5} cm, the fringe will be red, then the color will gradually change to violet. The colored fringes will soon disappear because at points far away from *O* there will be so many wavelengths (in the visible region) which will constructively interfere that we will observe uniform white illumination. For example, at a point *R*, such that $S_2R \sim S_1R = 30 \times 10^{-5}$ cm, wavelengths corresponding to $30 \times 10^{-5}/n$ ($n = 1, 2, \ldots$) will constructively interfere. In the visible region these wavelengths will be 7.5×10^{-5} cm (red), 6×10^{-5} cm (yellow), 5×10^{-5} cm (greenish yellow), and 4.3×10^{-5} cm (violet). Further, wavelengths corresponding to $30 \times 10^{-5}/\left(n + \frac{1}{2}\right)$ will destructively interfere; thus, in the visible region, the wavelengths 6.67×10^{-5} cm (orange), 5.5×10^{-5} cm (yellow), 4.6×10^{-5} cm (indigo) and 4.0×10^{-5} cm (violet)

will be absent. The color of such light, as seen by the unaided eye, will be white. Thus, with white light one gets a white central fringe at the point of zero path difference along with a few colored fringes on both the sides, the color soon fading off to white. While using a white light source, if we put a red (or green) filter in front of our eye, we will see the interference pattern corresponding to the red (or green) light.

As discussed above, when we observe an interference pattern using a white light source, we will see only a few colored fringes. However, if we put a red filter in front of our eye, the fringe pattern (corresponding to the red color) will suddenly appear. If we replace the red filter by a green filter in front of our eye, the fringe pattern corresponding to the green color will appear.

In the usual interference pattern with a nearly monochromatic source (such as a sodium lamp), a large number of interference fringes are obtained, and it is extremely difficult to determine the position of the central fringe. In many interference experiments it is necessary to determine the position of the central fringe, and as has been discussed above, this can be easily done by using white light as a source.

14.10 DISPLACEMENT OF FRINGES

We will now discuss the change in the interference pattern produced by introducing a thin transparent plate in the path of one of the two interference beams as shown in Fig. 14.20. Let t be the thickness of the plate, and let n be its refractive index. It is easily seen from the figure that light reaching point P from S_1 has to traverse a distance t in the plate and a distance $S_1P - t$ in air. Thus the time required for the light to reach from S_1 to point P is given by

$$\frac{S_1P - t}{c} + \frac{t}{v} = \frac{1}{c}\left(S_1P - t + nt\right)$$

$$= \frac{1}{c}\left[S_1P + (n-1)t\right] \qquad \textbf{(35)}$$

where $v\ \left(= c/n\right)$ represents the speed of light in the plate. Equation (35) shows that by introducing the thin plate the effective optical path increases by $(n-1)t$. Thus, when the thin plate is introduced, the central fringe (which corresponds to equal optical path from S_1 and S_2) is formed at point O' where

$$S_1O' + (n-1)t = S_2O'$$

Since [see Eq. (19)]

$$S_2O' - S_1O' \approx \frac{d}{D}OO'$$

therefore

$$(n-1)t = \frac{d}{D}OO' \qquad \textbf{(36)}$$

Thus the fringe pattern gets shifted by a distance Δ which is given by

$$\Delta = \frac{D(n-1)\,t}{d} \qquad \textbf{(37)}$$

The above principle enables us to determine the thickness of extremely thin transparent sheets (such as that of mica) by measuring the displacement of the central fringe. Further, if white light is used as a source, the displacement of the central fringe is easy to measure.

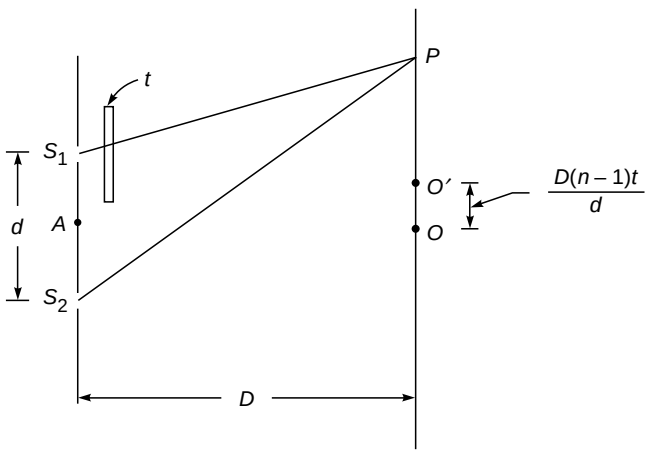


Fig. 14.20 If a thin transparent sheet (of thickness t) is introduced in one of the beams, the fringe pattern gets shifted by a distance $(n-1)tD/d$.

Example 14.10 In a double-slit interference arrangement one of the slits is covered by a thin mica sheet whose refractive index is 1.58. The distances S_1S_2 and AO (see Fig. 14.20) are 0.1 and 50 cm, respectively. Due to the introduction of the mica sheet the central fringe gets shifted by 0.2 cm. Determine the thickness of the mica sheet.

Solution: $\Delta = 0.2$ cm $\quad d = 0.1$ cm $\quad D = 50$ cm

Hence

$$t = \frac{d\,\Delta}{D(n-1)} = \frac{0.1 \times 0.2}{50 \times 0.58}$$

$$\approx 6.9 \times 10^{-4} \text{ cm}$$

Example 14.11 In an experimental arrangement similar to that discussed in Example 14.10, one finds that by introducing the mica sheet the central fringe occupies the position that was originally occupied by the eleventh bright fringe. If the source of light is a sodium lamp (λ = 5893 Å), determine the thickness of the mica sheet.

Solution: The point O' (see Fig. 14.20) corresponds to the eleventh bright fringe, thus

$$S_2O' - S_1O' = 11\lambda = (n-1)t = 0.58t$$

14.11 LLOYD'S MIRROR ARRANGEMENT

In this arrangement, light from a slit S_1 is allowed to fall on a plane mirror at grazing incidence (see Fig. 14.21). The light directly coming from slit S_1 interferes with the light reflected from the mirror, forming an interference pattern in the region BC of the screen. One may thus consider slit S_1 and its virtual image S_2 to form two coherent sources which produce the interference pattern. Note that at grazing incidence one really need not have a mirror; even a dielectric surface has very high reflectivity (see Chap. 23).

As can be seen from Fig. 14.21, the central fringe cannot be observed on the screen unless the latter is moved to the position $L'_1L'_2$, where it touches the end of the reflector. Alternatively, one may introduce a thin mica sheet in the path of the direct beam so that the central fringe appears in the region BC. (This is discussed in detail in Prob. 14.2.) Indeed, if the central fringe is observed with white light, it is found to be dark. This implies that the reflected beam undergoes a sudden phase change of π on reflection. Consequently, when point P on the screen is such that

$$S_2P - S_1P = n\lambda \qquad n = 0, 1, 2, 3, \ldots$$

we will get minima (i.e., destructive interference). On the other hand, if

$$S_2P - S_1P = \left(n + \frac{1}{2}\right)\lambda$$

we will get maxima.

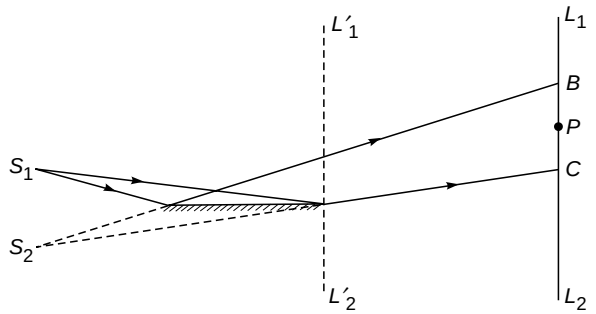


Fig. 14.21 The Lloyd's mirror arrangement.

In the next section, using the principle of optical reversibility, we will show that if there is an abrupt phase change of π when light gets reflected by a denser medium, then no such abrupt phase change occurs when reflection takes place at a rarer medium.

14.12 PHASE CHANGE ON REFLECTION

We will now investigate the reflection of light at an interface between two media, using the principle of optical reversibility. According to this principle, in the absence of any absorption, a light ray that is reflected or refracted will retrace its original path if its direction is reversed.[5]

Consider a light ray incident on an interface of two media of refractive indices n_1 and n_2 as shown in Fig. 14.22(a). Let the amplitude reflection and transmission coefficients be r_1 and t_1, respectively. Thus, if the amplitude of the incident ray is a, then the amplitudes of the reflected and refracted rays are ar_1 and at_1, respectively.

We now reverse the rays, and we consider a ray of amplitude at_1 incident on medium 1 and a ray of amplitude ar_1 incident on medium 2 as shown in Fig. 14.22(b). The ray of amplitude at_1 will give rise to a reflected ray of amplitude at_1r_2 and a transmitted ray of amplitude at_1t_2, where r_2 and t_2 are the amplitude reflection and transmission coefficients, respectively, when a ray is incident from medium 2 on medium 1. Similarly, the ray of amplitude ar_1 will give rise to a ray of amplitude ar_1^2 and a refracted ray of amplitude ar_1t_1. According to the principle of optical reversibility, the two rays of amplitudes ar_1^2 and at_1t_2 must combine to give the incident ray of Fig. 14.22(a); thus

$$ar_1^2 + at_1t_2 = a$$

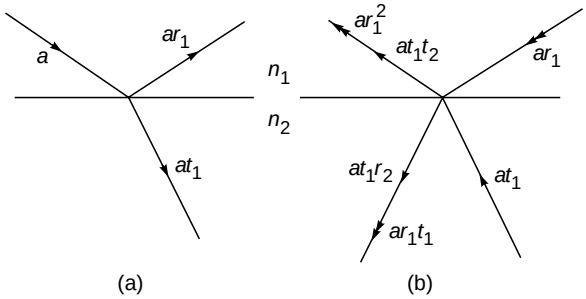


Fig. 14.22 (a) A ray traveling in a medium of refractive index n_1 incident on a medium of refractive index n_2. (b) Rays of amplitude ar_1 and at_1 incident on a medium of refractive index n_1.

[5] This principle is a consequence of time reversal invariance according to which processes can run either way in time; for more details see Refs. 3 and 8.

or

$$t_1 t_2 = 1 - r_1^2 \quad \textbf{(38)}$$

Further, the two rays of amplitudes at_1r_2 and ar_1t_1 must cancel each other, i.e.,

$$at_1r_2 + ar_1t_1 = 0$$

or

$$r_2 = -r_1 \quad \textbf{(39)}$$

Since we know from Lloyd's mirror experiment that an abrupt phase change of π occurs when light gets reflected by a denser medium, we may infer from Eq. (39) that no such abrupt phase change occurs when light gets reflected by a rarer medium. This is indeed borne out by experiments. Equations (38) and (39) are known as Stokes' relations.

In Chap. 24, we will calculate the amplitude reflection and transmission coefficients for plane waves incident on a dielectric and also on a conductor. It will be shown that the coefficients satisfy Stokes' relations; the phase change on reflection will also be discussed there.

Summary

- In 1801, Thomas Young devised an ingenious but simple method to lock the phase relationship between two sources of light. The trick lies in the division of a single wave front into two; these two split wave fronts act as if they emanated from two sources having a fixed phase relationship, and therefore when these two waves are allowed to interfere, a stationary interference pattern is obtained.
- For two coherent point sources, almost straight-line interference fringes are formed on some planes, and by measuring the fringe width (which represents the distance between two consecutive fringes) one can calculate the wavelength.
- On a plane which is normal to the line joining the two coherent point sources, the fringe pattern is circular.
- In Young's double-slit interference pattern, if we use a white light source, we get a white central fringe at the point of zero path difference along with a few colored fringes on both the sides, the color soon fading off to white. If we now introduce a very thin slice of transparent material (such as mica) in the path of one of the interfering beams, the fringes get displaced; and by measuring the displacement of fringes, we can calculate the thickness of the mica sheet.

Problems

14.1 In Young's double-hole experiment (see Fig. 14.6), the distance between the two holes is 0.5 mm, $\lambda = 5 \times 10^{-5}$ cm, and D = 50 cm. What will be the fringe width?

14.2 Figure 14.23 represents the layout of Lloyd's mirror experiment. Point S is a point source emitting waves of frequency 6×10^{14} s^{-1}. Points A and B represent the two ends of a mirror placed horizontally, and LOM represents the screen. The distances SP, PA, AB, and BO are 1 mm, 5 cm, 5 cm, and 190 cm, respectively. (a) Determine the position of the region where the fringes will be visible, and calculate the number of fringes. (b) Calculate the thickness of a mica sheet (n = 1.5) which should be introduced in the path of the direct ray so that the lowest fringe becomes the central fringe. The velocity of light is 3×10^{10} cm s^{-1}.

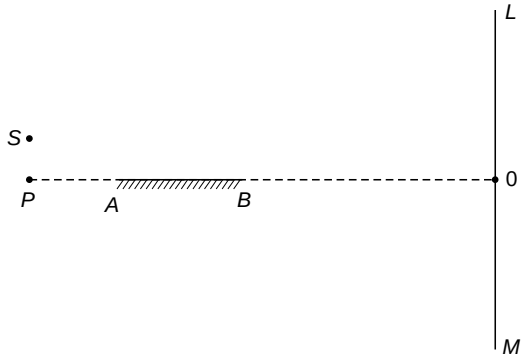


Fig. 14.23 For Prob. 14.2.

[**Ans:** (a) 2 cm, 40 fringes, (b) 38 μm]

14.3 **(a)** In Fresnel's biprism arrangement, show that $d = 2(n-1)a\alpha$, where a represents the distance from the source to the base of the prism (see Fig. 14.19), α is the angle of the biprism, and n is the refractive index of the material of the biprism.

(b) In a typical biprism arrangement b/a = 20, and for sodium light ($\lambda \simeq$ 5893 Å) one obtains a fringe width of 0.1 cm; here b is the distance between the biprism and the screen. Assuming n = 1.5, calculate the angle α.

[**Ans:** $\simeq$ 0.71°]

14.4 In Young's double-hole experiment, a thin mica sheet (n = 1.5) is introduced in the path of one of the beams. If the central fringe gets shifted by 0.2 cm, calculate the thickness of the mica sheet. Assume d = 0.1 cm and D = 50 cm.

14.5 To determine the distance between the slits in the Fresnel biprism experiment, one puts a convex lens in between the biprism and the eyepiece. Show that if $D > 4f$, one will obtain two positions of the lens where the image of the slits will be formed at the eyepiece; here f is the focal length of the convex lens, and D is the distance between the slit and the eyepiece. If d_1 and d_2 are the distances between the images (of the slits) as measured by the eyepiece, then show that $d = \sqrt{d_1 d_2}$. What would happen if $D < 4f$?

14.6 In Young's double-hole experiment, interference fringes are formed using sodium light which predominantly comprises two wavelengths (5890 and 5896 Å). Obtain the regions on the screen where the fringe pattern will disappear. You may assume d = 0.5 mm and D = 100 cm.

14.7 If one carries out Young's double-hole interference experiment using microwaves of wavelength 3 cm, discuss the nature of the fringe pattern if d = 0.1, 1, and 4 cm. You may assume D = 100 cm. Can you use Eq. (21) for the fringe width?

14.8 In Fresnel's two-mirror arrangement (see Fig. 14.18) show that points S, S_1, and S_2 lie on a circle and $S_1S_2 = 2b\theta$, where $b = MS$ and θ is the angle between the mirrors.

14.9 In the double-hole experiment using white light, consider two points on the screen, one corresponding to a path difference of 5000 Å and the other corresponding to a path difference of 40,000 Å. Find the wavelengths (in the visible region) which correspond to constructive and destructive interference. What will be the color of these points?

14.10 **(a)** Consider a plane which is normal to the line joining two point coherent sources S_1 and S_2 as shown in Fig. 14.14. If $S_1P - S_2P = \Delta$, then show that

$$y = \frac{1}{2\Delta}(d^2 - \Delta^2)^{1/2}[4D^2 + 4Dd + (d^2 - \Delta^2)]^{1/2}$$

$$\approx \frac{D}{\Delta}\sqrt{(d-\Delta)(d+\Delta)}$$

where the last expression is valid for $D >> d$.

(b) For λ = 0.5 μm, d = 0.4 mm and D = 20 cm; $S_1O - S_2O = 800\,\lambda$. Calculate the value of $S_1P - S_2P$ for point P to be the first dark ring and first bright ring.

[**Ans:** 0.39975 mm, 0.3995 mm]

14.11 In continuation of Prob. 14.10, calculate the radii of the first two dark rings for **(a)** D = 20 cm and **(b)** D = 10 cm.

[**Ans: (a)** ≈ 0.71 cm, **(b)** 1.22 cm]

14.12 In continuation of Prob. 14.10, assume that d = 0.5 mm, $\lambda = 5 \times 10^{-5}$ cm, and D = 100 cm. Thus the central (bright) spot will correspond to n = 1000. Calculate the radii of the first, second, and third bright rings which will correspond to n = 999, 998, and 997, respectively.

14.13 Using the expressions for the amplitude reflection and transmission coefficients [see Eqs. (67) to (72) of Chap. 24], show that they satisfy Stokes' relations.

14.14 Assume a plane wave incident normally on a plane containing two holes separated by a distance d. If we place a convex lens behind the slits, show that the fringe width, as observed on the focal plane of the lens, will be $f\lambda/d$, where f is the focal length of the lens.

14.15 In Prob. 14.14, show that if the plane (containing the holes) lies in the front focal plane of the lens, then the interference pattern will consist of exactly parallel straight lines. However, if the plane does not lie on the front focal plane, the fringe pattern will be hyperbolas.

14.16 In Young's double-hole experiment, calculate I/I_{max} where I represents the intensity at a point where the path difference is $\lambda/5$.

REFERENCES AND SUGGESTED READINGS

1. F. Graham Smith and T. A. King, *Optics and Photonics: An Introduction*, John Wiley, Chichester, 2000.
2. R. W. Ditchburn, *Light*, Academic Press, London, 1976.
3. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Chap. 52, Addison-Wesley, 1965.
4. M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, Cambridge, 2000.
5. E. Hecht and A. Zajac, *Optics*, Addison-Wesley, Reading, Mass., 1974.
6. R. S. Longhurst, *Geometrical and Physical Optics*, 2d ed., Longman, London, 1973.
7. D. E. Bailey and M. J. Welch, "Moire Fringes," *Proceedings of the Conference and Workshop on the Teaching of Optics*, (Eds. G. I. Opat, D. Booth, A. P. Mazzolini, and G. Smith, University of Melbourne, Australia, 1989.
8. A. Baker, *Modern Physics and Anti Physics*, Chap. 3, Addison-Wesley, Reading, Mass., 1970.
9. PSSC, *Physics*, D.C. Heath & Co. Boston, Mass., 1965.

Chapter Fifteen
INTERFERENCE BY DIVISION OF AMPLITUDE

Following a method suggested by Fizeau in 1868, Professor Michelson has produced what is perhaps the most ingenious and sensational instrument in the service of astronomy—the interferometer.

—Sir James Jeans, *The Universe Around Us*, Cambridge University Press, 1930

Important Milestones

1665	*In his treatise Micrographia, the British physicist Robert Hooke described his observations of the colors produced in flakes of mica, soap bubbles, and films of oil on water. He recognized that the color produced in mica flakes is related to their thickness but was unable to establish any definite relationship between thickness and color. Hooke supported a wave theory of light.*
1704	*"Newton's rings" were first observed by Boyle and Hooke—they are named after Newton because he had given an explanation using the corpuscular model which was later found to be unsatisfactory.*
1802	*Thomas Young gave a satisfactory explanation of Newton's rings based on wave theory.*
1881	*A. A. Michelson invented the Michelson interferometer. He was awarded the 1907 Nobel Prize in Physics "for his optical precision instruments and the spectroscopic and metrological investigations carried out with their aid." Michelson was America's first Nobel Prize winner in Science, and during the presentation ceremony of the Nobel Prize, the president of the Royal Swedish Academy of Sciences said, "Professor Michelson, your interferometer has rendered it possible to obtain a nonmaterial standard of length possessed of a degree of accuracy never hitherto attained. By its means we are enabled to ensure that the prototype of the meter has remained unaltered in length, and to restore it with absolute infallibility, supposing it were to get lost. . . . "*
1887	*A. A. Michelson and E. W. Morley carried out the famous Michelson–Morley experiment using the Michelson interferometer to detect the motion of the Earth with respect to the "Luminiferous Aether."*

15.1 INTRODUCTION

In Chap. 14 we discussed the interference pattern produced by division of a wave front; for example, light coming out of a pinhole was allowed to fall on two holes, and spherical waves emanating from these two holes produced the interference pattern. In this chapter we will consider the formation of interference pattern by division of amplitude; for example, if a plane wave falls on a thin film, then the wave reflected from the upper surface interferes with the wave reflected from the lower surface. Such studies have many practical applications and also explain phenomena such as the formation of beautiful colors produced by a soap film illuminated by white light.

15.2 INTERFERENCE BY A PLANE PARALLEL FILM WHEN ILLUMINATED BY A PLANE WAVE

If a plane wave is incident normally on a thin film of uniform thickness d (see Fig. 15.1), then the waves reflected from the upper surface interfere with the waves reflected from the lower surface; in this section we will study this interference pattern; why the film should be thin is explained in Sec. 15.7. To observe the interference pattern without obstructing the incident beam, we use a partially reflecting plate G as shown in Fig. 15.1. Such an arrangement also enables us to eliminate the direct beam from reaching the photographic plate P (or the eye). The plane wave may be produced by placing an illuminated pinhole at the focal point of a corrected lens; alternatively, it may just be a beam coming out of a laser.

Let the solid and the dashed lines in Fig. 15.2 represent the positions of the crests[1] (at any particular instant of time) corresponding to the waves reflected from the upper and lower surfaces of the film, respectively; in general, the wave reflected from the lower surface of the film will suffer multiple reflections—the effect of such multiple reflections is neglected (see Chap. 16). Clearly, the wave reflected from the lower surface of the film traverses an additional optical path of $2nd$,

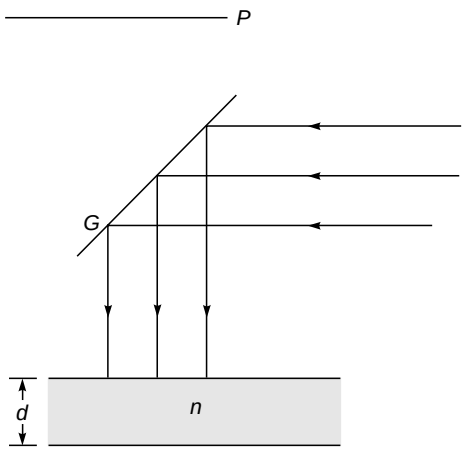


Fig. 15.1 The normal incidence of a parallel beam of light on a thin film of refractive index n and thickness d. G denotes a partially reflecting plate and P represents a photographic plate.

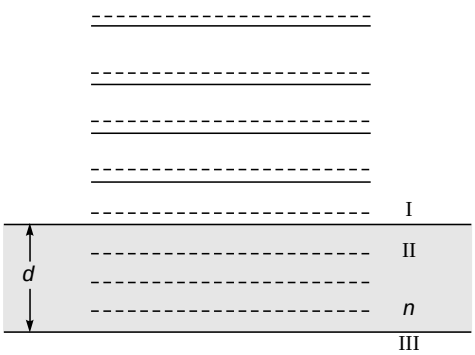


Fig. 15.2 The solid and the dashed lines represent the crests of the waves reflected from the upper surface and from the lower surface of the thin film. Notice that the distance between the consecutive crests inside the film is less than the corresponding distance in medium I.

where n represents the refractive index of the material of the film. Further, if the film is placed in air, then the wave reflected from the upper surface of the film will undergo a sudden change in phase of π (see Sec. 14.12), and as such the conditions for destructive or constructive interference will be given by

$$2nd = m\lambda \qquad \text{destructive interference} \qquad \textbf{(1a)}$$

$$= \left(m + \tfrac{1}{2}\right)\lambda \qquad \text{constructive interference} \qquad \textbf{(1b)}$$

where $m = 0, 1, 2, \ldots$ and λ represents the free space wavelength.

Thus, if we place a photographic plate at P (see Fig. 15.1), then the plate will receive uniform illumination; it will be dark when $2nd = m\lambda$ and bright when $2nd = \left(m + \frac{1}{2}\right)\lambda$, for $m = 0, 1, 2, \ldots$. Instead of placing the photographic plate, if we try to view the film (from the top) with the naked eye, then the film will appear to be uniformly illuminated.

The amplitudes of the waves reflected from the upper and lower surfaces will, in general, be slightly different; and as such the interference will not be completely destructive. However, with appropriate choice of the refractive indices of media II and III, the two amplitudes can be made very nearly equal (see Example 15.1).

For an air film between two glass plates (see Fig. 15.3) no phase change will occur on reflection at the glass-air interface; but a phase change of π will occur on reflection at the air-glass interface and the conditions for maxima and minima will remain the same. On the other hand, if

[1] Notice that the distance between consecutive crests in the film is less than the corresponding distance in air. This is so because the effective wavelength in a medium of refractive index n is λ/n.

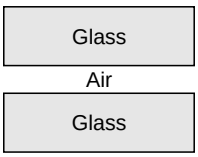


Fig. 15.3 Thin film of air formed between two glass plates.

medium I is crown glass ($n = 1.52$), medium II is an oil of refractive index 1.60, and medium III is flint glass ($n = 1.66$), then a phase change of π will occur at both the reflections and the conditions for maxima and minima will be

$$2nd = \left(m + \tfrac{1}{2}\right)\lambda \qquad \text{minima} \qquad \textbf{(2a)}$$

$$= m\lambda \qquad \text{maxima} \qquad \textbf{(2b)}$$

In general, whenever the refractive index of medium II lies in between the refractive indices of medium I and medium III, then the conditions of maxima and minima are given by Eqs. (2a) and (2b).

We next consider the oblique incidence of the plane wave on the thin film (see Fig. 15.4). Once again, the wave reflected from the upper surface of the film interferes with the wave reflected from the lower surface of the film. The latter traverses an additional optical path Δ, which is given by (see Fig. 15.5)

$$\Delta = n_2(BD + DF) - n_1 BC \qquad \textbf{(3)}$$

where C is the foot of the perpendicular from point F on BG. We will show in the next section that

$$\Delta = 2n_2 d \cos\theta' \qquad \textbf{(4)}$$

where θ' is the angle of refraction.

For a film placed in air, a phase change of π will occur when reflection takes place at point B, and as such, the conditions of destructive and constructive interference are given by

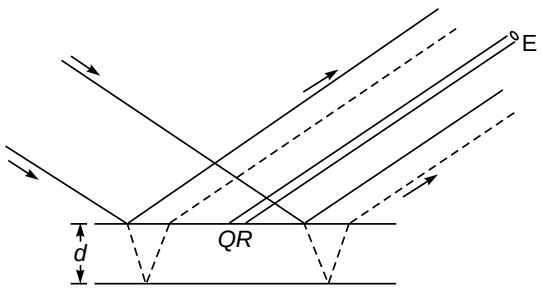


Fig. 15.4 The oblique incidence of a plane wave on a thin film. The solid and dashed lines denote the boundary of the wave reflected from the upper surface and from the lower surface of the film. The eye E receives the light reflected from the region QR.

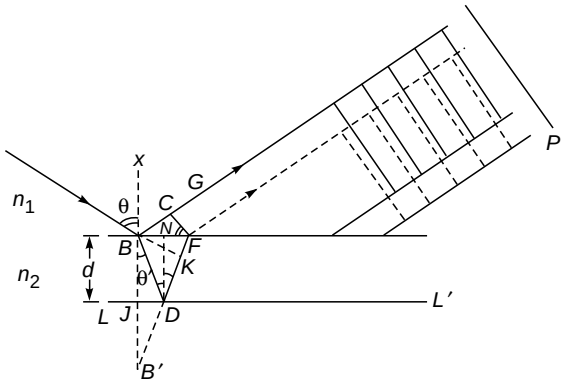


Fig. 15.5 Calculation of the optical path difference between the waves reflected from the upper surface of the film and from the lower surface of the film. The solid and the dashed lines represent the corresponding positions of the crests. P denotes a photographic plate.

$$\Delta = 2n_2 d\cos\theta' = m\lambda \qquad \text{minima} \qquad \textbf{(5a)}$$

$$= \left(m + \tfrac{1}{2}\right)\lambda \qquad \text{maxima} \qquad \textbf{(5b)}$$

If we place a photographic plate at P (see Fig. 15.5), it will receive uniform illumination; if we try to view the film with the naked eye (at position E — see Fig. 15.4), then only light rays reflected from a small position QR of the film will reach the eye. The image formed at the retina will be dark or bright depending on the value of Δ [(see Eq. (5)].

15.3 THE COSINE LAW

In this section we will show that the wave reflected from the lower surface of the film traverses an additional optical path which is given by the

$$\Delta\,[= n_2(BD + DF) - n_1 BC] = 2n_2 d\cos\theta' \qquad \textbf{(6)}$$

Let θ and θ' denote the angles of incidence and refraction, respectively. We drop a perpendicular BJ from point B on the lower surface LL' and extend BJ and FD to point B' where they meet (see Fig. 15.5). Clearly,

$$\angle JBD = \angle BDN = \angle NDF = \theta'$$

where N is the foot of the perpendicular drawn from point D on BF. Now

$$\angle BDJ = \frac{\pi}{2} - \theta'$$

and $$\angle B'DJ = \pi - \left[\left(\frac{\pi}{2} - \theta'\right) + \theta' + \theta'\right] = \frac{\pi}{2} - \theta'$$

Thus $$BD = BD' \quad \text{and} \quad BJ = JB' = d$$

or $BD + DF = B'D + DF = B'F$

Hence $$\Delta = n_2 B'F - n_1 BC \quad \textbf{(7)}$$

Now $\angle CFB = \angle CBX = \theta$

$$BC = BF \sin\theta = \frac{KF}{\sin\theta'} \sin\theta = \frac{n_2}{n_1} KF \quad \textbf{(8)}$$

where K is the foot of the perpendicular from B on $B'F$. Substituting the above expression for BC in Eq. (7), we get

$$\Delta = n_2 B'F - n_2 KF = n_2 B'K$$

or $$\Delta = 2n_2 d \cos\theta' \quad \textbf{(9)}$$

which is known as the *cosine law*.

15.4 NONREFLECTING FILMS

One of the important applications of the thin film interference phenomenon discussed in Sec. 15.2 lies in reducing the reflectivity of lens surfaces; we discuss this in this section. However, for a quantitative understanding of the phenomenon, we will have to assume that when a light beam (propagating in a medium of refractive index n_1) is incident normally on a dielectric of refractive index n_2, then the amplitudes of the reflected and the transmitted beams are related to that of the incident beam through the following relations [see Fig. 15.6(a)]:

$$a_r = \frac{n_1 - n_2}{n_1 + n_2} a_i \quad \textbf{(10a)}$$

$$a_t = \frac{2n_1}{n_1 + n_2} a_i \quad \textbf{(10b)}$$

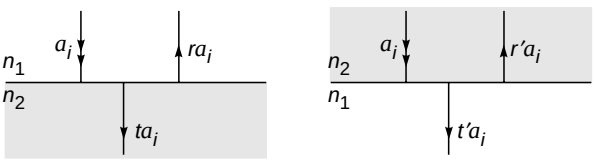


Fig. 15.6 (a) If a plane wave of amplitude a_i, propagating in a medium of refractive index n_1, is incident normally on a medium of refractive index n_2, then the amplitudes of the reflected and the transmitted beams are a_r and a_t, respectively. Similarly, (b) corresponds to the case when the beam (propagating in a medium of refractive index n_2) is incident on a medium of refractive index n_1.

where a_i, a_r, and a_t are the amplitudes of the incident beam, reflected beam, and transmitted beam, respectively. Notice that when $n_2 > n_1$, amplitude a_r is negative, showing that when a reflection occurs at a denser medium, a phase change of π occurs. The amplitude reflection and transmission coefficients r and t are, therefore, given by

$$r = \frac{n_1 - n_2}{n_1 + n_2} \quad \textbf{(11a)}$$

$$t = \frac{2n_1}{n_1 + n_2} \quad \textbf{(11b)}$$

Equations (10) and (11) can be derived using electromagnetic theory; see Eqs. (67) to 72 (with $\theta_1 = \theta_2 = 0$) in Sec.24.2. If r' and t' are the reflection and transmission coefficients where light propagating in a medium of refractive index n_2 is incident on a medium of refractive index n_1 [see Fig. 15.6(b)], then

$$r' = \frac{n_2 - n_1}{n_2 + n_1} = -r \quad \textbf{(12)}$$

$$t' = \frac{2n_2}{n_1 + n_2} \quad \textbf{(13)}$$

and

$$1 - tt' = 1 - \frac{4n_1 n_2}{(n_1 + n_2)^2} = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2 = r^2 \quad \textbf{(14)}$$

Equations (13) and (14) represent Stokes' relations (see Sec. 14.12).

We will now discuss the application of the thin film interference phenomenon in reducing the reflectivity of lens surfaces. We all know that in many optical instruments (such as a telescope) there are many interfaces, and the loss of intensity due to reflections can be severe. For example, for near-normal incidence, the reflectivity of the crown glass surface in air is

$$\left(\frac{n-1}{n+1}\right)^2 = \left(\frac{1.5-1}{1.5+1}\right)^2 \simeq 0.04$$

i.e., 4% of the incident light is reflected. For a dense flint glass $n \simeq 1.67$, and about 6% of light is reflected. Thus, if we have a large number of surfaces, the losses at the interfaces can be considerable. To reduce these losses, lens surfaces are often coated with a $\lambda/4n$ thick nonreflecting film; the refractive index of the film is less than that of the lens. For example, glass ($n = 1.5$) may be coated with a MgF_2 film (see Fig. 15.7), and the film thickness d should be such that[2]

$$2n_f d = \frac{1}{2}\lambda$$

[2] Since the refractive index of the nonreflecting film is greater than that of air and less than that of the glass, abrupt phase change of π occurs at both the reflections. Consequently, when $2nd \cos\theta' = m\lambda$, there will be constructive interference and when $2nd \cos\theta' = \left(m + \frac{1}{2}\right)\lambda$, there will be destructive interference.

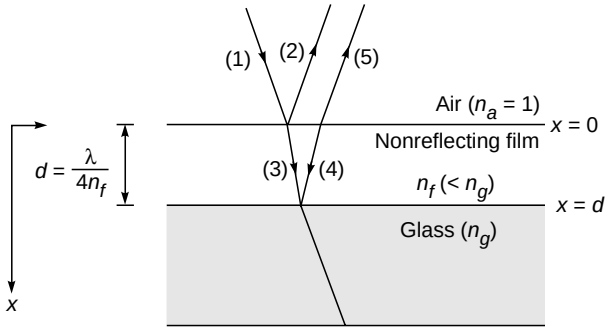


Fig. 15.7 If a film (having a thickness of $\lambda/4n_f$ and having refractive index less than that of the glass) is coated on the glass, then waves reflected from the upper surface of the film destructively interfere with the waves reflected from the lower surface of the film. Such a film is known as a nonreflecting film.

or
$$d = \frac{\lambda}{4n_f} \qquad \textbf{(15)}$$

where we have assumed near-normal incidence [i.e., cos θ′ ≅ 1; see Eq. (9)] and n_f represents the refractive index of the film; for MgF_2, $n_f = 1.38$. Thus, if we assume λ to be 5.0×10^{-5} cm (which roughly corresponds to the center of the visible spectrum), we will have

$$d = \frac{5.0 \times 10^{-5} \text{ cm}}{4 \times 1.38} \approx 0.9 \times 10^{-5} \text{ cm}$$

Figure 15.8 shows a comparison between an eyeglass lens without antireflective coating (top) and a lens with antireflective coating (bottom). Note the reflection of the photographer in the top lens and the tinted reflection in the bottom. We note the following points:

1. Let n_a, n_f, and n_g be the refractive indices of air, nonreflecting film, and glass, respectively. If a is the amplitude of the incident wave, then the amplitudes of the reflected and refracted waves (the corresponding rays shown as 2 and 3 in Fig. 15.7) are

$$-\frac{n_f - n_a}{n_f + n_a}a \qquad \text{and} \qquad \frac{2n_a}{n_f + n_a}a$$

respectively (we have assumed near-normal incidence). The amplitudes of the waves corresponding to rays 4 and 5 are

$$-\frac{2n_a}{n_f + n_a}\frac{n_g - n_f}{n_g + n_f}a$$

and
$$-\frac{2n_a}{n_f + n_a}\frac{n_g - n_f}{n_g + n_f}\frac{2n_f}{n_f + n_a}a$$

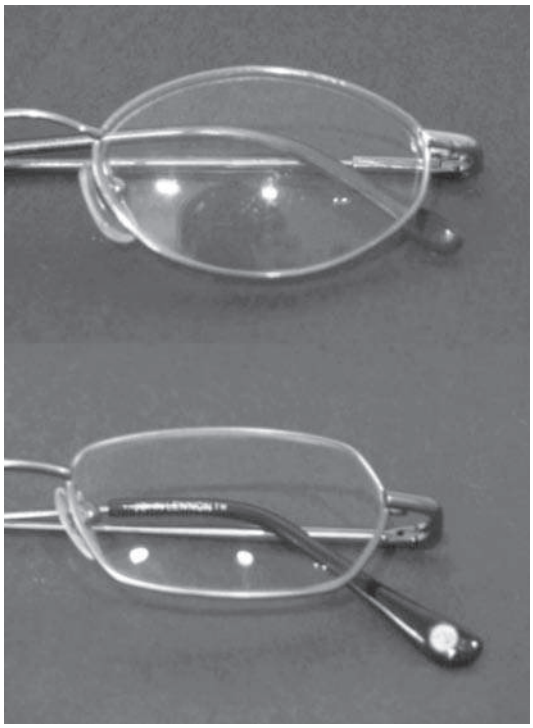


Fig. 15.8 Comparison between an eyeglass lens without antireflective coating (top) and a lens with antireflective coating (bottom). Note the reflection of the photographer in the top lens and the tinted reflection in the bottom. The photograph was taken by Justin Lebar; used with permission from Mr. Lebar. A color photograph appears in the insert at the back of the book.

respectively. Now, for complete destructive interference, the waves corresponding to rays 2 and 5 should have the same amplitude, i.e.,

$$-\frac{n_f - n_a}{n_f + n_a}a = -\frac{2n_a}{n_f + n_a}\frac{n_g - n_f}{n_g + n_f}\frac{2n_f}{n_f + n_a}a \qquad \textbf{(16)}$$

or
$$\frac{n_f - n_a}{n_f + n_a} = \frac{n_g - n_f}{n_g + n_f} \qquad \textbf{(17)}$$

where we have used the fact that $4n_a n_f/(n_f + n_a)^2$ is very nearly equal to unity; for $n_a = 1$ and $n_f = 1.4$,

$$\frac{4n_a n_f}{(n_f + n_a)^2} \approx 0.97$$

On simplification we obtain

$$n_f = \sqrt{n_a n_g} \qquad \textbf{(18)}$$

If the first medium is air, then $n_a = 1$, and with $n_g = 1.66$ (dense flint glass) n_f should be 1.29, and when $n_g = 1.5$

(light crown glass), n_f should be 1.22. We note that the refractive indices of magnesium fluoride and cryolite are 1.38 and 1.36, respectively. Now for a $\lambda/4n$ thick film, the reflectivity will be about

$$\left(\frac{n_f - n_a}{n_f + n_a} - \frac{n_g - n_f}{n_g + n_f}\right)^2 \qquad \textbf{(19)}$$

Thus, for $n_a = 1$, $n_f = 1.38$, and $n_g = 1.5$, the reflectivity will be about 1.3%. In the absence of the film, the reflectivity would have been about 4%. The reduction of reflectivity is much more pronounced for the dense flint glass. This technique of reducing the reflectivity is known as *blooming*.

2. The film is nonreflecting only for a particular value of λ; in Eq. (15) λ was assumed to be 5000 Å. For a polychromatic light, the film's nonreflecting property will be falling off when λ is greater or less than the above value. However, the effect is not serious. For example, for the MgF_2 film on crown glass at 5000 Å, the reflectivity rises by about 0.5% as one goes to either the red or the violet end of the visible spectrum. In Sec. 15.4.2 we will discuss why we should use a $\lambda/4n$ thick film and not $3\lambda/4n$ or $5\lambda/4n$ thick film, although the latter will also give destructive interference for the chosen wavelength.
3. As in the case of Young's double-slit experiment there is no loss of energy; there is merely a redistribution of energy. The energy appears mostly in the transmitted beam.

15.4.1 Mathematical Expressions for the Reflected Waves

We will carry out a bit of mathematical analysis for the antireflecting film shown in Fig. 15.7. We assume that $n_g > n_f > n_a$ and that the x axis is pointing downward with $x = 0$ at the upper surface of the film. The displacement associated with the incident wave (propagating in the $+x$ direction) is given by

$$y_1 = a\cos(\omega t - k_a x); \qquad k_a = \frac{\omega}{c} n_a \qquad \textbf{(20)}$$

Thus at $x = 0$, $y_1 = a\cos\omega t$. The reflected wave (shown as 2) is therefore

$$y_2 = -a|\,r_1\,|\cos(\omega t + k_a x) \qquad \textbf{(21)}$$

where

$$|\,r_1\,| = \left|\frac{n_f - n_a}{n_f + n_a}\right| \qquad \textbf{(22)}$$

is a positive quantity. The minus sign in Eq. (21) represents the sudden phase change of π at $x = 0$. The transmitted wave (shown as 3) is given by

$$y_3 = at_1\cos(\omega t - k_f x) \qquad k_f = \frac{\omega}{c} n_f \qquad \textbf{(23)}$$

where

$$t_1 = \frac{2n_a}{n_f + n_a} \qquad \textbf{(24)}$$

Thus the displacement at $x = d$ (associated with wave 3) is

$$y_3 = at_1\cos(\omega t - k_f d) \qquad \textbf{(25)}$$

Therefore, the wave reflected from the lower surface (wave 4, which would be propagating in the negative x direction) is given by

$$y_4 = -at_1|\,r_2\,|\cos[\omega t + k_f(x - 2d)]$$
$$|\,r_2\,| = \left|\frac{n_g - n_f}{n_g + n_f}\right| \qquad \textbf{(26)}$$

where the phase factor is adjusted such that at $x = +d$ we obtain the phase given by Eq. (25). Wave 5 is therefore given by

$$y_5 = -at_1|\,r_2\,|t_2\cos(\omega t + k_a x - 2k_f d) \qquad \textbf{(27)}$$

Assuming the amplitudes of y_2 and y_5 to be approximately the same, destructive interference (between y_2 and y_5) occurs if

$$2k_f d = \pi, 3\pi, \ldots \qquad \textbf{(28)}$$

or

$$d = \frac{\lambda_f}{4}, \frac{3\lambda_f}{4}, \frac{5\lambda_f}{4}, \ldots \qquad \lambda_f = \frac{\lambda}{n_f} \qquad \textbf{(29)}$$

15.4.2 Rigorous Expressions for Reflectivity

In the above section we considered two-beam interference and neglected multiple reflections at the lower and upper surfaces. The effect of multiple reflections will be discussed in Sec. 16.2; however, such an effect is automatically taken into account when we solve Maxwell's equations incorporating the appropriate boundary conditions. In Sec. 24.4 we will carry out such an analysis and will show that the reflectivity (at normal incidence) of a dielectric film of the type shown in Fig. 15.7 is given by [see Eq. (97) of Chap. 24][3]

$$R = \frac{r_1^2 + r_2^2 + 2r_1 r_2\cos 2\delta}{1 + r_1^2 r_2^2 + 2r_1 r_2\cos 2\delta} \qquad \textbf{(30)}$$

[3] Equation (30) is actually valid even for oblique incidence with r_1, r_2, and δ defined appropriately (see Sec. 24.4).

where

$$r_1 = \frac{n_a - n_f}{n_a + n_f} \quad \text{and} \quad r_2 = \frac{n_f - n_g}{n_f + n_g} \tag{31}$$

represent the Fresnel reflection coefficients at the first and second interface, respectively, and

$$\delta = \frac{2\pi}{\lambda} n_f d \tag{32}$$

d is the thickness of the film, and, as before, λ is the free space wavelength. Elementary differentiation shows us that $dR/d\delta = 0$ when $\sin 2\delta = 0$. Indeed for $r_1 r_2 > 0$,

$$\cos 2\delta = -1 \quad \text{(minima)} \tag{33}$$

represents the condition for minimum reflectivity, and when this condition is satisfied, the reflectivity is given by

$$R = \left(\frac{r_1 - r_2}{1 - r_1 r_2}\right)^2 = \left(\frac{n_a n_g - n_f^2}{n_a n_g + n_f^2}\right)^2 \tag{34}$$

where we have used Eq. (31). Thus the film is nonreflecting when

$$n_f = \sqrt{n_a n_g}$$

consistent with Eq. (18). Now, the condition $\cos 2\delta = -1$ implies

$$2\delta = \frac{4\pi}{\lambda} n_f d = (2m + 1)\pi \quad m = 0, 1, 2, \ldots \tag{35}$$

or

$$d = \frac{\lambda}{4n_f}, \frac{3\lambda}{4n_f}, \frac{5\lambda}{4n_f}, \ldots \tag{36}$$

In Fig. 15.9 we have plotted the reflectivity as a function of δ for

$$n_a = 1 \qquad n_g = 1.5 \tag{37}$$

and

$$n_f = \sqrt{n_a n_g} \simeq 1.225$$

As expected, R is maximum ($\approx 4\%$) when $\delta = 0, \pi, 2\pi, \ldots$, and the film is antireflecting ($R = 0$) when $\delta = \pi/2, 3\pi/2, \ldots$, implying $d = \lambda/4n_f, 3\lambda/4n_f, \ldots$. As an example, suppose that we wish to make the film antireflecting at $\lambda = 6000$ Å; then from Eq. (26) the thickness of the film could be

$$1224.7 \text{ Å} \quad \text{or} \quad 3674.2 \text{ Å} \quad \text{or} \quad 6123.7 \text{ Å}, \ldots$$

In Fig. 15.9(b) we have plotted the reflectivity as a function of wavelength for $d = 1224.7$ and 3674.2 Å. As can be seen, for $d = \lambda/4n_f$, the minimum is broad and the reflectivity small for the entire range of the visible spectrum. Thus for antireflecting coating, the smallest film thickness is always preferred. For $n_a = 1$, $n_g = 1.5$, and $n_f = 1.38$, the reflectivity [according to Eq. (34)] comes out to be 1.4%, which is quite close to the result obtained by using the approximate theory described earlier [see Eq. (19)].

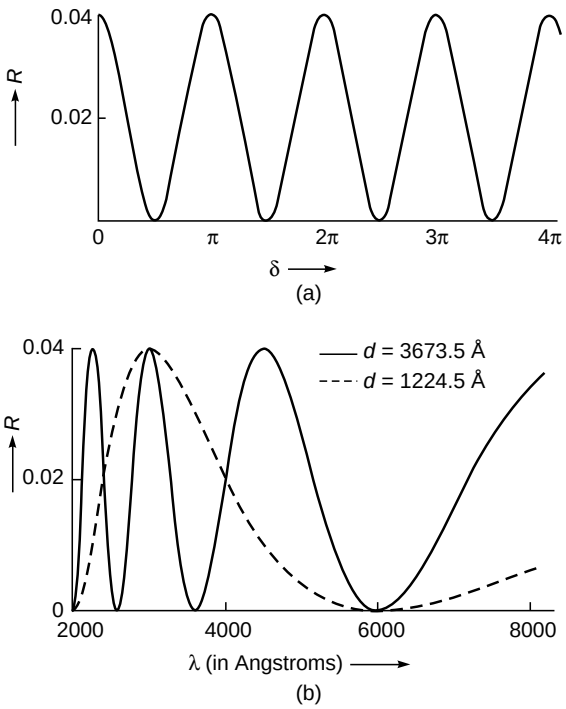


Fig. 15.9 (a) Variation of the reflectivity of a film as a function of δ ($= 2\pi n_f d/\lambda$) for $n_a = 1$, $n_g = 1.5$, and $n_f = \sqrt{n_a n_g} \simeq 1.225$. Notice that the reflectivity is zero for $\delta = \pi/2, 3\pi/2, 5\pi/2, \ldots$. (b) Wavelength variation of the reflectivity for a film of thickness 1224.5 Å (dashed curve) and of thickness 3673.5 Å (solid curve) with $n_a = 1$, $n_g = 1.5$, and $n_f = \sqrt{n_a n_g} \approx 1.225$. Notice that both films are non-reflecting at 6000 Å.

15.5 HIGH REFLECTIVITY BY THIN FILM DEPOSITION

Another important application of the thin film interference phenomenon is the converse of the procedure just discussed; i.e., the glass surface is coated by a thin film of suitable material to increase the reflectivity. The film thickness is again $\lambda/4n_f$, where n_f represents the refractive index of the film; however, the film is such that its refractive index is greater than that of the glass. Consequently, an abrupt phase change of π occurs only at the air-film interface, and the beams reflected from the air-film interface and the film-glass interface constructively interfere. For example, if we

consider a film of refractive index 2.37 (zinc sulfide), then the reflectivity is $(2.37 - 1)^2/(2.37 + 1)^2$, i.e., about 16%. In the presence of a glass surface of refractive index 1.5 (light crown glass), the reflectivity will become (see the analysis in Sec. 15.4)

$$\left[-\frac{2.37-1}{2.37+1}-\frac{4\times1\times2.37}{(3.37)^2}\times\frac{2.37-1.5}{2.37+1.5}\right]^2$$

which gives about 35%. Note that if the difference between the refractive indices of the film and the glass is increased, then the reflectivity will also increase.

We can again use Eq. (30) to calculate the high reflectivity obtained by thin film deposition. Indeed when $n_a < n_f$ and $n_f > n_g$, $r_1 r_2 < 0$ [(see Eq. (31)] and

$$\cos 2\delta = -1 \qquad \text{(maxima)} \tag{38}$$

represents the condition for *maximum* reflectivity. The *maximum* value of the reflectivity is given by

$$R = \left(\frac{r_1 - r_2}{1 - r_1 r_2}\right)^2 \tag{39}$$

For $n_a = 1.0$, $n_f = 2.37$, and $n_g = 1.5$, we have

$$r_1 \simeq -0.407 \qquad r_2 \simeq 0.225$$

Elementary calculations show that the reflectivity is about 33% which compares well with the value of 35% obtained by using the approximate theory described earlier.

15.6 REFLECTION BY A PERIODIC STRUCTURE[4]

In Sec. 15.4 we showed that a film of thickness $\lambda/4n_f$, where λ is the free space wavelength and n_f is the film refractive index (which lies between the refractive indices of the two surrounding media), acts as an antireflection layer. This is due to the destructive interference occurring between the waves reflected from the top and bottom interfaces. In Sec. 15.5 we showed that if the refractive index of the film was smaller (or greater) than those of both the surrounding media, then in such a case, in addition to the phase difference due to the additional path traveled by the wave reflected from the lower interface, there would be an extra phase difference of π between the two reflected waves. Thus, in such a case a film of thickness $\lambda/4n_f$ would increase the reflectivity rather than reduce it.

We now consider a medium consisting of alternate layers of high and low refractive indices of $n_0 + \Delta n$ and $n_0 - \Delta n$ of equal thickness d [see Fig. 15.10(a)]. Such a medium is called

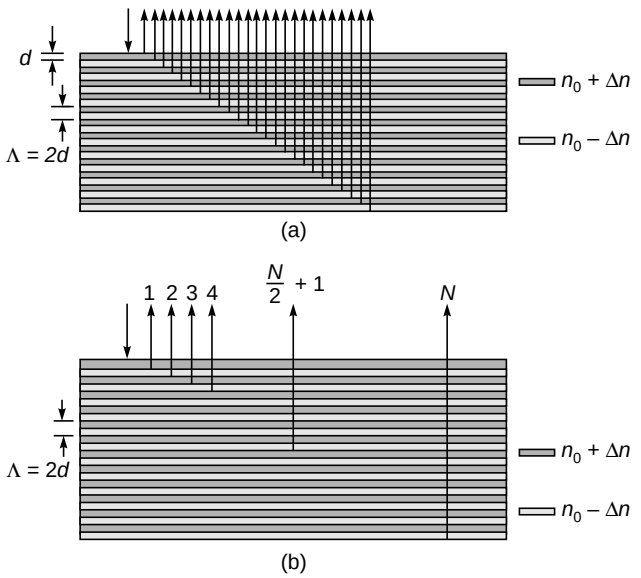


Fig. 15.10 (a) Reflection from a periodic structure consisting of alternate layers of refractive indices $n_0 + \Delta n$ and $n_0 - \Delta n$, each of thickness $d = \lambda_B/4n_0$. (b) If we choose a wavelength $\lambda_B + \Delta\lambda$ such that reflections from layer 1 and layer $N/2 + 1$ are out of phase, reflections from layer 2 and layer $N/2 + 2$ are out of phase, etc., and finally reflections from layers $N/2$ and N are out of phase, then the reflectivity will be zero.

[4] This section has been very kindly written by Prof. K. Thyagarajan.

a periodic medium, and the spatial period of the refractive index variation is given by

$$\Lambda = 2d$$

Now if $\Delta n << n_0$ and if we choose the thickness of each layer to be

$$d = \frac{\lambda}{4n_0} \approx \frac{\lambda}{4(n_0 + \Delta n)} \approx \frac{\lambda}{4(n_0 - \Delta n)}$$

then the reflections arising out of individual reflections from the various interfaces will all be in phase and should result in a strong reflection. Thus for strong reflection at a chosen (free space) wavelength λ_B, the period of the refractive index variation should be

$$\Lambda = 2d = \frac{\lambda_B}{2n_0} \qquad \textbf{(40)}$$

This is referred to as the Bragg condition and is very similar to the Bragg diffraction of X-rays from various atomic layers (see Sec. 18.9). Equation (40) corresponds to the Bragg condition for normal incidence. The quantity λ_B is often referred to as the Bragg wavelength.

As an example, we consider a periodic medium comprising alternate layers of refractive indices 1.51 and 1.49; i.e., $n_0 = 1.50$ and $\Delta n = 0.01$. If we require a strong reflectivity at $\lambda = \lambda_B = 5500$ Å, then the required periodicity is

$$\Lambda = \frac{5500}{2 \times 1.5} \text{ Å} \approx 1833 \text{ Å}$$

If the periodic medium is made up of 100 layers (i.e., 50 periods), then we may approximate the total resultant amplitude to be

$$100 \times \frac{\Delta n}{n_0} \approx \frac{1}{1.5}$$

where $\Delta n/n_0$ is the amplitude reflection coefficient at each interface. The above estimation is only an approximation which is valid when $N\,\Delta n/n_0 << 1$, i.e., for small reflectivities; here we are just trying to obtain a crude estimate of the total reflectivity. Thus the reflectivity at 5500 Å should be

$$R \approx \left(\frac{1}{1.5}\right)^2 \quad \Rightarrow \quad R \approx 44\% \qquad \textbf{(41)}$$

Figure 15.11 shows an actual calculated value of the reflectivity as a function of wavelength (using rigorous electromagnetic theory—see Ref. 6) for a periodic medium with $n_0 = 1.5$, $\Delta n = 0.01$, and $d = \lambda_B/4n_0$, and consisting of 100 layers. Note that the actual calculation predicts a reflectivity of about 33% which compares well with our crude estimate of 44%!

One notices from Fig. 15.11 that as we move away from the central wavelength ($\lambda_B = 2n_0\Lambda$), the reflectivity of the periodic medium falls off sharply. One can indeed obtain an

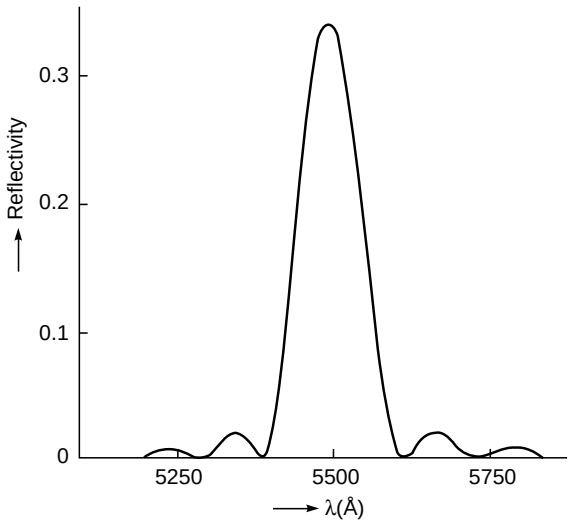


Fig. 15.11 The exact variation of reflectivity with wavelength of a 100-layer periodic structure with $n_0 = 1.5$, $\Delta n = 0.01$, and $\Lambda = 2d = 1833$ Å. The peak reflectivity appears at $\lambda = \lambda_B = 4n_0d$. (Adapted from Ref. 6.)

approximate expression for the wavelength deviation $\Delta\lambda$ from λ_B which will produce a zero reflectivity. To do this, we first note that at λ_B ($= 2n_0\Lambda$), the waves reflected from each of the N individual layers are all in phase, leading to a strong reflection. If we move away from λ_B, then the individual waves reflected from the various layers will not be in phase, and thus the reflectivity reduces. If we choose a wavelength $\lambda_B + \Delta\lambda$ such that the reflections from layer 1 and layer $N/2 + 1$, from layers 2 and $N/2 + 2$, and so on, up to the reflections from layers $N/2$ and N are out of phase [see Fig. 15.10(b)], then the reflectivity will be zero. For reflection from each of the top $N/2$ layers, there is a reflection from a corresponding lower $N/2$ layer which is out of phase. (The argument is very similar to that used for obtaining the direction of minima in the diffraction pattern of a slit—see Sec. 18.2 and Fig. 18.5). Thus when we move from λ_B to $\lambda_B + \Delta\lambda$, the waves reflected from the first layer and $(N/2 + 1)$st layer should have an additional phase difference of π. Thus,

$$\frac{2\pi}{\lambda_B} n_0 \frac{N\Lambda}{2} - \frac{2\pi}{(\lambda_B + \Delta\lambda)} n_0 \frac{N\Lambda}{2} = \pi \qquad \textbf{(42)}$$

where the first term on the LHS is simply the phase difference at λ_B between reflections 1 and $N/2 + 1$ due to the extra path traveled by the latter wave, and the second term is that at $\lambda_B + \Delta\lambda$. Assuming $\Delta\lambda << \lambda_B$, we have

$$\frac{2\pi}{\lambda_B^2} \frac{n_0 N\Lambda}{2} \Delta\lambda = \pi$$

or $$\frac{\Delta\lambda}{\lambda_B} \approx \frac{\lambda_B}{n_0 N\Lambda} = \frac{\lambda_B}{2n_0 L} \qquad \textbf{(43)}$$

where we used Eq. (40) and $L = N\Lambda/2$ is the total thickness of the periodic medium. For the example shown in Fig. 15.11, we have

$$\Delta\lambda \approx 110 \text{ Å} \qquad \textbf{(44)}$$

which compares very well with the actual value in Fig. 15.11. Thus if the incident wave is polychromatic (such as white light), the reflected light may have a high degree of monochromaticity. This is indeed the principle used in white light holography.

The periodic medium discussed above finds wide applications in high-reflectivity multilayer coatings, volume holography, fiber Bragg gratings, etc.

15.6.1 Fiber Bragg Gratings

A periodic structure discussed above has a very important application in the working of a fiber Bragg grating (usually abbreviated FBG). We will discuss the optical fiber in Chap. 27; it may suffice here to mention that an optical fiber is a cylindrical structure consisting of a central dielectric core cladded by a material of slightly lower refractive index (see Fig. 27.7). The guidance of the light beam takes place because of total internal reflections at the core-cladding interface (see Chaps. 27 and 29 for details). The cladding material is pure silica, and the core is usually silica doped with germanium; the doping results in a slightly higher refractive index. Now, when a germanium-doped silica core fiber is exposed to ultraviolet radiation (with wavelength around 0.24 μm), the refractive index of the germanium-doped region increases; this is due to the phenomenon known as *photosensitivity* which was discovered by Kenneth Hill in 1974. The refractive index increase can be as large as 0.001 in the core of the fiber. If the fiber is exposed to a pair of interfering UV beams (see Fig. 15.12), then we obtain an interference pattern similar to that shown in Fig.14.11(b). In regions of constructive interference, the refractive index increases. Since the fringe width depends on the angle between the interfering beams, the period of the grating can be controlled by choosing the angle between the interfering beams (see Example 14.5). Thus exposing a germanium-doped silica fiber to the interference pattern formed between two UV beams leads to the formation of a periodic refractive index variation in the core of the fiber.

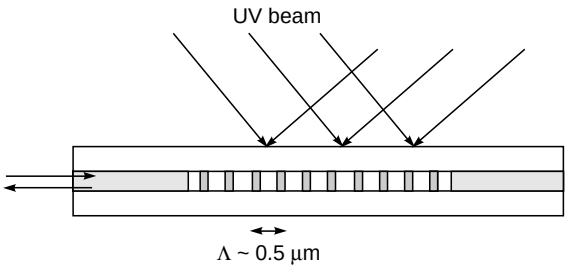


Fig. 15.12 A fiber Bragg grating (usually abbreviated FBG) is produced by allowing two beams to produce an interference pattern.

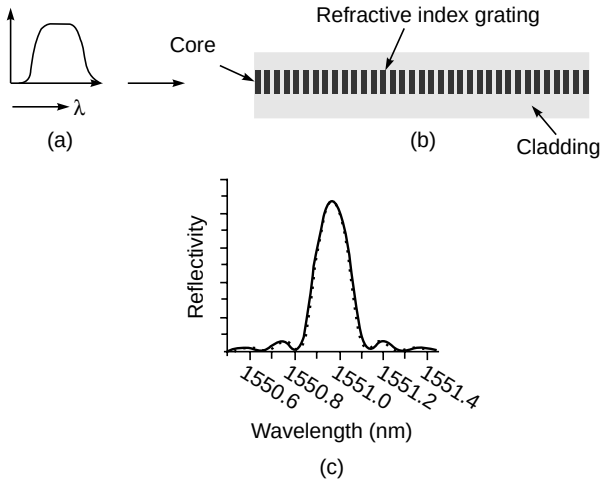


Fig. 15.13 (a) The broad spectrum of the light wave incident on the FBG shown in (b). (c) The spectrum of the reflected wave; solid line shows the calculated spectrum; and the dots show the experimentally measured values of the FBG fabricated at CGCRI, Kolkata. (Figure courtesy Dr. S. Bhadra and Dr. S. Bandyopadhyay of CGCRI, Kolkata.)

We consider a polychromatic beam incident on the fiber as shown in Fig. 15.13. As discussed above, the reflection from the periodic structure will add up in phase when

$$\lambda = \lambda_B = 2\Lambda n_0 \qquad \text{Bragg condition} \qquad \textbf{(45)}$$

where Λ represents the period of the refractive index variation (see Fig. 15.12). Figure 15.13(a) shows the frequency spectrum of the incident polychromatic beam, and the corresponding spectrum of the reflected beam is shown in Fig. 15.13(b). Figure 15.13(c) shows a typical frequency spectrum of the reflected wave; solid line shows the calculated spectrum [using Eq.(3) of App. C], and the dashed curve shows the experimentally measured values. For a silica fiber $n_0 \approx 1.46$, and for the periodic structure to be reflecting at λ = 1550 nm we must have

$$\Lambda = \frac{\lambda_B}{2n_0} = \frac{1550 \text{ nm}}{2 \times 1.46} \approx 0.531 \text{ μm} \qquad \textbf{(46)}$$

The corresponding peak reflectivity is given by

$$R_p = \tanh^2\left(\frac{\pi \Delta n L}{\lambda_B}\right) \approx 0.855 \qquad \textbf{(47)}$$

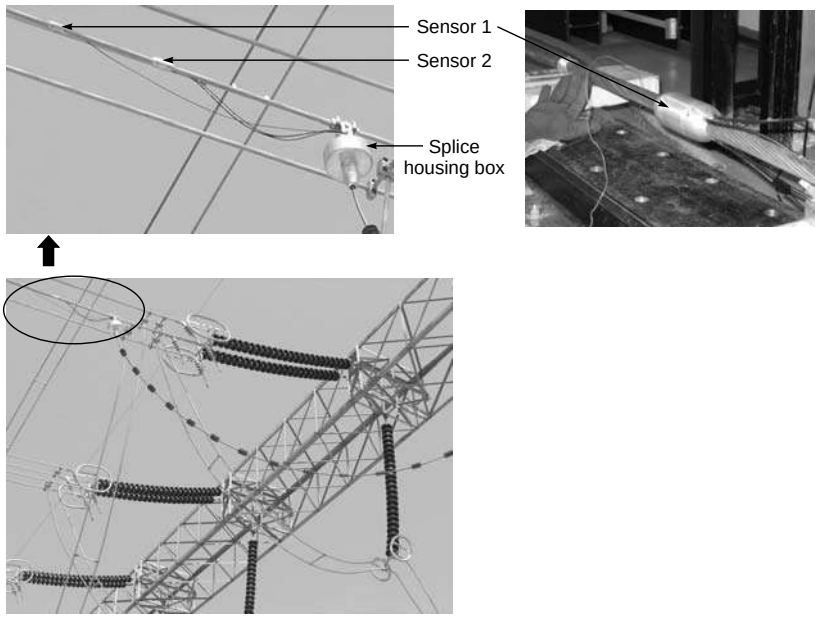


Fig. 15.14 FBG-based temperature sensor system on 400 kV power conductor at Subhashgram substation (near Kolkata) of Powergrid Corporation of India. Photo courtesy of Dr. Kamal Dasgupta and Dr. Tarun Gangopadhyay, CGCRI, Kolkata. Color photographs appear in the insert at the back of the book.

Fig. 15.15 The substation of Powergrid Corporation of India (near Kolkata, India) where the FBG temperature sensors have been installed. In the photograph, the author is with Dr. Tarun Gangopadhyay and Dr. Kamal Dasgupta of CGCRI, Kolkata. A color photograph appears in the insert at the back of the book. Photo courtesy of Dr. Kamal Dasgupta and Dr. Tarun Gangopadhyay, CGCRI, Kolkata.

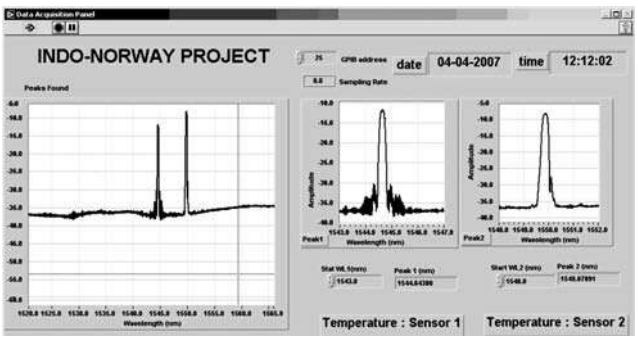


Fig. 15.16 A typical reflection spectrum from the two FBG sensors shown in Fig. 15.14. (Photo courtesy of Dr. Kamal Dasgupta and Dr. Tarun Gangopadhyay, CGCRI, Kolkata.)

where we have assumed $\Delta n \approx 4 \times 10^{-4}$ and $L = 2$ mm. The corresponding bandwidth is given by (see Appendix C)

$$\frac{\Delta\lambda_0}{\lambda_B} \approx \frac{\lambda_B}{2n_0 L}\sqrt{1+\left(\frac{\Delta n L}{\lambda_B}\right)^2} \qquad \textbf{(48)}$$

giving $\Delta\lambda_0 \approx 0.5$ nm. As can be seen from the above equations, that the bandwidth (i.e., the monochromaticity of the reflected wave) and the peak reflectivity are determined by Δn and L.

Because of the extremely small bandwidth of the reflected spectrum, FBGs are being extensively used as sensors. For example, a small increase in the temperature will increase the period of the grating which will result in an increase of the peak wavelength. Because silica is a dielectric material, FBG-based temperature sensors become particularly useful in places where there is high voltage. Figure 15.14 shows the FBG-based temperature sensor system on a 400 kV power conductor at an electric power substation (see Fig. 15.15). Figure 15.16 shows a typical reflection spectrum and the temperature recorded from the two FBG sensors shown in

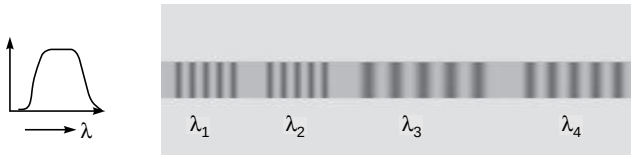


Fig. 15.17 (a) The broad spectrum of the light wave incident on a fiber on which four gratings have been written, as shown in (b). Each grating has a slightly different period because of which each will have peak reflectivity at a different wavelength.

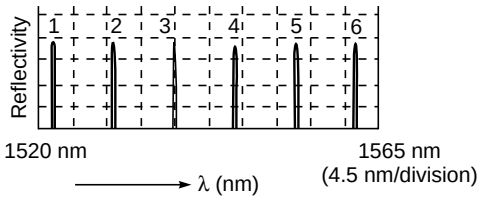


Fig. 15.18 The actual spectrum of the reflected wave from a fiber on which six gratings have been written, each having a slightly different period. The wavelengths at which peak reflectivity occurs are 1522.030 nm, 1529.915 nm, 1537.950 nm, 1545.955 nm, 1553.990 nm, and 1561.895 nm. The gratings were fabricated at CGCRI, Kolkata. (Figure courtesy of Dr. Kamal Dasgupta of CGCRI, Kolkata.)

Fig. 15.14; for the two sensors, peak reflectivity occurs at 1544.6438 and 1545.8789 nm, respectively.

One of the main advantages of the FBG sensor is the fact that several gratings can be written on a single fiber, as shown in Fig. 15.17. Each grating has a different period and therefore a specific wavelength at which peak reflectivity occurs. If such a distributed sensor is put inside a bridge, one can measure the strain corresponding to the particular region. In fact for many newly constructed bridges, FBG sensors are put at various places. Figure 15.18 shows the actual spectrum of the reflected light beam from a fiber on which six gratings have been written, each having a slightly different period. The wavelengths at which peak reflectivity occur are

1522.030 nm with 3 dB bandwidth of 0.240 nm
1529.915 nm with 3 dB bandwidth of 0.230 nm
1537.950 nm with 3 dB bandwidth of 0.240 nm
1545.955 nm with 3 dB bandwidth of 0.230 nm
1553.990 nm with 3 dB bandwidth of 0.240 nm
1561.895 nm with 3 dB bandwidth of 0.230 nm

The 3 dB bandwidth means that, for example, for the first grating, the reflectivity will fall by 50% at $\lambda \approx 1521.910$ and 1522.150 nm. Each grating has a length of 1 cm. Thus for the first grating with $\lambda_B = 1522.030$ nm, we get

$$\Lambda = \frac{\lambda_B}{2n_0} = \frac{1522.03 \text{ nm}}{2 \times 1.46} \approx 0.5212 \ \mu\text{m}$$

Further, assuming $L \approx 0.01$ m and $n_0 \approx 1.46$, Eq. (48) gives (one has to be careful with the units!)

$$\frac{0.240}{1522} \approx \frac{1.522 \times 10^{-6}}{1.46 \times 0.01} \sqrt{1 + \left(\frac{\Delta n \times 0.01}{1.522 \times 10^{-6}} \right)^2}$$

giving $\Delta n \approx 1.7 \times 10^{-4}$.

15.7 INTERFERENCE BY A PLANE PARALLEL FILM WHEN ILLUMINATED BY A POINT SOURCE

In Sec. 15.2 we considered the incidence of a parallel beam of light on a thin film and discussed the interference produced by the waves reflected from the upper and lower surfaces of the film. We will now consider the illumination of the film by a point source of light. Once again, to observe the film without obstructing the incident beam, we will use a partially reflecting plate *G* as shown in Fig. 15.19. However, to study the interference pattern, we may *assume* the point source *S* to be right above the film (see Fig. 15.20) such that the distance *SK* (in Fig. 15.20) is equal to *SA* + *AK* (in Fig. 15.19); *KA* (in Fig. 15.19) and *KS* (in Fig. 15.20) are normal to the film.

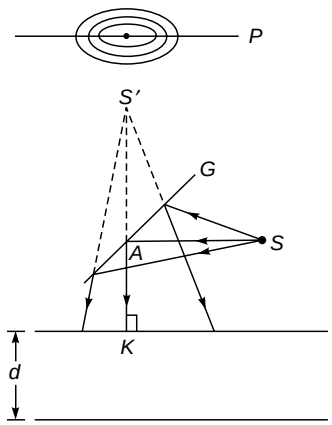


Fig. 15.19 Light emanating from a point source *S* is allowed to fall on a thin film of thickness *d*. *G* is a partially reflecting plate, and *P* represents the photographic plate. On the photographic plate, circular fringes are obtained.

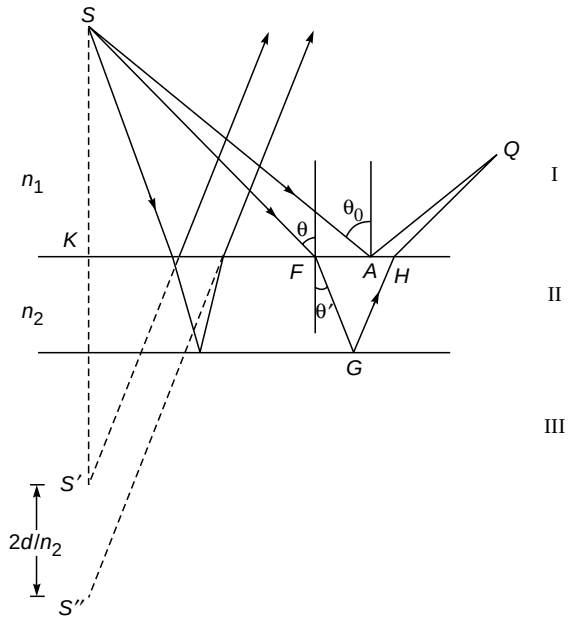


Fig. 15.20 If light emanating from a point source S is incident on a thin film, then the interference pattern produced in region I is approximately same as would have been produced by two coherent point sources S' and S'' (separated by a distance $2d/n_2$, where d represents the thickness of the film and n_2 represents the refractive index of the film).

Obviously, the waves reflected from the upper surface of the film will appear to emanate from point the S' where

$$KS' = KS \qquad \textbf{(49)}$$

(see Fig. 15.20). Further, simple geometric considerations will show that the waves reflected from the lower surface appear to emanate from point S'', where

$$KS'' \simeq KS + 2d/n_2 \qquad \textbf{(50)}$$

(see Fig. 15.20). Equation (50) is valid only for near-normal incidence; this is a consequence of the fact that the image of a point source produced by a plane refracting surface is not perfect. Thus, at least for near-normal incidence, the interference pattern produced in region I (see Fig. 15.20) will be very nearly the same as that produced by two point coherent sources S' and S'' (which is the double-hole experiment of Young discussed in the previous chapter). This is not identical to Young's pattern because S'' is not a perfect image of point S. For large angles of incidence, the waves reflected from the lower surface will appear to emanate from a point which will be displaced from S''. Thus, if we put a photographic plate P (see Fig. 15.19), we will, in general, obtain interference fringes. The intensity of an arbitrary point Q (in Fig. 15.20) will be determined by the following relations:

$$\Delta = \left(m + \frac{1}{2}\right)\lambda \qquad \text{maxima} \qquad \textbf{(51a)}$$

$$= m\lambda \qquad \text{minima} \qquad \textbf{(51b)}$$

where

$$\Delta = [n_1 SF + n_2(FG + GH) + n_1 HQ] - [n_1(SA + AQ)] \qquad \textbf{(52)}$$

represents the optical path difference and we have assumed that in one of the reflections, an abrupt phase change of π occurs; n_1 and n_2 are the refractive indices of media I and II, respectively. The above conditions are rigorously correct; i.e., valid even for large angles of incidence. Further, it can be shown that for near-normal incidence,

$$\Delta \simeq 2n_2 d \cos\theta' \qquad \textbf{(53)}$$

A more rigorous calculation shows (see Ref. 7)

$$\Delta \simeq 2n_2 d \cos\theta' \left[1 - \frac{n_1^2 \sin\theta \cos\theta}{n_2^2 - n_1^2 \sin^2\theta}\left(\frac{\theta_0 - \theta}{2}\right)\right] \qquad \textbf{(54)}$$

where the angles θ, θ_0, and θ' are defined in Fig. 15.20.

Now, if we put a photographic plate [parallel to the surface of the film (see Fig. 15.20)], we will obtain dark and bright concentric rings (see Example 14.6).[5] On the other hand, if we view the film with the naked eye then, for a given position of the eye, we will be able to see only a very small portion of the film. From examples, with the eye at the position E and the point source at S, only a portion of the film around the point B will be visible [see Fig. 15.21(a)], and this point will appear to be dark or bright as the optical path difference

$$\Delta = n_1 SQ + n_2(QA + AB) - n_1 SB$$

is $m\lambda$ or $\left(m + \frac{1}{2}\right)\lambda$. Further, using a method similar to the one described in Sec. 15.3, we obtain

$$\Delta \simeq 2n_2 d \cos\theta' \qquad \textbf{(55)}$$

Instead of looking at the film, if the eye is focused at infinity, then the interference is between the rays which are derived from a single incident ray by reflection from the upper and lower surfaces of the film [see Fig. 15.21(b)]. For example, rays PM and QR, which focus at the point O of the retina, are derived from the single ray SP, and rays $P'M'$ and $Q'R'$, which focus at a different point O' on the retina, are derived from ray SP'. Since the angles of refraction θ'_1 and θ'_2 (for these two sets of rays) will be different, points O and O' will, in general, not have the same intensity.

[5] If the point source is taken far away, then it can be easily seen that the rings will spread out, and in the limit of the point source being taken to infinity (i.e., incidence of a parallel beam), the photographic plate will be uniformly illuminated.

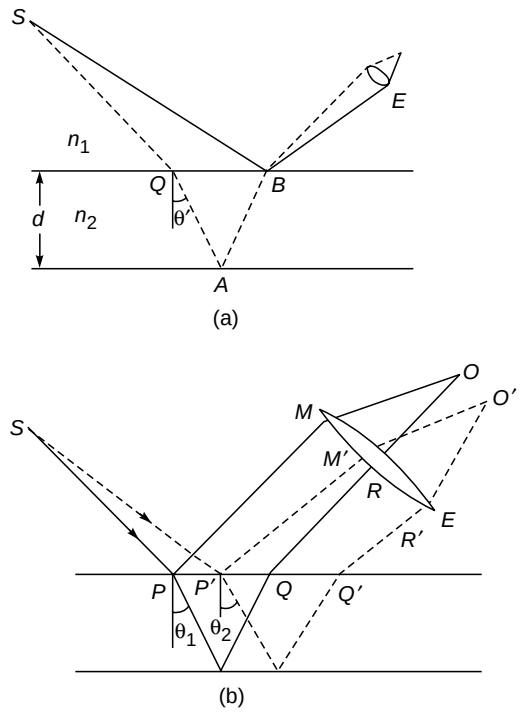


Fig. 15.21 Light emanating from a point source S is incident on a thin film; (a) if the film is viewed by the naked eye E then the point B will appear to be dark if the optical path $[\{n_1\, SQ + n_2\, (QA + AB)\} - n_1\, SB]$ is $m\lambda$, and bright if the optical path is $\left(m + \frac{1}{2}\right)\lambda$. (b) If the eye is focused for infinity then it receives parallel rays from different directions corresponding to different values of the angles of refraction θ′ (and hence different values of the optical path difference).

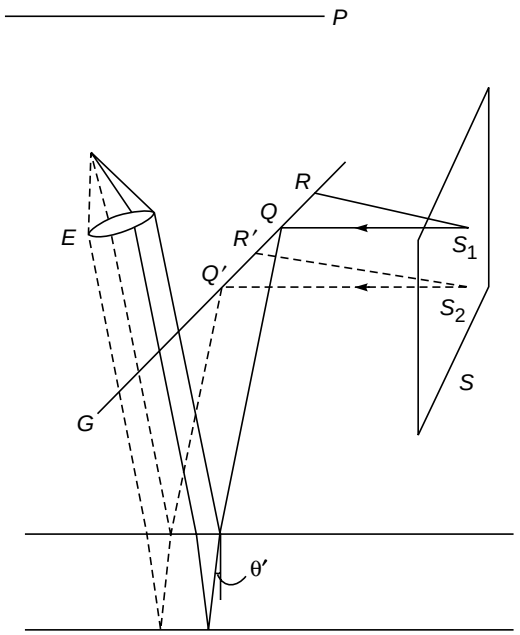


Fig. 15.22 Light emanating from an extended source illuminates a thin film. G represents the partially reflecting plate, and P represents the photographic plate. The eye E is focused at infinity.

We next consider the illumination by an extended source of light S (see Fig. 15.22). Such an extended source may be produced by illuminating a ground glass plate by a sodium lamp. Each point on the extended source will produce its own interference pattern on the photographic plate P; these will be displaced with respect to one another. Consequently, no definite fringe pattern will appear on the photographic plate. However, if we view the film with the naked eye, rays from all points of the film will reach the eye. If the eye is focused at infinity, then parallel light coming in a particular direction reaching the eye would have originated from nearby points of the extended source, and the intensity produced on the retina would depend on the value of $2nd \cos \theta'$ which is the same for all parallel rays such as S_1Q, S_2Q', etc. (see Fig. 15.22). Rays emanating in a different direction (such as S_1R, S_2R', etc.) would correspond to a different value of θ′ and would focus at a different point on the retina. Since θ′ is constant over the circumference of a cone (whose axis is normal to the film and whose vertex is at the eye), the eye will see dark and bright concentric rings, with the center lying along the direction $\theta' = 0$. Such fringes, produced by a film of uniform thickness, are known as Haidinger fringes. They are also known as fringes of equal inclination because the changes in the optical path are due to the changes in the direction of incidence and hence in the value of θ′. In Sec. 15.10 we will discuss the Michelson interferometer where such fringes are observed.

15.8 INTERFERENCE BY A FILM WITH TWO NONPARALLEL REFLECTING SURFACES

Untill now we have assumed the film to be of uniform thickness. We will now discuss the interference pattern produced by a film of varying thickness. Such a film may be produced by a wedge which consists of two nonparallel plane surfaces [see Fig. 15.23(a)].

We first consider a parallel beam of light incident normally on the upper surface of the film [see Fig. 15.23(a)]. In Fig. 15.23(b) the successive positions of the crests (at a particular instant of time) reflected from the upper surface and from the lower surface of the film are shown by solid and dashed lines, respectively. Obviously, a photographic plate P will record straight-line interference fringes which will be parallel to the edge of the wedge (the edge is the line passing

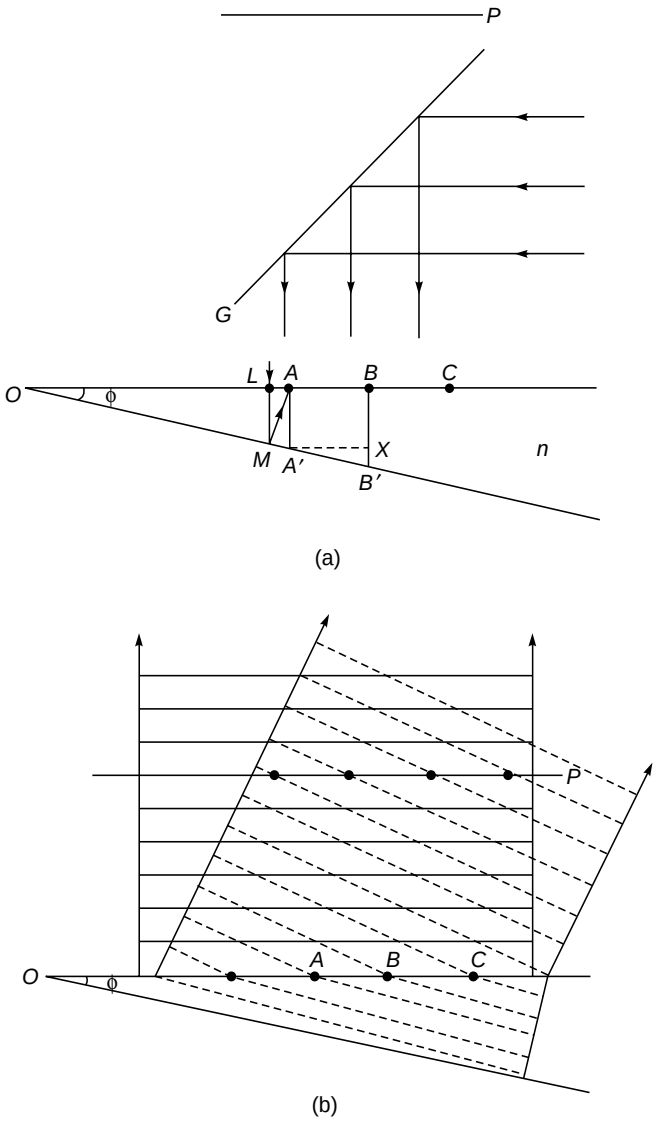


Fig. 15.23 (a) A parallel beam of light incident on a wedge. (b) The solid and the dashed lines represent the positions of the crests (at a particular instant of time) corresponding to the waves reflected from the upper surface and from the lower surface respectively. The maxima will correspond to the intersection of the solid and dashed lines. The fringes will be perpendicular to the plane of the paper.

through point O and perpendicular to the plane of the paper). The dots in the figure indicate the positions of maxima. To find the distance between two consecutive fringes on the film, we note that for point A to be bright[6]

$$n(LM + MA) = \left(m + \tfrac{1}{2}\right)\lambda \qquad m = 0, 1, 2, \ldots \qquad \textbf{(56)}$$

[see Fig. 15.23(a)]. However, when the wedge angle ϕ is very small (which is indeed the case for practical systems),

$$LM + MA \approx 2AA'$$

where AA' represents the thickness of the film at A. Thus the condition for point A to be bright is

$$2nAA' \approx \left(m + \tfrac{1}{2}\right)\lambda \qquad \textbf{(57)}$$

Similarly, the next bright fringe will occur at point B where

$$2nBB' \approx \left(m + \frac{3}{2}\right)\lambda \qquad \textbf{(58)}$$

Thus $$2n(BB' - AA') \approx \lambda$$

or $$XB' \approx \lambda/2n \qquad \textbf{(59)}$$

But $$XB' = A'X \tan\phi$$

or $$A'X = \beta \approx \frac{\lambda}{2n\phi} \qquad \textbf{(60)}$$

where β represents the fringe width and we have assumed ϕ to be small. Such fringes are commonly referred to as *fringes of equal thickness*.

On the other hand, for a point source, the fringe pattern will be similar to the parallel film case; i.e., for near-normal incidence, the pattern will be very nearly the same as that produced by two sources S' and S'' (Fig. 15.24). (Notice that

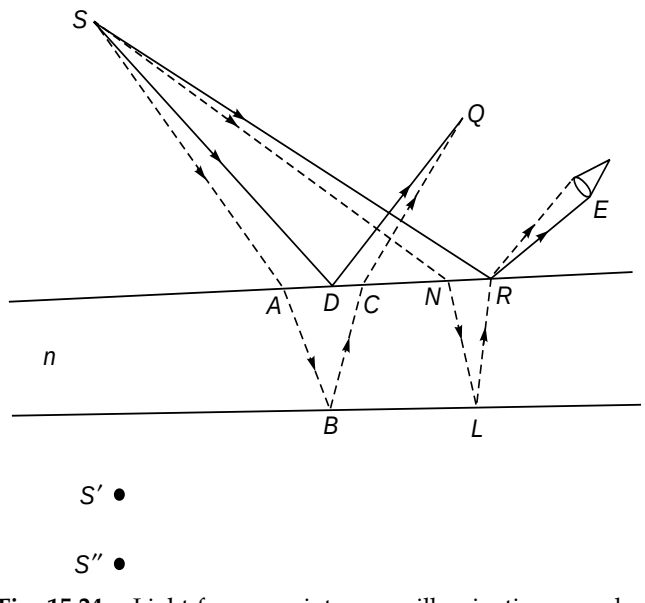


Fig. 15.24 Light from a point source illuminating a wedge. E represents the lens of the eye.

[6] We are assuming here that the beam undergoes a sudden phase change of π when it gets reflected by the upper surface. The expression for the fringe width [Eq. (60)] is, however, independent of this condition.

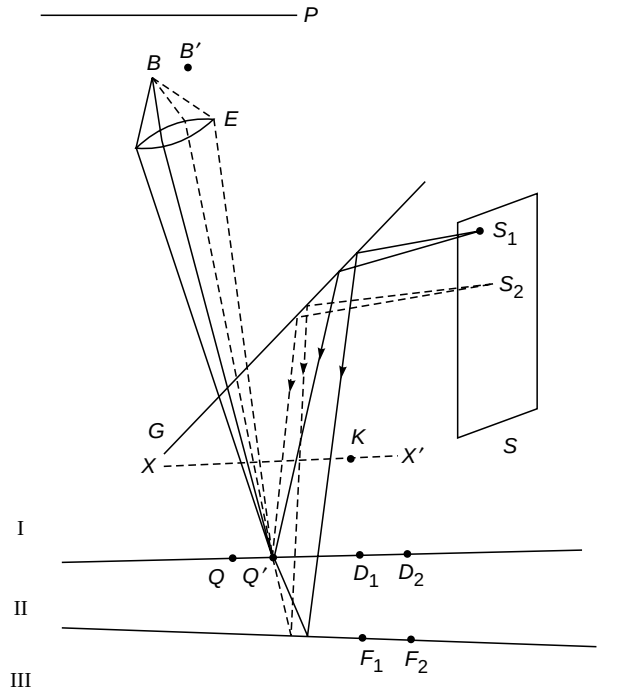


Fig. 15.25 Localized interference fringes produced by an extended source *S*. Fringes will be seen only when the eye is focused on the upper surface of the film.

point S'' is not vertically below S'; this is a consequence of the fact that the two surfaces of the film are not parallel.) The intensity of an arbitrary point Q will be determined by the following equations:

$$[SA + n(AB + BC) + CQ] - (SD + DQ) = \begin{cases} \left(m + \frac{1}{2}\right)\lambda & \text{maxima} \\ m\lambda & \text{minima} \end{cases} \quad \textbf{(61)}$$

If we view the film with the naked eye (say, at position *E*—see Fig. 15.24), then only a small portion of the film (around the point *R*) will be visible and the point *R* will be bright or dark as the optical path difference $[SN + n(NL + LR)] - SR$ is $\left(m + \frac{1}{2}\right)\lambda$ or $m\lambda$, respectively. One can similarly discuss the case when the eye is focused for infinity.

We next consider the illumination by an extended source *S* as shown in Fig. 15.25. Since the extended source can be assumed to consist of a large number of independent point sources, each point source will produce its own pattern on a photographic plate *P*; consequently, no definite fringe pattern will be observed.[7] However, if we view the film with a camera (or with the naked eye) and if the camera is focused on the upper surface of the film, then a particular point on the film will appear dark or bright depending on whether $2nd$ is $m\lambda$ or $\left(m + \frac{1}{2}\right)\lambda$ (see Fig. 15.25)—we are assuming near-normal incidence. The interference at point *Q* may occur due to light coming from different points on the extended source; but if the incidence is near normal, then the intensity at point *Q* will be determined entirely by the thickness of the film there. Similarly, the intensity at point Q' will be determined by the thickness of the film at Q'; however, point Q' will be focused at a different point B' on the retina of the eye. The fringes formed by a wedge will be straight lines parallel to the edge of the film OO' (Fig. 15.26). It should be emphasized that all along we are assuming near-normal incidence and that the wedge angle is extremely small. These assumptions are indeed valid for practical systems.

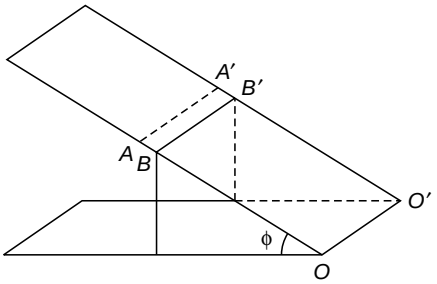


Fig. 15.26 The fringes formed by a wedge will be parallel to the edge OO'.

Note that if we focus the camera on a plane XX', which is slightly above the film, then no definite interference pattern will be observed. This follows from the fact that the light waves reaching the point *K* from S_2 undergo reflection at points D_2 and F_2, and the light waves reaching *K* from S_1 undergo reflection at points D_1 and F_1. Since the thickness of the film is not uniform, the waves reaching *K* from S_1 may produce brightness, whereas the waves reaching from S_2 may produce darkness. Thus, to view the fringes, one must focus the camera on the upper surface of the film, and in this sense, the fringes are said to be localized. It is left as an exercise for the reader to verify that if the camera is focused for infinity, no definite interference pattern will be recorded.

Until now we have assumed the film to be "thin"; the question now arises as to how thin the film should be. To obtain an interference pattern, there should be a definite phase relationship

[7] There is, however, one exception to this. When the extended source is taken to a very large distance, then the light rays reaching the plate *G* will be approximately parallel and an interference pattern (of low contrast) will be formed on plate *P*. The same phenomenon will also occur if, instead of moving the extended source, we take the plate *P* far away from the wedge.

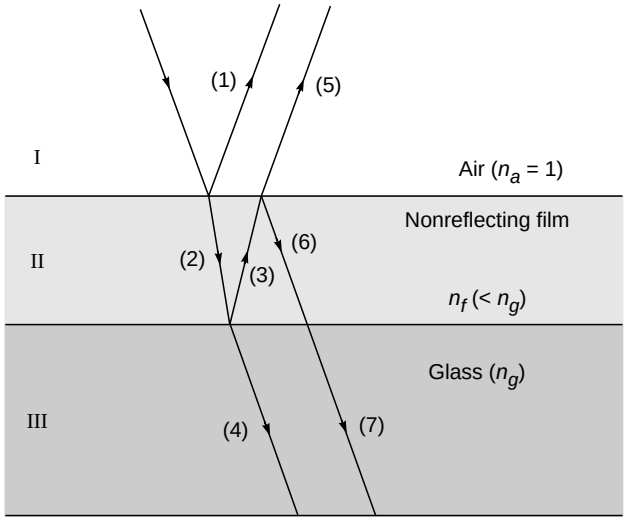


Fig. 15.27 In general, whereas the amplitudes of rays 1 and 5 are nearly the same, the amplitudes of 2 and 6 are quite different.

between the waves reflected from the upper surface of the film and from the lower surface of the film. Thus the path difference Δ (= $2nd\cos\theta'$) should be small compared to the coherence length.[8] For example, if we are using the D_1 line of an ordinary sodium lamp ($\lambda = 5.890 \times 10^{-5}$ cm), the coherence length is of the order of 1 cm, and for fringes to be visible, Δ should be much less than 1 mm. There is no particular value of Δ for which the fringes disappear; but as the value of Δ increases, the contrast of the fringes becomes poorer. A laser beam has a very high coherence length, and fringes can be visible even for path differences much greater than 1 m. On the other hand, if we use a white light source, no fringes will be visible for $\Delta \gtrsim 2 \times 10^{-4}$ cm (see Sec. 14.9).

Interference also occurs in region III (see Fig. 15.27) between the directly transmitted beam and the beam which comes out of the film after suffering two reflections, first from the lower surface and then from the upper surface of the film. However, the two amplitudes will be very different, and the fringes will have very poor contrast (see Example 15.1).

Example 15.1 Consider a film of refractive index 1.36 in air. Assuming near-normal incidence ($\theta \approx 0$), show that whereas the amplitudes of the reflected rays 1 and 5 (Fig. 15.27) are nearly equal, the amplitudes of the transmitted rays 4 and 7 are quite different. (This is the reason why the fringes observed in transmission have very poor contrast.)

Solution: Let the amplitude of the incident ray be a, and let the amplitudes of rays 1, 2, 3, ... be denoted by a_1, a_2, ... etc. Using Eqs. (10a) and (10b), we get

$$a_1 = \frac{1-n}{1+n}a = -\frac{0.36}{2.36}a \approx -0.153a$$

$$a_2 = \frac{2}{1+n}a = \frac{2}{2.36}a \approx 0.847a$$

$$a_3 = \frac{n-1}{n+1}a_2 = \frac{0.36}{2.36} \times 0.874a \approx 0.129a$$

$$a_5 = \frac{2n}{n+1}a_3 = \frac{2 \times 1.36}{2.36} \times 0.129a \approx 0.149a$$

$$a_4 = \frac{2n}{1+n}a_2 = \frac{2 \times 1.36}{2.36} \times 0.847a \approx 0.977a$$

$$a_7 = \frac{2n}{n+1}a_6 = \frac{2n}{n+1} \cdot \frac{n-1}{n+1}a_3 = \frac{2 \times 1.36 \times .36}{(2.36)^2}a_3$$

$$\approx 0.023a$$

We first note that the sign of a_5 is opposite to that of a_1 which is a consequence of the fact that a sudden phase change of π occurs when the ray gets reflected at point B. Further the magnitude of a_5 is nearly equal to that of a_1. On the other hand, $|a_7| << |a_4|$. This is the reason why the interference fringes formed in transmission have poor contrast.

15.9 COLORS OF THIN FILMS

We saw in Sec. 15.8 that if light from an extended monochromatic source (such as a sodium lamp) is incident normally on a wedge, then equally spaced dark and bright fringes will be observed. The distance between two consecutive bright (or dark) fringes is determined by the wedge angle, the wavelength of light, and the refractive index of the film. If we use a polychromatic source (such as an incandescent lamp), we will observe colored fringes. Further, if instead of a wedge we have a film of arbitrarily varying thickness, we will again observe fringes, each fringe representing the locus of constant film thickness (see Fig. 15.28). This is indeed what we see when sunlight falls on a soap bubble or on a thin film of oil on water. If the optical path difference between the waves reflected from the upper surface of the film and from the lower surface of the film exceeds a few wavelengths, the interference pattern will be washed out due to the overlapping of interference patterns of many colors and no fringes will be seen (see

[8] Coherence length is defined in Sec. 17.1. If a source remains coherent for a time τ, then the coherence length L will be about $c\tau_c$, where c is the speed of light in free space. Thus for $\tau_c \sim 10^{-10}$ s, $L \sim 3$ cm.

Fig. 15.28 A typical fringe pattern produced by an air film formed between two glass surfaces (which are not optically flat) and placed in contact with each other. Whenever the thickness of the air film is $m\lambda/2$, we obtain a dark fringe; and when the thickness is $\left(m+\frac{1}{2}\right)\lambda/2$, we obtain a bright fringe. Each fringe describes a focus of equal thickness of the film. (Photograph courtesy Prof. R. S. Sirohi.)

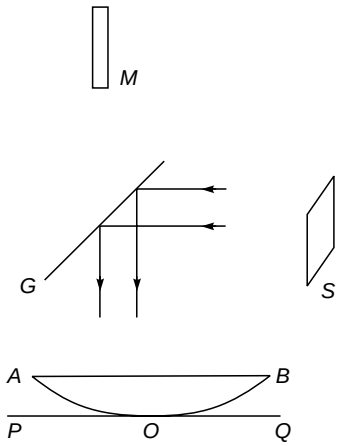


Fig. 15.29 An arrangement for observing Newton's rings. Light from an extended source *S* is allowed to fall on a thin film formed between the plano-convex lens *AOB* and the plane glass plate *POQ*. *M* represents a traveling microscope.

Sec. 14.9). Thus, to see the fringes with white light, the film should not be more than few wavelengths thick.

15.10 NEWTON'S RINGS

If we place a planoconvex lens on a plane glass surface, a thin film of air is formed between the curved surface of the lens (*AOB*) and the plane glass plate (*POQ*)—see Fig. 15.29. The thickness of the air film is zero at the point of contact *O* and increases as one moves away from the point of contact. If we allow monochromatic light (such as from a sodium lamp) to fall on the surface of the lens, then the light reflected from the surface *AOB* interferes with the light reflected from the surface *POQ*. For near-normal incidence (and considering points very close to the point of contact) the optical path difference between the two waves is very nearly equal to $2nt$, where n is the refractive index of the film and t is the thickness of the film. Thus, whenever the thickness of the air film satisfies the condition

$$2nt = \left(m+\tfrac{1}{2}\right)\lambda \qquad m = 0, 1, 2, \ldots \qquad \textbf{(62)}$$

we will have maxima. Similarly the condition

$$2nt = m\lambda \qquad \textbf{(63)}$$

will correspond to minima. Since the convex side of the lens is a spherical surface, the thickness of the air film will be constant over a circle (whose center will be at *O*), and we will obtain concentric dark and bright rings. These rings are known as Newton's rings.[9] Note that to observe the fringes, the microscope (or the eye) has to be focused on the upper surface of the film (see the discussion in Sec. 15.7).

The radii of various rings can be easily calculated. As mentioned earlier, the thickness of the air film will be constant over a circle whose center is at the point of contact *O*. Let the radius of the *m*th dark ring be r_m, and if t is the thickness of the air film where the *m*th dark ring appears to be formed, then

$$r_m^2 = t(2R - t) \qquad \textbf{(64)}$$

where R represents the radius of curvature of the convex surface of the lens (see Fig. 15.30). Now $R \approx 100$ cm and

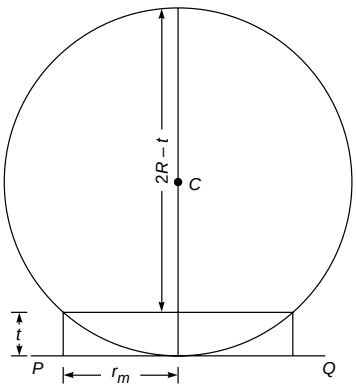


Fig. 15.30 r_m represents the radius of the *m*th dark ring; the thickness of the air film (where the *m*th dark ring is formed) is t.

[9] Boyle and Hooke had independently observed the fringes earlier, but Newton was the first to measure their radii and make an analysis. The proper explanation was given by Thomas Young. Also see Milestones in the beginning of this chapter.

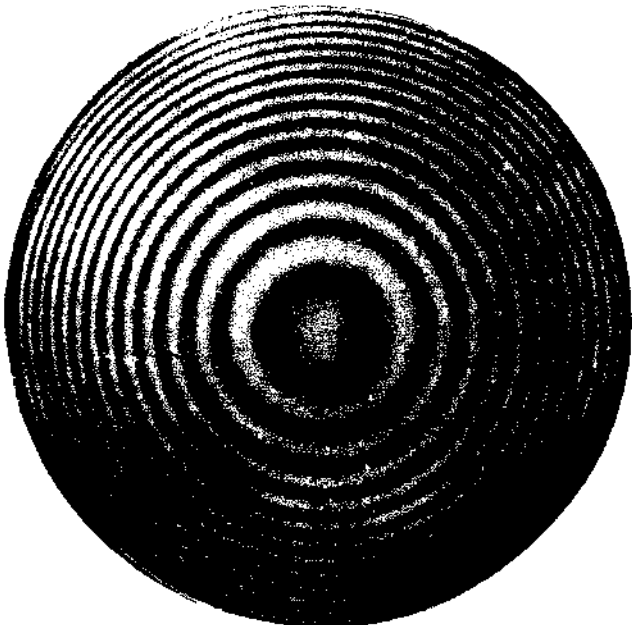


Fig. 15.31 Newton's rings as observed in reflection. The rings observed with transmitted light are of much poorer contrast. (Photograph courtesy Dr. G. Bose.)

$t \lesssim 10^{-3}$ cm. Thus we may neglect t in comparison to $2R$ to obtain

$$r_m^2 \approx 2Rt$$

or
$$2t \approx \frac{r_m^2}{R} \qquad \textbf{(65)}$$

Substituting this in Eq. (63), we get

$$r_m^2 \approx m\lambda R\,; \qquad m = 0, 1, 2, \ldots \qquad \textbf{(66)}$$

which implies that the radii of the rings vary as the square root of natural numbers. Thus the rings will become close to each other as the radius increases (see Fig. 15.31). Between the two dark rings there will be a bright ring whose radius will be $\sqrt{m+\frac{1}{2}}\,\lambda R$.

Newton's rings can be easily observed in the laboratory by using an apparatus as shown in Fig. 15.29. Light from an extended source (emitting almost monochromatic light, such as a sodium lamp) is allowed to fall on a glass plate which partially reflects the beam. This reflected beam falls on the planoconvex lens–glass plate arrangement, and Newton's rings can be easily observed by viewing directly or through a traveling microscope M. Actually, one really need not have a planoconvex lens; the rings will be visible even if a biconvex lens is used.

Typically for $\lambda = 6 \times 10^{-5}$ cm and $R = 100$ cm

$$r_m = 0.0774\sqrt{m} \quad \text{cm} \qquad \textbf{(67)}$$

Thus the radii of the first, second, and third dark rings are approximately 0.0774, 0.110, and 0.134 cm, respectively. Notice that the spacing between the second and third dark rings is smaller than the spacing between the first and second dark rings.

Equation (63) predicts that the central spot should be dark. Normally, with the presence of minute dust particles the point of contact is really not perfect, and the central spot may not be perfectly dark. Thus while carrying out the experiment, one should measure the radii of the mth and the $(m+p)$th ring ($p \approx 10$) and take the difference in the squares of the radii ($r_{m+p}^2 - r_m^2 = p\lambda R$), which is indeed independent of m. Usually, the diameter can be more accurately measured, and in terms of the diameters the wavelength is given by

$$\lambda = \frac{D_{m+p}^2 - D_m^2}{4pR} \qquad \textbf{(68)}$$

The radius of curvature can be accurately measured with the help of a spherometer, and therefore by carefully measuring the diameters of dark (or bright) rings one can experimentally determine the wavelength.

If a liquid of refractive index n is introduced between the lens and the glass plate, the radii of the dark rings are given by

$$r_m = \sqrt{\frac{m\lambda R}{n}} \qquad \textbf{(69)}$$

Equation (69) may be compared with Eq. (66). Further, if the refractive indices of the material of the lens and of the glass plate are different and if the refractive index of the liquid lies in between the two values, the central spot will be bright as in Fig. 15.31 and Eq. (69) gives the radii of the bright rings.

An important practical application of the principle involved in the Newton rings experiment lies in the determination of the optical flatness of a glass plate. Consider a glass surface placed on another surface whose flatness is known. If a monochromatic light beam is allowed to fall on this combination and the reflected light is viewed by a microscope, then, in general, dark and bright patches will be seen (Fig. 15.28). The space between the two glass surfaces forms an air film of varying thickness, and whenever this thickness becomes $m\lambda/2$, we see a dark spot; and when this thickness becomes $\left(m+\frac{1}{2}\right)\lambda/2$, we see a bright spot. Two consecutive dark fringes will be separated by the air film whose thickness will differ by $\lambda/2$. Consequently, by measuring the distance between consecutive dark and bright fringes, one can calculate the optical flatness of a glass plate.

When we observe Newton's rings by using a white light source, we will have a situation similar to that discussed in Sec. 14.9; i.e., we will see only a few colored fringes. However, if we put a red filter in front of the naked eye, the fringe pattern (corresponding to the red color) will suddenly appear. If we

replace the red filter by a green filter in front of the eye, the fringe pattern corresponding to the green color will appear; this is similar to the discussion in Sec. 14.9.

Example 15.2 Consider the formation of Newton's rings by monochromatic light of $\lambda = 6.4 \times 10^{-5}$ cm. Assume the point of contact to be perfect. Now slowly raise the lens vertically above the plate. As the lens moves gradually away from the plate, discuss the ring pattern as seen through the microscope. Assume the radius of the convex surface to be 100 cm.

Solution: Since the point of contact is perfect, the central spot will be dark, the first dark ring will form at P where $PA = \lambda/2$, and the radius of this ring OA will be $\sqrt{\lambda R}$ (= 0.080 cm); see Fig. 15.32(a). Similarly, the radius of the second dark ring will be $OB = \sqrt{2\lambda R}$ (= 0.113 cm). If we now raise the lens by $\lambda/4$ (= 1.6×10^{-5} cm), then $2t$ corresponding to the central spot would be $\lambda/2$ and instead of the dark spot at the center we will now have a bright spot. The radii of the first and the second dark rings will be

$$OA_1 = \left(\frac{1}{2}\lambda R\right)^{1/2} = 0.0566 \text{ cm}$$

and

$$OB_1 = \left(\frac{3}{2}\lambda R\right)^{1/2} = 0.098 \text{ cm}$$

respectively [see Fig. 15.32(b)]. If the lens is further moved by $\lambda/4$ (see Fig. 15.32(c)], then the first dark ring collapses to the center and the central spot will be dark. The ring which was originally at Q now shifts to Q_2; similarly the ring at R [Fig. 15.32(a)] collapses to R_2 [Fig. 15.32(c)].

Thus, as the lens is moved upward, the rings collapse to the center. Hence if we can measure the distance by which the lens is moved upward and also count the number of dark spots that have collapsed to the center, we can determine the wavelength. For example, in the present case, if the lens is moved by 6.4×10^{-3} cm, 200 rings will collapse to the center. If one carries out this experiment, it will be observed that the 200th dark ring will slowly converge to the center, and when the lens has moved by exactly 6.4×10^{-3} cm, it has exactly come to the center.

Example 15.3 Consider the formation of Newton's rings when two closely spaced wavelengths are present; for example, the D_1 and D_2 lines of sodium (λ_1 = 5890 Å and λ_2 = 5896 Å). What will be the effect of the presence of these two wavelengths as the lens is gradually moved away from the plate? What will happen if the sodium lamp is replaced by a white light source?

Solution: We will first assume that the lens is in contact with the plane glass plate [see Fig. 15.32(a)]. Since the two wavelengths are very close, the bright and dark rings of λ_1 superpose on the bright and dark rings of λ_2, respectively. This can be easily seen by calculating the radii of the ninth dark and bright ring for each wavelength.

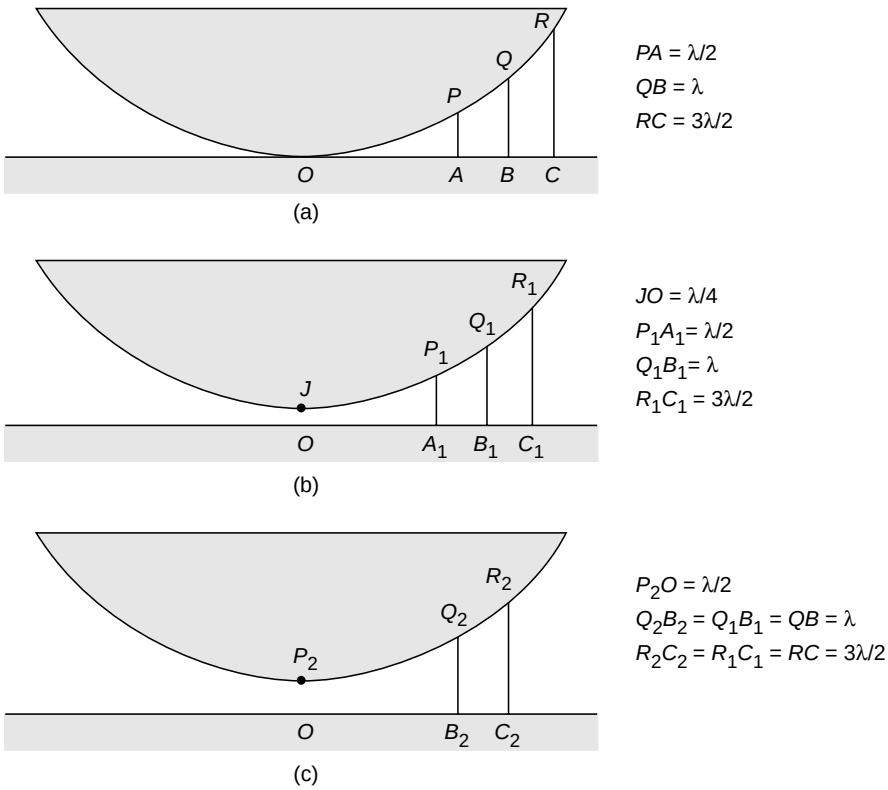


Fig. 15.32 The rings collapse to the center as the lens is moved away from the plate.

For $\lambda = 5.890 \times 10^{-5}$ cm,

$$\text{Radius of ninth bright ring} = \sqrt{\left(9+\tfrac{1}{2}\right)\lambda R}$$
$$= \sqrt{9.5 \times 5.890 \times 10^{-5} \times 100}$$
$$= 0.236548 \text{ cm}$$
$$\text{Radius of ninth dark ring} = \sqrt{9\lambda R}$$
$$= 0.230239 \text{ cm}$$

Similarly, for $\lambda = 5.896 \times 10^{-5}$ cm,

$$\text{Radius of ninth bright ring} = \sqrt{9.5 \times 5.896 \times 10^{-5} \times 100}$$
$$= 0.236669 \text{ cm}$$

and

$$\text{Radius of ninth dark ring} = \sqrt{9 \times 5.896 \times 10^{-3}}$$
$$= 0.230356 \text{ cm}$$

Thus the rings almost exactly superpose on each other. However, for large values of m, the two ring patterns may produce uniform illumination. To be more specific, when the air-film thickness t is such that

$$2t = m\lambda_1 = \left(m+\tfrac{1}{2}\right)\lambda_2$$

or
$$\frac{2t}{\lambda_2} - \frac{2t}{\lambda_1} = \frac{1}{2} \qquad \textbf{(70)}$$

then around that point the fringe system will completely disappear; i.e., the bright ring for wavelength λ_1 will fall on the dark ring for wavelength λ_2, and conversely. Thus the contrast will be zero, and no fringe pattern will be visible. Rewriting Eq. (70), we get

$$2t\frac{\lambda_1 - \lambda_2}{\lambda_1\lambda_2} = \frac{1}{2}$$

or
$$2t = \frac{1}{2}\frac{\lambda_1\lambda_2}{\Delta\lambda} \approx \frac{1}{2}\frac{(5.893\times10^{-5})^2}{6\times10^{-8}}$$
$$\approx 3 \times 10^{-2} \text{ cm}$$

This will correspond to $m \approx 500$.

We will see the effect of the same phenomenon if we slowly raise the convex lens in the upward direction as we had considered in Example 15.2. Let t_0 be the vertical distance through which the lens has been raised (see Fig. 15.33), and let t_0 be such that it satisfies the following equation:

$$\frac{2t_0}{\lambda_2} - \frac{2t_0}{\lambda_1} = \frac{1}{2}$$

or
$$t_0 = \frac{\lambda_1\lambda_2}{4(\lambda_1 - \lambda_2)}$$

Thus, if point J (see Fig. 15.33) corresponds to a dark spot for λ_1, then it will correspond to a bright spot for λ_2, and conversely. Further, the nearby dark rings for λ_1 will almost fall at the same place as the bright rings for λ_2, and the interference pattern will be washed out. Thus viewing from a microscope, we will not be able to see any ring pattern. Now, if the lens is further moved upward by a distance t_0, then we will have

$$\frac{2t_1}{\lambda_2} - \frac{2t_1}{\lambda_1} = 1 \qquad \textbf{(71)}$$

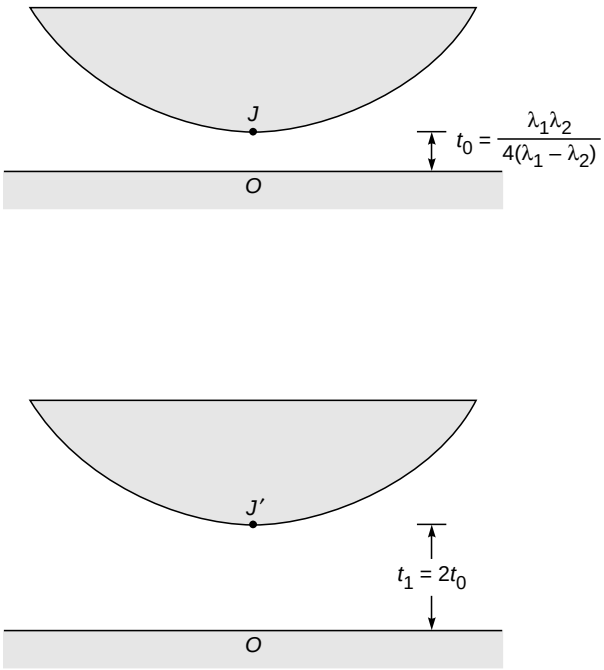


Fig. 15.33 In the Newton rings experiment, if the light consists of two closely spaced wavelengths λ_1 and λ_2 (such as the D_1 and D_2 lines of sodium), then if the lens is separated by a distance t_0 $\left(= \lambda_1\lambda_2 / 4(\lambda_1 - \lambda_2)\right)$, interference fringes will be washed out. The fringes will reappear when the distance is $2t_0$.

where $t_1 = 2t_0$. Consequently, if point J' corresponds to a dark spot for λ_1, then it will also correspond to a dark spot for λ_2. The fringe pattern will reappear but now with a slightly weaker contrast (see also Chap. 17).

In this way if we continue to move the lens upward, the fringe system will reappear every time the lens is moved up by a distance $2t_0$ $\left(\approx \tfrac{1}{2}\lambda_1\lambda_2 / \Delta\lambda\right)$. This principle is used in a Michelson interferometer to measure the small wavelength difference $\Delta\lambda$ between two closely spaced lines (such as the D_1 and D_2 lines of sodium).

For complete disappearance of the fringe pattern the intensities of the two lines λ_1 and λ_2 should be the same.

Another corollary of the above experiment consists in finding the change in the interference pattern (as we move up the convex lens) when we consider a single line of wavelength λ, but which has a width of $\Delta\lambda$. Thus we should assume all wavelengths between λ and $\lambda + \Delta\lambda$ to exist. By finding the approximate height at which the fringes disappear, we can calculate $\Delta\lambda$. The coherence length L is related to $\Delta\lambda$ through the following relation (see Sec. 17.2):

$$L \sim \frac{\lambda^2}{\Delta\lambda} \qquad \textbf{(72)}$$

15.11 THE MICHELSON INTERFEROMETER

A schematic diagram of the Michelson interferometer is shown in Fig. 15.34; S represents a light source (which may be a sodium lamp) and L represents a ground glass plate so that an extended source of almost uniform intensity is formed. G_1 is a beam splitter; i.e., a beam incident on G_1 gets partially reflected and partially transmitted. M_1 and M_2 are good-quality plane mirrors having very high reflectivity. One of the mirrors (usually M_2) is fixed and the other (usually M_1) is capable of moving away from or toward the glass plate G_1 along an accurately machined track by means of a screw. In the normal adjustment of the interferometer, mirrors M_1 and M_2 are perpendicular to each other and G_1 is at 45° to the mirror.

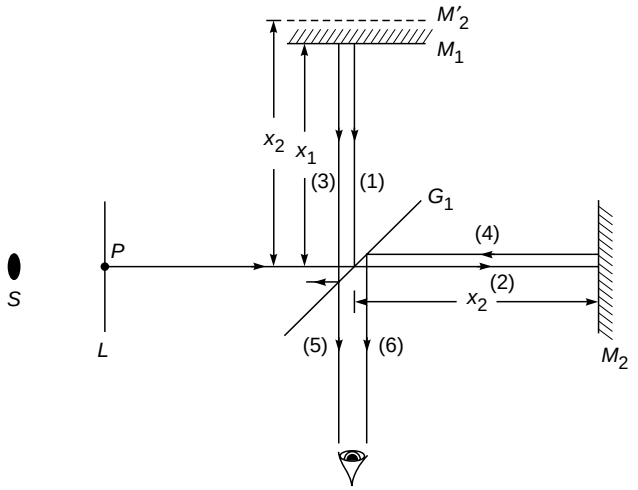


Fig. 15.34 Schematic of the Michelson interferometer.

Waves emanating from a point P get partially reflected and partially transmitted by the beam splitter G_1, and the two resulting beams are made to interfere in the following manner: The reflected wave (shown as 1 in Fig. 15.34) undergoes a further reflection at M_1, and this reflected wave gets (partially) transmitted through G_1; this is shown as 5 in the figure. The transmitted wave (shown as 2 in Fig. 15.34) gets reflected by M_2 and gets (partially) reflected by G_1 and results in the wave shown as 6 in the figure. Waves 5 and 6 interfere in a manner exactly similar to that shown in Fig. 15.22. This can be easily seen from the fact that if x_1 and x_2 are the distances of mirrors M_1 and M_2 from the plate G_1, then to the eye the waves emanating from point P will appear to get reflected by two parallel mirrors (M_1 and M_2' — see Fig. 15.34) separated by a distance $x_1 \sim x_2$. As discussed in Sec. 15.7, if we use an extended source, then no definite interference pattern will be obtained on a photographic plate placed at the position of the eye. Instead, if we have a camera focused for infinity, then on the focal plane we will obtain circular fringes, each circle corresponding to a definite value of θ (see Figs. 15.22 and 15.35); the circular fringes will look like the ones shown in Fig. 15.36. Now, if the beam splitter is just a simple glass plate, the beam reflected from mirror M_2 will undergo an abrupt phase change of π (when getting reflected by the beam splitter), and since the extra path that one of the beams will traverse will be $2(x_1 \sim x_2)$, the condition for destructive interference will be

$$2d \cos \theta = m\lambda$$

where $m = 0, 1, 2, 3, \ldots$ and

$$d = x_1 \sim x_2$$

and the angle θ represents the angle that the rays make with the axis (which is normal to the mirrors as shown in Fig. 15.35). Similarly, the condition for a bright ring is

$$2d \cos \theta = \left(m + \tfrac{1}{2}\right)\lambda$$

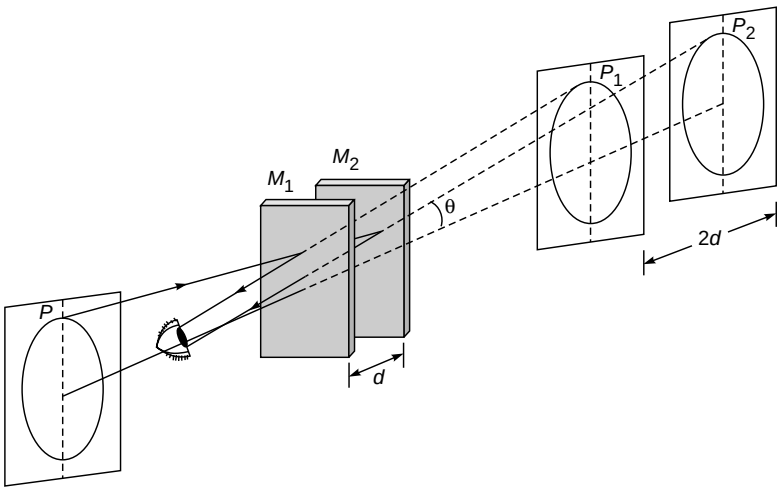


Fig. 15.35 A schematic of the formation of circular fringes (Adapted from Ref. 7).

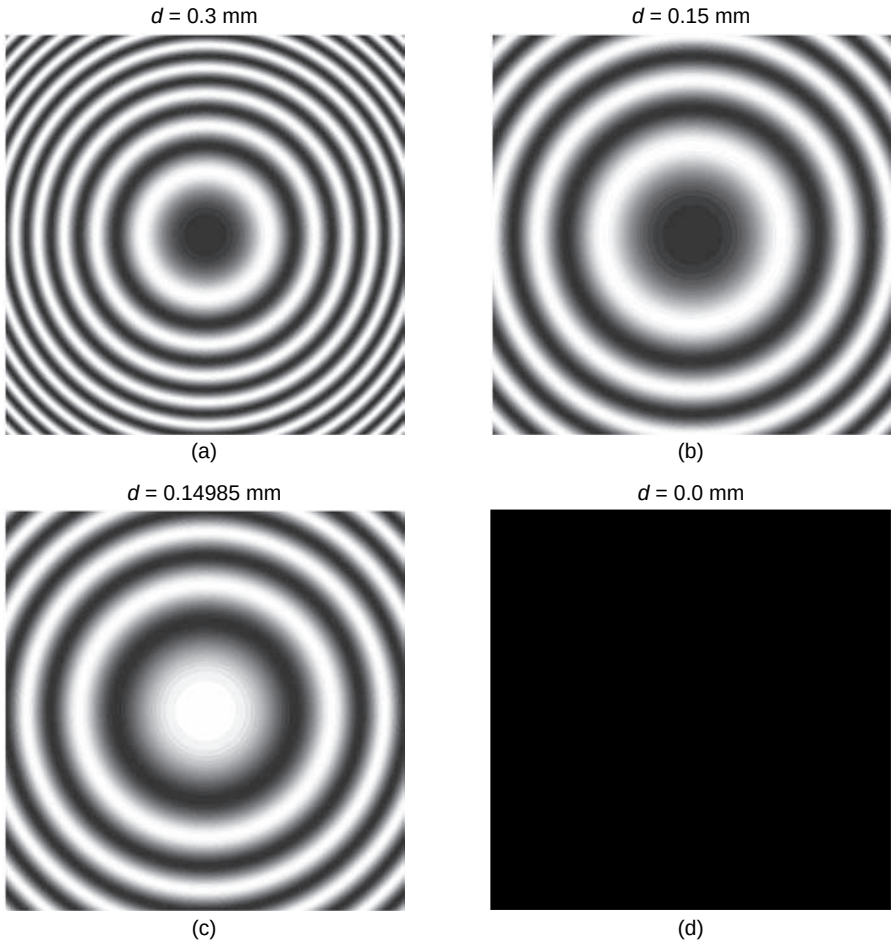


Fig. 15.36 Computer-generated interference pattern produced by a Michelson interferometer.

For example, for $\lambda = 6 \times 10^{-5}$ cm if $d = 0.3$ mm, the angles at which the dark rings will occur are

$$\theta = \cos^{-1}\frac{m}{500}$$

$$= 0°, 2.56°, 3.62°, 4.44°, 5.13°, 5.73°, 6.28°, \ldots$$

corresponding to $m = 1000, 999, 998, 997, 996, 995, \ldots$. Thus the central dark ring in Fig. 15.36(a) corresponds to $m = 1000$, the first dark ring corresponds to $m = 999$, etc. If we now reduce the separation between the two mirrors so that $d = 0.15$ mm, the angles at which the dark rings occur will be [see Fig.15.36(b)]

$$\theta = \cos^{-1}\frac{m}{500} = 0°, 3.62°, 5.13°, 6.28°, 7.25°, \ldots$$

where the angles now correspond to $m = 500, 499, 498, 497, 496, 495, \ldots$. Thus as we start reducing the value of d, the fringes will appear to collapse at the center and the fringes become less closely placed. Note that if d is now slightly decreased, say, from 0.15 to 0.14985 mm, then

$$2d = 499.5\lambda$$

the dark central spot in Fig.15.36(b) (corresponding to $m = 500$) would disappear and the central fringe will become bright. Thus, as d decreases, the fringe pattern tends to collapse toward the center. (Conversely, if d is increased, the fringe pattern will expand.) Indeed, if N fringes collapse to the center as mirror M_1 moves by a distance d_0, then we must have

$$2d = m\lambda$$

$$2(d - d_0) = (m - N)\lambda$$

where we have set $\theta' = 0$ because we are looking at the central fringe. Thus

$$\lambda = \frac{2d_0}{N} \qquad \textbf{(73)}$$

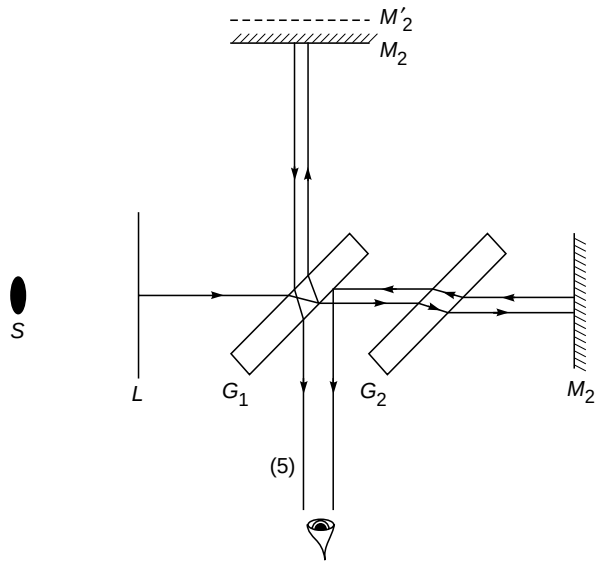


Fig. 15.37 In an actual interferometer there is also a compensating plate G_2.

This provides us with a method for the measurement of the wavelength. For example, in a typical experiment, if 1000 fringes collapse to the center as the mirror is moved through a distance of 2.90×10^{-2} cm, then

$$\lambda = 5800 \text{ Å}$$

The above method was used by Michelson for the standardization of the meter. He found that the red cadmium line ($\lambda = 6438.4696$ Å) is one of the ideal monochromatic sources, and as such this wavelength was used as a reference for the standardization of the meter. In fact he defined the meter by the following relation:

1 m = 1,553,164.13 red cadmium wavelengths

the accuracy is almost 1 part in 10^9.

In an actual Michelson interferometer, the beam splitter G_1 consists of a plate (which may be about $\frac{1}{2}$ cm thick), the back surface of which is partially silvered, and the reflections occur at the back surface as shown in Fig. 15.37. It is immediately obvious that beam 5 traverses the glass plate three times, and to compensate for this additional path, one introduces a "compensating plate" G_2 which is exactly of the same thickness as G_1. The compensating plate is not really necessary for a monochromatic source because the additional path $2(n-1)t$ introduced by G_1 can be compensated by moving mirror M_1 by a distance $(n-1)t$, where n is the refractive index of the material of the glass plate G_1.

However, for a white light source it is not possible to simultaneously satisfy the zero path-difference condition for all wavelengths, since the refractive index depends on wavelength. For example, for $\lambda = 6560$ and 4861 Å, the refractive index of crown glass is 1.5244 and 1.5330, respectively. If we are using a 0.5 cm thick crown glass plate as G_1, then M_1 should be moved by 0.2622 cm for $\lambda = 6560$ Å and by 0.2665 cm for $\lambda = 4861$ Å, the difference between the two positions corresponding to over 100 wavelengths! Thus, if we have a continuous range of wavelengths from 4861 to 6560 Å, the path difference between any pair of interfering rays (see Fig. 15.34) will vary so rapidly with wavelength that we would observe only a uniform white light illumination. However, in the presence of the compensating plate G_2, we would observe a few colored fringes around the point corresponding to zero path difference (see Sec. 14.9).

The Michelson interferometer can also be used in the measurement of two closely spaced wavelengths. Let us assume that we have a sodium lamp which emits predominantly two closely spaced wavelengths 5890 and 5896 Å. The interferometer is first set corresponding to the zero path difference.[10] Near $d = 0$, both the fringe patterns will overlap. If mirror M_1 is moved away from (or toward) plate G_1 through a distance d, then the maxima corresponding to the wavelength λ_1 will not, in general, occur at the same angle as λ_2. Indeed, if the distance d is such that

$$\frac{2d}{\lambda_1} - \frac{2d}{\lambda_2} = \frac{1}{2} \qquad \textbf{(74)}$$

and if $2d \cos\theta' = m\lambda_1$, then $2d \cos\theta' = \left(m + \frac{1}{2}\right)\lambda_2$. Thus, the maxima of λ_1 will fall on the minima of λ_2, and conversely, and the fringe system will disappear. It can be easily seen that if

$$\frac{2d}{\lambda_1} - \frac{2d}{\lambda_2} = 1 \qquad \textbf{(75)}$$

then interference pattern will again reappear. In general, if

$$\frac{2d}{\lambda_1} - \frac{2d}{\lambda_2}$$

is 1/2, 3/2, 5/2, . . . , we will have disappearance of the fringe pattern; and if it is equal to 1, 2, 3, . . . , then the interference pattern will appear.

Instead of two discrete wavelengths, if the source consists of all wavelengths lying between λ and $\lambda + \Delta\lambda$, then no interference pattern will be observed if

$$\frac{2d}{\lambda} - \frac{2d}{\lambda + \Delta\lambda/2} \gtrsim \frac{1}{2}$$

or

$$2d \gtrsim \frac{\lambda^2}{\Delta\lambda} \qquad \textbf{(76)}$$

[10] The zero path difference is easily obtained by using white light where only a few colored fringes, around $d = 0$, will be visible.

In this case the fringes will not reappear because we have a continuous range of wavelengths rather than two discrete wavelengths (see Sec. 17.2).

Michelson and Morley carried out the famous Michelson–Morley experiments using the Michelson interferometer to detect the motion of the earth with respect to the "ether"—this experiment is discussed in Sec. 30.10.

Example 15.4 For a sodium lamp, the distance traversed by the mirror between two successive disappearances is 0.289 mm. Calculate the difference in the wavelengths of the D_1 and D_2 lines. Assume λ = 5890 Å.

Solution: When the mirror moves through a distance 0.289 mm, the additional path introduced is 0.578 mm. Thus

$$\frac{0.578}{\lambda} - \frac{0.578}{\lambda + \Delta\lambda} = 1$$

or
$$\Delta\lambda \simeq \frac{\lambda^2}{0.578} = \frac{(5890 \times 10^{-7})^2}{0.578} \text{ mm}$$
$$\simeq 6 \text{ Å}$$

Summary

- If a plane wave is incident normally on a thin film of uniform thickness d, then the waves reflected from the upper surface interfere with the waves reflected from the lower surface. Indeed, for a film of thickness $\lambda/4n_f$ (where λ is the free space wavelength and n_f is the film refractive index which lies between the refractive indices of the two surrounding media), the wave reflected from the upper surface interferes destructively with the wave reflected from the lower surface, and therefore the film acts as an antireflection layer.
- A medium consisting of a large number of alternate layers of high and low refractive indices of $n_0 + \Delta n$ and $n_0 - \Delta n$ of equal thickness d is called a periodic medium, and the spatial period of the refractive index variation is denoted by Λ (= $2d$). For $\Delta n << n_0$, if $d \approx \lambda/4n_0$ (where λ is the free space wavelength), the reflections arising out of the individual reflections from the various interfaces will all be in phase and will result in a strong reflection. Thus for strong reflection at a chosen (free space) wavelength λ_B, the period of the refractive index variation should be
$$\Lambda = 2d = \frac{\lambda_B}{2n_0}$$
This is referred to as the Bragg condition. This is the principle of operation of fiber Bragg gratings.
- If we place a planoconvex lens on a plane glass surface, a thin film of air is formed between the curved surface of the lens and the plane glass plate. If we allow monochromatic light (such as from a sodium lamp) to fall (almost normally) on the surface of the lens, then the light reflected from the curved surface interferes with the light reflected from the plane surface. Since the convex side of the lens is a spherical surface, the thickness of the air film will be constant over a circle and we will see concentric dark and bright rings. These rings are known as Newton's rings. The radii of the concentric rings are such that the difference between the square of the radii of successive fringes is very nearly a constant.
- The Michelson interferometer was used by Michelson for the standardization of the meter. He found that the red cadmium line (λ = 6438.4696 Å) is one of the ideal monochromatic sources, and as such this wavelength was used as a reference for the standardization of the meter. In fact he defined the meter as
$$1 \text{ m} = 1{,}553{,}164.13 \text{ red cadmium wavelengths}$$
the accuracy is almost 1 part in 10^9.
- The Michelson interferometer can also be used in the measurement of two closely spaced wavelengths.

Problems

15.1 A glass plate of refractive index 1.6 is in contact with another glass plate of refractive index 1.8 along a line such that a wedge of 0.5° is formed. Light of wavelength 5000 Å is incident vertically on the wedge, and the film is viewed from the top. Calculate the fringe spacing. The whole apparatus is immersed in an oil of refractive index 1.7. What will be the qualitative difference in the fringe pattern, and what will be the new fringe width?

15.2 Two plane glass plates are placed on top of each other, and on one side a cardboard is introduced to form a thin wedge of air. Assuming that a beam of wavelength 6000 Å is incident normally, and that there are 100 interference fringes per centimeter, calculate the wedge angle.

15.3 Consider a nonreflecting film of refractive index 1.38. Assume that its thickness is 9×10^{-6} cm. Calculate the wavelengths (in the visible region) for which the film will be nonreflecting. Repeat the calculations for the thickness of the film to be 45×10^{-6} cm. Show that both films will be nonreflecting for a particular wavelength, but only the former one will be suitable. Why?

15.4 In the Newton rings arrangement, the radius of curvature of the curved side of the planoconvex lens is 100 cm. For $\lambda = 6 \times 10^{-5}$ cm, what will be the radii of the 9th and 10th bright rings?

15.5 In the Newton rings arrangement, the radius of curvature of the curved surface is 50 cm. The radii of the 9th and 16th dark rings are 0.18 and 0.2235 cm, respectively. Calculate the wavelength.

[*Hint*: The use of Eq. (66) will give wrong results. Why?]

[Ans: 5015 Å]

15.6 In the Newton rings arrangement, if the incident light consists of two wavelengths of 4000 and 4002 Å, calculate the

distance (from the point of contact) at which the rings will disappear. Assume that the radius of curvature of the curved surface is 400 cm.

[**Ans:** 4 cm]

15.7 In Prob. 15.6 if the lens is slowly moved upward, calculate the height of the lens at which the fringe system (around the center) will disappear.

[**Ans:** 0.2 mm]

15.8 An equiconvex lens is placed on another equiconvex lens. The radii of curvature of the two surfaces of the upper lens are 50 cm, and those of the lower lens are 100 cm. The waves reflected from the upper and lower surfaces of the air film (formed between the two lenses) interfere to produce Newton's rings. Calculate the radii of the dark rings. Assume $\lambda = 6000$ Å.

[**Ans:** $0.0447\sqrt{m}$ cm]

15.9 In the Michelson interferometer arrangement, if one of the mirrors is moved by a distance 0.08 mm, 250 fringes cross the field of view. Calculate the wavelength.

[**Ans:** 6400 Å]

15.10 The Michelson interferometer experiment is performed with a source which consists of two wavelengths of 4882 and 4886 Å. Through what distance does the mirror have to be moved between two positions of the disappearance of the fringes?

[**Ans:** 0.298 mm]

15.11 In the Michelson interferometer experiment, calculate the various values of θ' (corresponding to bright rings) for $d = 5 \times 10^{-3}$ cm. Show that if d is decreased to 4.997×10^{-3} cm, the fringe corresponding to $m = 200$ disappears. What will be the corresponding values of θ'? Assume $\lambda = 5 \times 10^{-5}$ cm.

REFERENCES AND SUGGESTED READINGS

1. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, 1975.
2. M. Cagnet, M. Francon, and S. Mallick, *Atlas of Optical Phenomena*, Springer-Verlag, Berlin, 1971.
3. E. F. Cave and L. V. Holroyd, "Inexpensive Michelson Interferometer," *American Journal Physics*, Vol 23, p. 61, 1955.
4. A. H. Cook, *Interference of Electromagnetic Waves*, Clarendon Press, Oxford, 1971.
5. M. Francon, *Optical Interferometry*, Academic Press, New York, 1966.
6. A. K. Ghatak and K. Thyagarajan, *Optical Electronics*, Cambridge University Press, London, 1989. Reprinted by Foundation Books, New Delhi.
7. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, McGraw-Hill Book Co., New York, 1976.
8. V. Oppenheim and J. H. Jaffe, "Interference in an Optical Wedge," *American Journal Physics*, Vol. 24, p. 610, 1956.
9. J. Sladkova, *Interference of Light*, Iliffe Books Ltd., London, 1968.
10. W. H. Steel, *Interferometry*, Cambridge University Press, London, 1967.
11. S. Tolansky, *An Introduction to Interferometry*, Longmans Green and Co., London, 1955.

Chapter Sixteen

MULTIPLE-BEAM INTERFEROMETRY

When two Undulations . . . coincide either perfectly or very nearly in Direction, their joint effect is a Combination of the Motions belonging to each.[1]

—Thomas Young (1801)

Important Milestone

1899 *Marie Fabry and Jean Perot invented the Fabry–Perot interferometer which is characterized by a very high resolving power.*

16.1 INTRODUCTION

In the last two chapters, we have been discussing interference between two beams which are derived from a single beam either by division of wave front or by division of amplitude. In this chapter, we will discuss interference involving many beams which are derived from a single beam by multiple reflections (division of amplitude). Thus, for example, if a plane wave falls on a plane parallel glass plate, then the beam undergoes multiple reflections at the two surfaces and a large number of beams of successively diminishing amplitude emerge on both sides of the plate. These beams (on either side) interfere to produce an interference pattern at infinity. We will show that the fringes so formed are much sharper than those by two-beam interference and, therefore, the interferometers involving multiple-beam interference have a high resolving power and hence find applications in high-resolution spectroscopy.

16.2 MULTIPLE REFLECTIONS FROM A PLANE PARALLEL FILM

We consider the incidence of a plane wave on a plate of thickness h (and of refractive index n_2) surrounded by a medium of refractive index n_1 as shown in Fig. 16.1; as we will discuss later, the Fabry–Perot interferometer consists of two partially reflecting mirrors (separated by a fixed distance h) placed in air so that $n_1 = n_2 = 1$.

Let A_0 be the (complex) amplitude of the incident wave. The wave will undergo multiple reflections at the two interfaces as shown in Fig. 16.1(a). Let r_1 and t_1 represent the amplitude reflection and transmission coefficients, respectively, when the wave is incident from n_1 toward n_2, and let r_2 and t_2 represent the corresponding coefficients when the wave is incident from n_2 toward n_1. Thus the amplitude of the successive reflected waves will be

$$A_0 r_1,\ A_0 t_1 r_2 t_2 e^{i\delta},\ A_0 t_1 r_2^3 e^{2i\delta},\ \ldots$$

where
$$\delta = \frac{2\pi}{\lambda_0}\Delta = \frac{4\pi n_2 h \cos\theta_2}{\lambda_0} \qquad \textbf{(1)}$$

represents the phase difference (between two successive waves emanating from the plate) due to the additional path traversed by the beam in the film (see Sec. 15.1) and in Eq. (1), θ_2 is the angle of refraction inside the film (of refractive index n_2), h the film thickness, and λ_0 is the free space wavelength. Thus the resultant (complex) amplitude of the reflected wave will be

$$\begin{aligned} A_r &= A_0[r_1 + t_1 t_2 r_2 e^{i\delta}(1 + r_2^2 e^{i\delta} + r_2^4 e^{2i\delta} + \cdots)] \\ &= A_0\left(r_1 + \frac{t_1 t_2 r_2 e^{i\delta}}{1 - r_2^2 e^{i\delta}}\right) \end{aligned} \qquad \textbf{(2)}$$

[1] The author found this quotation in Ref. 1.

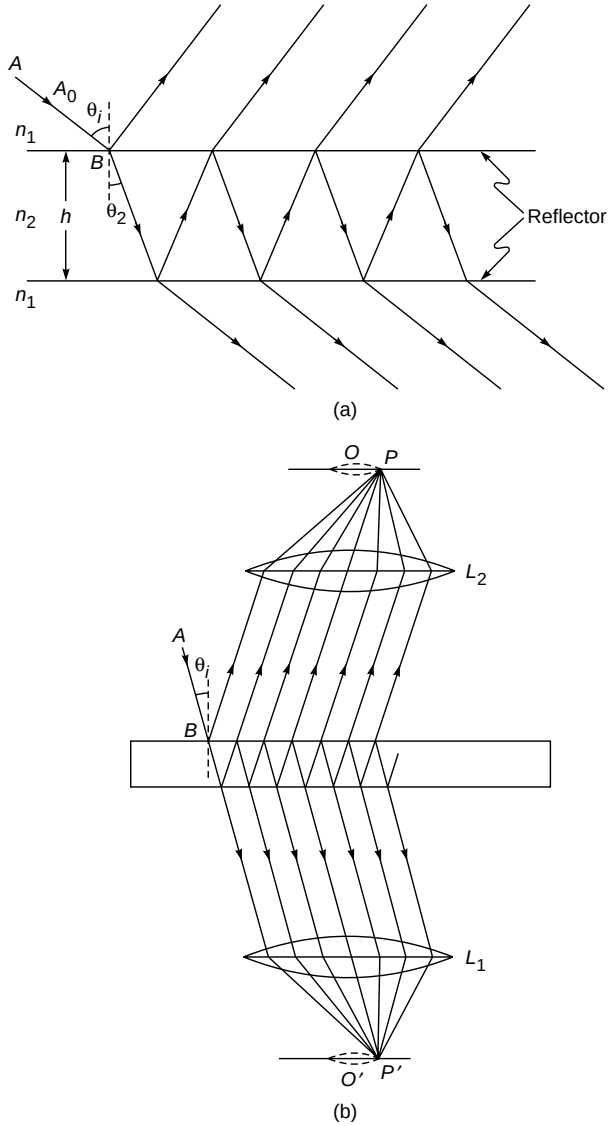


Fig. 16.1 (a) Reflection and transmission of a beam of amplitude A_0 incident at an angle θ_i on a film of refractive index n_2 and thickness h. (b) Any ray parallel to AB will focus at the same point P. If ray AB is rotated about the normal at B, then point P will rotate on the circumference of a circle centered at point O; this circle will be bright or dark depending on the value of θ_i. Rays incident at different angles will focus at different distances from point O, and one will obtain concentric bright and dark rings for an extended source.

Now, if the reflectors are lossless, the reflectivity and the transmittivity at each interface are given by (see Sec. 14.12)

$$R = r_1^2 = r_2^2$$
$$\tau = t_1 t_2 = 1 - R$$

(We are reserving the symbol T for the transmittivity of the Fabry–Perot etalon.) Thus

$$\frac{A_r}{A_0} = r_1\left[1 - \frac{(1-R)e^{i\delta}}{1 - Re^{i\delta}}\right]$$

where we have used the fact that $r_2 = -r_1$. Thus the reflectivity of the Fabry–Perot etalon is given by

$$\mathcal{R} = \left|\frac{A_r}{A_0}\right|^2 = R\left|\frac{1-e^{i\delta}}{1-Re^{i\delta}}\right|^2$$

$$= R\,\frac{(1-\cos\delta)^2 + \sin^2\delta}{(1-R\cos\delta)^2 + R^2\sin^2\delta}$$

$$= \frac{4R\sin^2\delta/2}{(1-R)^2 + 4R\sin^2\delta/2}$$

or
$$\mathcal{R} = \frac{F\sin^2\delta/2}{1 + F\sin^2\delta/2} \qquad \textbf{(3)}$$

where

$$F = \frac{4R}{(1-R)^2} \qquad \textbf{(4)}$$

is called the coefficient of Finesse. One can immediately see that when $R << 1$, F is small and the reflectivity is proportional to $\sin^2(\delta/2)$. The same intensity distribution is obtained in the two-beam interference pattern (see Sec. 14.7); we obtained $\sin^2(\delta/2)$ instead of $\cos^2(\delta/2)$ because of the additional phase change of π in one of the reflected beams.

Similarly, the amplitude of the successive transmitted waves will be

$$A_0 t_1 t_2,\ A_0 t_1 t_2 r_2^2 e^{i\delta},\ A_0 t_1 t_2 r_2^4 e^{2i\delta},\ \ldots$$

where, without any loss of generality, we have assumed the first transmitted wave to have zero phase. Thus the resultant amplitude of the transmitted wave will be given by

$$A_t = A_0 t_1 t_2\,(1 + r_2^2 e^{i\delta} + r_2^4 e^{2i\delta} + \cdots)$$

$$= A_0\,\frac{t_1 t_2}{1 - r_2^2 e^{i\delta}} = A_0\,\frac{1-R}{1-Re^{i\delta}}$$

Thus the transmittivity T of the film is given by

$$T = \left|\frac{A_t}{A_0}\right|^2 = \frac{(1-R)^2}{(1-R\cos\delta)^2 + R^2\sin^2\delta}$$

or
$$T = \frac{1}{1 + F\sin^2\delta/2} \qquad \textbf{(5)}$$

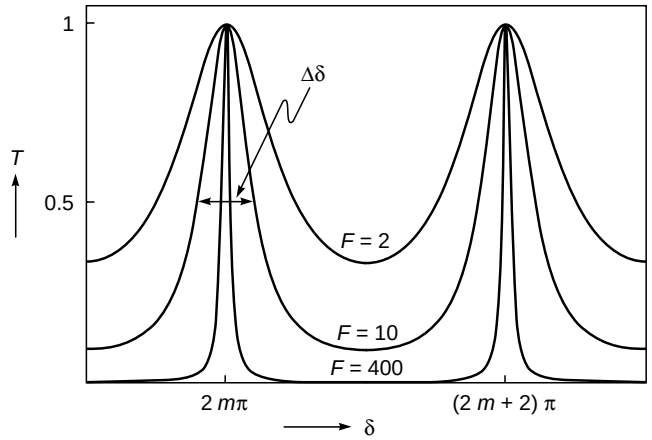


Fig. 16.2 The transmittivity of a Fabry–Perot etalon as a function of δ for different values of *F;* the value of *m* is usually large. The transmission resonances become sharper as we increase the value of *F*. The FWHM is denoted by Δδ.

It is immediately seen that the reflectivity and the transmittivity of the Fabry–Perot etalon add to unity. Further,

$$T = 1$$

when

$$\delta = 2m\pi \qquad m = 1, 2, 3, \ldots \tag{6}$$

In Fig. 16.2 we have plotted the transmittivity as a function of δ for different values of *F*. To get an estimate of the width of the transmission resources, let

$$T = \frac{1}{2} \qquad \text{for } \delta = 2m\pi \pm \frac{\Delta\delta}{2}$$

Thus

$$F \sin^2 \frac{\Delta\delta}{4} = 1 \tag{7}$$

The quantity Δδ represents the FWHM (full width at half maximum). In almost all cases, Δδ <<< 1, and therefore, to a very good approximation, it is given by

$$\Delta\delta \approx \frac{4}{\sqrt{F}} = \frac{2(1-R)}{\sqrt{R}} \tag{8}$$

Thus the transmission resources become sharper as the value of *F* increases (see Fig. 16.2).

16.3 THE FABRY–PEROT ETALON

In this section, we will discuss the Fabry–Perot interferometer which is based on the principle of multiple-beam interferometry discussed in Sec. 16.2. The interferometer (as shown in Fig. 16.3) consists of two plane glass (or quartz) plates which are coated on one side with a partially reflecting metallic film[2] (of aluminum or silver) of about 80% reflectivity. These two plates are kept in such a way that they enclose a plane parallel slab of air between their coated surfaces. If the reflecting glass plates are held parallel to each other at a fixed separation, we have a *Fabry–Perot etalon*. In fact, we may neglect the presence of the plates and consider only the reflection (and transmission) by the metallic film; further, if the plates are parallel, the rays will not undergo any deviation.

In a typical experiment, light from a broad source is collimated by a lens and is passed through the Fabry–Perot etalon as shown in Fig. 16.3. Thus, if we consider light of a specific wavelength λ_0, the incident light will be completely transmitted (i.e., $T = 1$) if the angle of incidence is such that

$$\delta = \frac{4\pi}{\lambda_0} n_2 h \cos\theta_2 = 2m\pi \tag{9}$$

or

$$\cos\theta_2 = \frac{m\lambda_0}{2n_2 h} \tag{10}$$

For large values of *F*, when θ_2 is slightly different from the value given by the above equation, the transmittivity will be very small. Hence, for a given wavelength, at the focal plane of lens *L*, we will obtain a fringe pattern consisting of concentric rings—each bright ring will correspond to a particular value of *m*. The sharpness of the bright rings (and hence the resolving power of the etalon) will increase with the value of *F*.

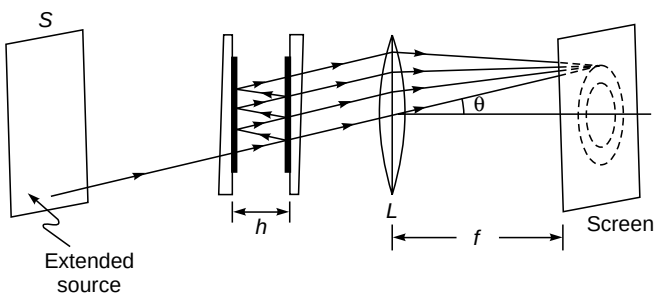


Fig. 16.3 The Fabry–Perot etalon.

[2] In the visible region of the spectrum, silver is the best metal to coat with (the reflectivity is about 0.97 in the red region and decreases to about 0.90 in the blue region). But beyond the blue region, the reflectivity falls rapidly. Aluminum is usually employed below 4000 Å.

Example 16.1 As an example, we assume an etalon with $n_2 = 1$, $h = 1$ cm, and $F = 400$ ($F = 400$ implies $R \approx 0.905$; i.e., each mirror of the etalon has about 90% reflectivity). In Fig. 16.4 we have plotted the intensity variation with θ for $\lambda_0 = 5000$ and 4999.98 Å. The actual fringe pattern (as obtained on the focal plane of a lens of focal length 25 cm) is shown in Fig. 16.5. Now, for

$$\lambda_0 = \lambda_1 = 5000 \text{ Å}$$

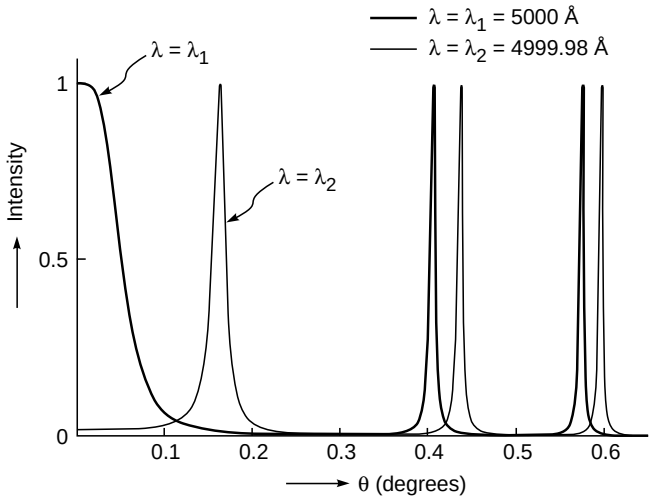


Fig. 16.4 The variation of intensity with θ for a Fabry–Perot interferometer with $n_2 = 1$, $h = 1.0$ cm, and $F = 400$, corresponding to $\lambda_0 = 5000$ Å (= λ_1) and $\lambda_0 = 4999.98$ Å (= λ_2).

Fig. 16.5 The (computer generated) ring pattern as obtained (on the focal plane of a lens) in a Fabry–Perot etalon with $n_2 = 1$, $h = 1.0$ cm and $F = 400$, corresponding to $\lambda_0 = 5000$ Å (=λ_1) and $\lambda_0 = 4999.98$ Å (=λ_2).

Eq. (9) gives us

$$\theta_2 = \cos^{-1} \frac{m}{40000}$$

Thus bright rings will form at

$$\theta_2 = 0°,\ 0.41°,\ 0.57°,\ 0.70°,\ \ldots$$

corresponding to m = 40,000; 39,999; 39,998; 39,997; . . . , respectively. This is shown as the thick curve in Fig. 16.4. On the other hand, for

$$\lambda_0 = \lambda_2 = 4999.98 \text{ Å}$$

we get

$$\theta_2 = \cos^{-1} \frac{m}{40000.16}$$

Thus bright rings will form at

$$\theta_2 = 0.162°,\ 0.436°,\ 0.595°,\ \ldots$$

corresponding to m = 40,000; 39,999; and 39,998 respectively. This is shown as the thin curve in Fig. 16.4. The corresponding ring patterns as obtained on the focal plane of the lens are shown in Fig. 16.5; from the figure we can see that the two spectral lines having a small wavelength difference of 0.02 Å are quite well resolved by the etalon. In the figure, the central bright spot and the first ring correspond, respectively, to $\lambda_0 = 5000$ and 4999.98 Å, both corresponding to $m = 40000$. The next two closely spaced rings correspond to $m = 39999$ for the two wavelengths.

16.3.1 Flatness of the Coated Surfaces

To have sharp fringes, the coated surfaces should be parallel to a very high degree of accuracy. Indeed, the coated surfaces should be flat within about λ/50, where λ is the wavelength of light. To see this, we assume that in the above example h is increased by λ/20 (= 250 Å = 2.5×10^{-6} cm):

$$h = 1 + 2.5 \times 10^{-6} = 1.0000025 \text{ cm}$$

For $\lambda_0 = 5000$ Å, we will have

$$\theta_2 = \cos^{-1} \frac{m}{40,000.1}$$

and bright rings will form at

$$\theta_2 = 0.128°, 0.425°, 0.587°, \ldots$$

If we compare the results obtained in Example 16.1, we find that if there is a variation in the spacing by about λ/20, the fringes corresponding to the wavelengths 5000 and 4999.98 Å will start overlapping. Thus the coated surfaces should be parallel within a very small fraction of the wavelength. Further, the two uncoated surfaces of each plate are made to have a slight angle between them (~1 to 10 minutes, see Fig. 16.3) so that one could avoid the unwanted fringes formed due to multiple reflections in the plate itself.

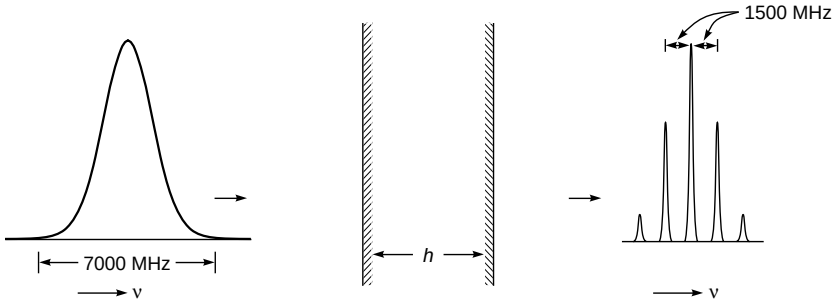


Fig. 16.6 A beam having a spectral width of about 7000 MHz (around $\nu_0 = 6 \times 10^{14}$ Hz) is incident normally on a Fabry–Perot etalon with h = 10 cm and $n_2 = 1$. The output has five narrow spectral lines.

16.3.2 Modes of the Fabry-Perot Cavity

We consider a polychromatic beam incident normally ($\theta_2 = 0$) on a Fabry–Perot etalon with air between the reflecting plates ($n_2 = 1$); see Fig. 16.6. In terms of the frequency

$$\nu = \frac{c}{\lambda_0}$$

Eq. (9) tells us that transmission resonance will occur when

$$\nu = \nu_m = m\frac{c}{2h} \qquad \textbf{(11)}$$

where m is an integer. The above equation represents the different (longitudinal) modes of the (Fabry–Perot) cavity. For h = 10 cm, the frequency spacing of two adjacent modes is given by

$$\delta\nu = \frac{c}{2h} = 1500 \text{ MHz}$$

For an incident beam having a central frequency of

$$\nu = \nu_0 = 6 \times 10^{14} \text{ Hz}$$

and a spectral width[3] of 7000 MHz, the output beam will have frequencies

$$\nu_0 \qquad \nu_0 \pm \delta\nu \qquad \text{and} \qquad \nu_0 \pm 2\delta\nu$$

as shown in Fig. 16.6. One can readily calculate from Eq. (11) that the five lines correspond to

m = 399,998; 399,999; 400,000; 400,001 and 400,002

Figure 16.7 shows a typical output of a multilongitudinal (MLM) laser diode. The wavelength spacing between two modes is about 0.005 μm.

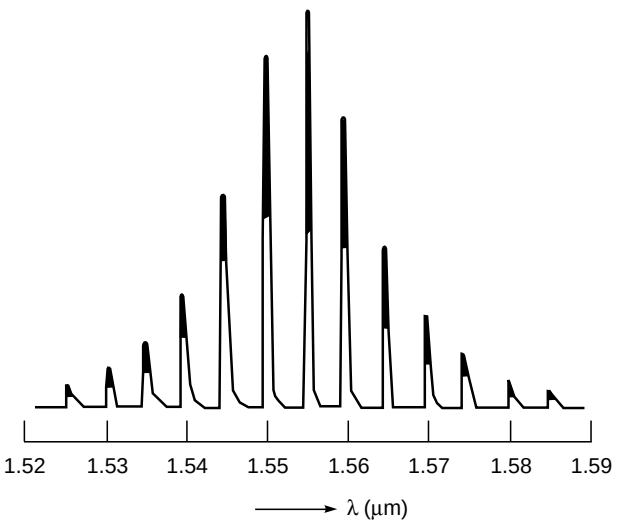


Fig. 16.7 Typical output spectrum of a Fabry–Perot multilongitudinal mode laser diode; the wavelength spacing between two modes is about 0.005 μm (After Ref. 9).

16.4 THE FABRY-PEROT INTERFEROMETER

If one of the mirrors is kept fixed while the other is capable of moving to change the separation between the two mirrors, the system is called a Fabry–Perot interferometer. For a beam incident normally on the interferometer, we vary the separation h and measure the intensity variation on the focal plane of lens L as shown in Fig. 16.8. Such an arrangement is usually referred to as a scanning Fabry–Perot interferometer. Since the separation h is varied, we write it as

$$h = h_0 + x \qquad \textbf{(12)}$$

If the incident beam is monochromatic, a typical variation of intensity at point P is shown in Fig. 16.9. The figure corresponds to the frequency of the incident beam being

$$\nu = \nu_0 = 6 \times 10^{14} \text{ Hz}$$

[3] For $\nu_0 = 6 \times 10^{14}$ Hz, $\lambda_0 = 5000$ Å and a spectral width of 7000 MHz would imply $\left|\Delta\lambda_0/\lambda_0\right| = \left|\Delta\nu/\nu_0\right| = 7 \times 10^9/6 \times 10^{14} \approx 1.2 \times 10^{-5}$, giving $\Delta\lambda_0 \approx 0.06$ Å. Thus a frequency spectral width of 7000 MHz (around $\nu_0 = 6 \times 10^{14}$ Hz) implies a wavelength spread of only 0.06 Å.

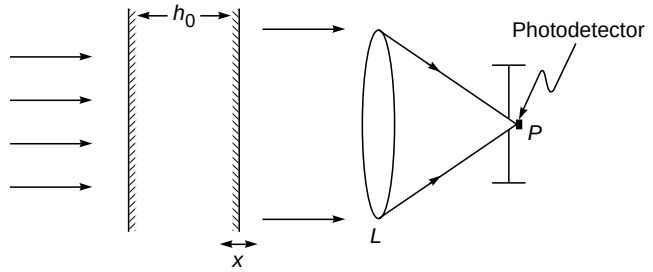


Fig. 16.8 A scanning Fabry–Perot interferometer. The intensity variation is recorded (by a photodetector) on the focal plane of lens *L*.

For $h_0 = 10$ cm, $n_2 = 1$, and $\cos\theta_2 = 1$, we get

$$\delta = \frac{4\pi\nu_0(h_0 + x)}{c}$$

$$= 800000\pi\left(1 + \frac{x}{h_0}\right)$$

Thus transmission resonances will occur for

$$\delta = 800000\pi, 800002\pi, 800004\pi, \ldots$$

which will occur when

$$x = 0, 250 \text{ nm}, 500 \text{ nm}, \ldots$$

respectively. The two curves in Fig. 16.9 correspond to $F = 100$ and $F = 1000$. Notice that the transmission resonances become sharper if we increase the value of F. Figure 16.10 shows variation of intensity at point P when the incident beam has two frequencies separated by 300 MHz. Obviously, the two frequencies are well resolved.

If the frequency of the incident beam is increased by $c/(2h_0)$, i.e., if

$$\nu = \nu_0 + \frac{c}{2h_0}$$

then one can easily show that transmission resonances will occur at the same values of x, and the corresponding values of δ will be $800{,}002\pi$ (corresponding to $x = 0$), $800{,}004\pi$ (corresponding to $x = 250$ nm), etc. Indeed, if

$$\nu = \nu_0 \pm p\frac{c}{2h_0} \qquad p = 1, 2, 3, \ldots$$

we will have the same T versus x curve. The quantity

$$\Delta\nu_s = \frac{c}{2h_0} \tag{13}$$

is known as the free spectral range (FSR) of the interferometer. Thus when the spectrum has widely separated wavelength components, we have overlapping of orders.

16.5 RESOLVING POWER

We will first consider the resolving power corresponding to a beam incident normally on a scanning Fabry–Perot interferometer. This will be followed by the case corresponding to the Fabry–Perot etalon.

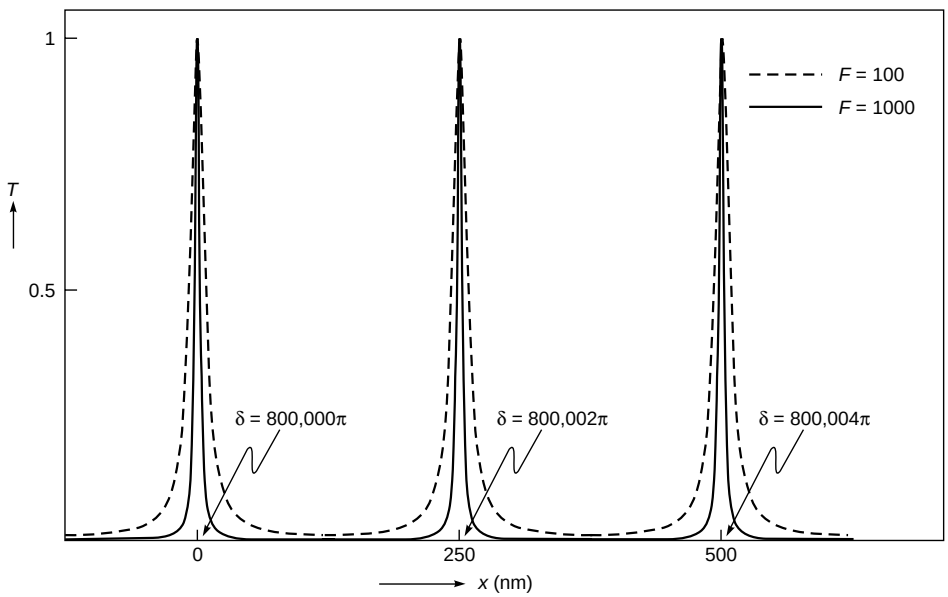


Fig. 16.9 Variation of intensity at point P with x (see Fig. 16.8) for a monochromatic beam incident normally on a scanning Fabry–Perot interferometer; the solid curve corresponds to $F = 1000$, and the dashed curve corresponds to $F = 100$.

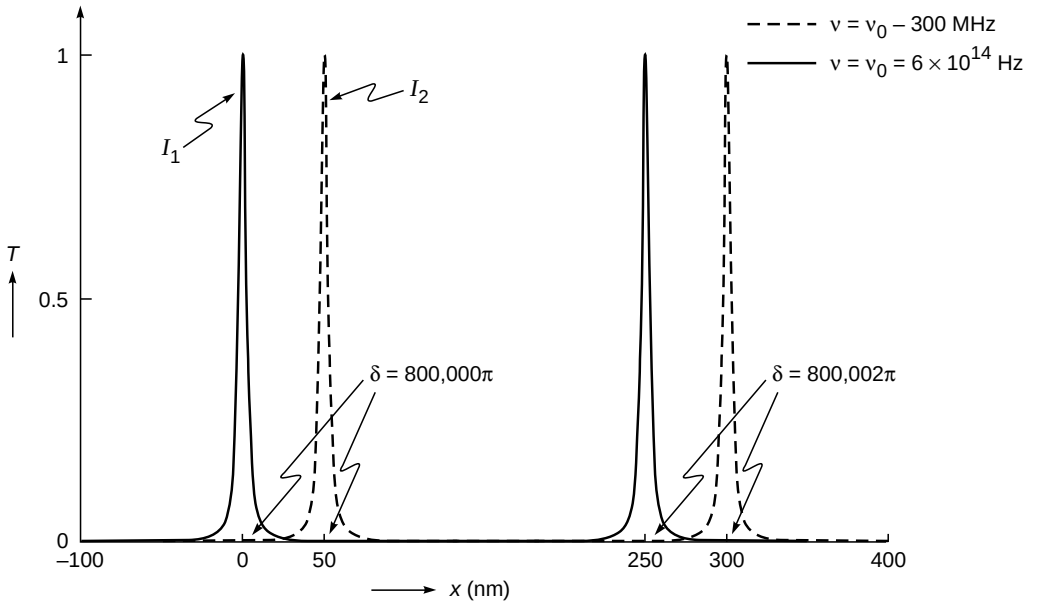


Fig. 16.10 Variation of intensity at point P with x (see Fig. 16.8) when the incident beam has two frequencies separated by 300 MHz.

16.5.1 Resolving Power of a Scanning Fabry–Perot Interferometer

We consider the presence of two frequencies ν_1 and ν_2 of equal intensity in the beam incident normally on a scanning Fabry–Perot interferometer. For the two frequencies to be just resolved, we assume that the half intensity point of ν_1 falls on the half intensity point of ν_2 as shown in Fig. 16.11. When this happens, the minimum of the resultant intensity distribution (shown as the dashed curve in Fig. 16.11) is about 74% of the corresponding maximum value. Now, as discussed in Sec. 16.2, if the half intensity point occurs at

$$\delta = \delta_{1/2} = 2m\pi \pm \frac{\Delta\delta}{2} \qquad \textbf{(14)}$$

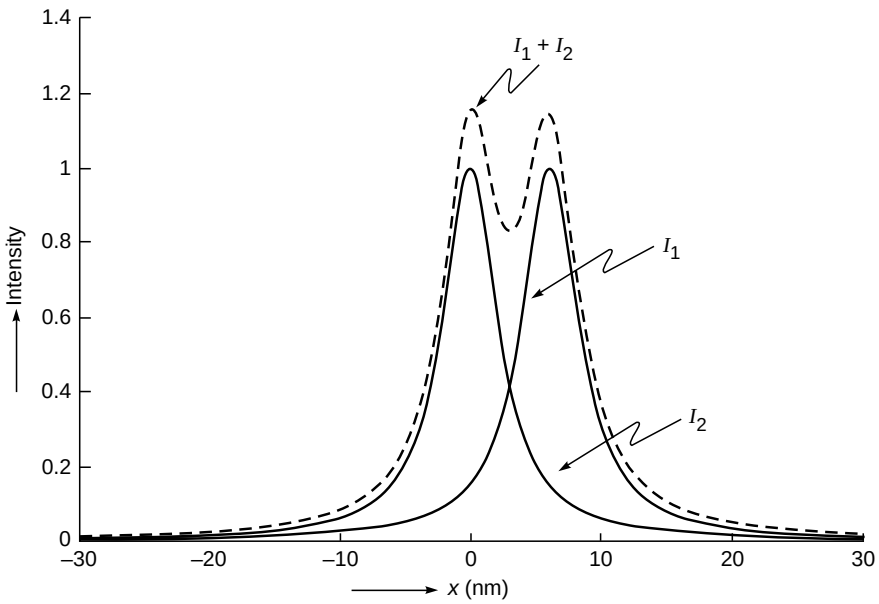


Fig. 16.11 The individual intensity variations I_1 and I_2 in the presence of two frequencies ν_1 and ν_2 and the total intensity variation $I_1 + I_2$ when the two frequencies are just resolved.

then

$$\Delta\delta \approx \frac{4}{\sqrt{F}} \tag{15}$$

[see Eq. (8)]. Consider the frequency ν_1. If the intensity maximum occurs at $h = h_1$ then

$$\delta_1 = \frac{4\pi h_1 \nu_1}{c} = 2m\pi \tag{16}$$

Let the intensity maximum for $\nu = \nu_2$ $(= \nu_1 + \Delta\nu_1)$ occur at

$$h = h_2 = h_1 + \Delta h_1$$

Thus

$$\delta_2 = \frac{4\pi (h_1 + \Delta h_1)(\nu_1 + \Delta\nu_1)}{c} = 2m\pi \tag{17}$$

Using Eqs. (16) and (17) and neglecting the second-order term $\Delta h_1 \Delta\nu_1$, we get

$$\nu_1 \Delta h_1 + h_1 \Delta\nu_1 = 0$$

or

$$\Delta h_1 = -\frac{h_1}{\nu_1}\Delta\nu_1 \tag{18}$$

Equation (18) implies that for Δh_1 to be positive, $\Delta\nu_1$ should be negative. Now, for the frequency ν_1, let the half intensity point occur at $h = h_1 + \delta h_1$ (the corresponding value of δ will be $2m\pi + \frac{1}{2}\Delta\delta_1$; thus using Eq. (16), we find

$$\frac{4\pi\nu_1 \delta h_1}{c} = \frac{1}{2}\Delta\delta_1 \approx \frac{2}{\sqrt{F}} \tag{19}$$

or

$$\Delta h_1 \approx \frac{c}{2\pi\nu_1\sqrt{F}} \tag{20}$$

For the two frequencies to be just resolved

$$\Delta h_1 = 2\delta h_1 \approx \frac{c}{\pi\nu_1\sqrt{F}} \tag{21}$$

Using Eq. (18), we get for the resolving power

$$\left|\frac{\nu_1}{\Delta\nu_1}\right| = \frac{h_1}{\Delta h_1} = \frac{\pi h_1 \nu_1 \sqrt{F}}{c}$$

Or dropping the subscript, we get

$$\text{Resolving power} = \left|\frac{\nu}{\Delta\nu}\right| = \frac{\pi h\nu\sqrt{F}}{c} \tag{22}$$

Or, in terms of the wavelength,

$$\text{Resolving power} = \left|\frac{\lambda_0}{\Delta\lambda_0}\right| = \frac{\pi h\sqrt{F}}{\lambda_0} \tag{23}$$

For $h = 1$ cm, and $\lambda_0 = 6 \times 10^{-5}$ cm

$$\Delta\lambda \approx \begin{cases} 0.013\,\text{Å} & \text{for } F = 80 \\ 0.006\,\text{Å} & \text{for } F = 360 \end{cases}$$

16.5.2 Resolving Power of a Fabry–Perot Etalon

We consider light from a broad source incident on a Fabry–Perot etalon as shown in Fig. 16.3. We once again consider the presence of two wavelengths λ_1 and λ_2 of equal intensity. Now, $T = 1$ if the angle of incidence is such that [see Eq. (9)]

$$\delta = \frac{4\pi\nu}{c} h\mu = 2m\pi \tag{24}$$

where $\mu = \cos\theta$, and for the sake of simplicity, we have dropped the subscript on μ and θ. We can now have arguments very similar to those in Sec. 16.5.1 except now h is fixed and μ $(= \cos\theta)$ is varied. Thus, if the mth-order intensity maxima for $\nu = \nu_1$ and $\nu = \nu_2$ $(= \nu_1 + \Delta\nu_1)$ occur at $\mu = \mu_1$ and $\mu = \mu_2$ $(= \mu_1 + \Delta\mu_1)$, then

$$\delta_1 = \frac{4\pi\nu_1 h\mu_1}{c} = 2m\pi \tag{25}$$

and

$$\delta_2 = \frac{4\pi h(\nu_1 + \Delta\nu_1)(\mu_1 + \Delta\mu_1)}{c} = 2m\pi \tag{26}$$

Thus, neglecting the second-order term we get

$$\Delta\mu_1 = -\frac{\mu_1}{\nu_1}\Delta\nu_1 \tag{27}$$

Now, for the frequency ν_1, let the half intensity point occur at $\mu = \mu_1 + \delta\mu_1$ (the corresponding value of δ will be $2m\pi + \frac{1}{2}\Delta\delta_1$); thus using Eq. (24) gives

$$\frac{4\pi\nu_1 h\delta\mu_1}{c} = \frac{1}{2}\Delta\delta_1 \approx \frac{2}{\sqrt{F}} \tag{28}$$

or

$$\delta\mu_1 \approx \frac{c}{2\pi\nu_1 h\sqrt{F}} \tag{29}$$

As discussed earlier, for the two frequencies to be just resolved, we assume that the half intensity point of ν_1 falls on the half intensity point of ν_2, giving

$$\Delta\mu_1 = 2\delta\mu_1 \approx \frac{c}{\pi\nu_1 h\sqrt{F}} \tag{30}$$

Using Eq. (27), we get

$$\text{Resolving power} = \left|\frac{\nu_1}{\Delta\nu_1}\right| = \frac{\mu_1}{\Delta\mu_1} = \frac{\pi\nu_1 h\sqrt{F}\mu_1}{c} \tag{31}$$

Or, in terms of the wavelength,

$$\text{Resolving power} = \left|\frac{\lambda_0}{\Delta\lambda_0}\right| = \frac{\pi h\sqrt{F}\cos\theta}{\lambda_0} \tag{32}$$

Thus for $F = 360$ $(R = 0.9)$, $h = 1$ cm, and $\lambda_0 = 5000$ Å,

$$\left|\frac{\lambda_0}{\Delta\lambda_0}\right| \approx 1.2 \times 10^6$$

where we have assumed normal incidence. The above equation gives

$$\Delta\lambda_0 \approx 0.004 \text{ Å}$$

Thus a Fabry–Perot instrument can resolve wavelengths differing by about 10^{-3} Å. This is in contrast to a grating (say, having 25,000 grooves) which resolves up to about 0.1 Å at $\lambda = 5000$ Å and a prism (made of dense flint glass with 5 cm base) which resolves only up to about 1 Å at 5000 Å. Note that in the above analysis, we have considered two monochromatic lines at λ and $\lambda + \Delta\lambda$. In general, the lines at the two wavelengths λ and $\lambda + \Delta\lambda$ themselves will have a wavelength spread, and this restricts the use of such high resolving powers.

When the Fabry–Perot interferometer is used to analyze spectra with closely spaced lines, then the distance between the adjacent maxima is greater than the displacement between the system of rings of the spectral lines. But when the spectrum has widely separated wavelength components, then the displacement between the rings may be greater than the separation between adjacent maxima. The results in the "overlapping" of orders (see also the discussion at the end of Sec. 16.4). The difference in wavelength $\Delta\lambda_s$ which corresponds to a displacement of one order is called the spectral range of the interferometer. Thus we can write

$$\Delta\lambda_s = \frac{\lambda^2}{2nh\cos\theta} \qquad \textbf{(33)}$$

This becomes, for near normal incidence ($\theta \approx 0$),

$$\Delta\lambda_s = \frac{\lambda^2}{2nh} \qquad \textbf{(34)}$$

which is found to be inversely proportional to h. This is in contrast to the resolving power which depends directly on h [see Eqs. (31) and (32)].

When the spectrum is complex consisting of a number of widely separated wavelength components, each with a hyperfine structure, then one can separate the different wavelength components by employing the Fabry–Perot interferometer along with a spectrograph as shown in Fig. 16.12(a). The light emerging from source S is rendered parallel by lens L_1. The interference pattern formed by the Fabry–Perot interferometer (marked by FP in the figure) is made to fall on the slit of the spectrograph. The spectrograph separates the spectral components, and one obtains in plane P images of the slit, each crossed by fringes as shown in Fig. 16.12(b).

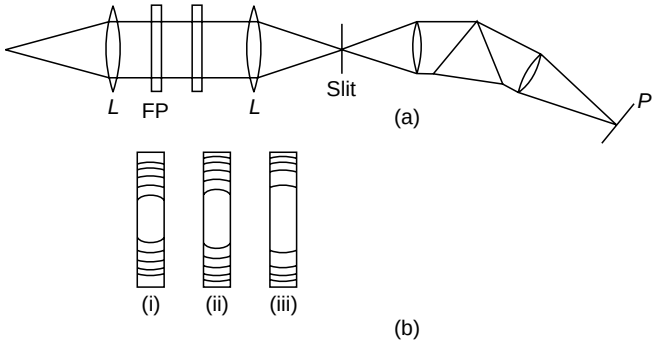


Fig. 16.12 (a) A Fabry–Perot interferometer used in conjunction with a spectrograph. (b) The interlaced fringes formed in the plane of the slit are separated by the prism. For example, (i), (ii), and (iii) may correspond to the lines in the red, yellow, and green regions, respectively, as observed on plane P.

16.6 THE LUMMER–GEHRCKE PLATE[4]

We saw in Sec. 16.2 that the sharpness of fringes (and hence the resolving power) of a multiple-beam interferometer increases as the reflectivity R of the plate increases. But one cannot use every thick coating of metals to increase the reflectivity as the intensity of the beam would be reduced considerably due to absorption in metallic coatings. This difficulty can be overcome by the use of the phenomenon of total internal reflection (instead of metallic reflection); this is used in the Lummer–Gehrcke plate.

A Lummer–Gehrcke plate is a plane parallel made of glass (or quartz), on one end of which a small right-angle prism of the same material is fixed (see Fig. 16.13). The angle of the prism is chosen in such a way that the rays incident normally on the surface of the prism hit the two surfaces of the plate at an angle slightly less than the critical angle.[5] Since the two surfaces are parallel, all successive reflections will occur at the same (near critical) angle. Most of the light will be reflected with a little fraction being transmitted at each reflection. Thus, there will emerge from the upper and lower surfaces of the plate a series of waves which will finally interfere to produce interference fringes in plane P (see Fig. 16.13). Notice that the prism suppresses the externally reflected beam. In plane P, one obtains fringe patterns on either side of the plate. The fringes are approximately straight lines parallel to the plate surfaces.

[4] Sections 16.6 and 16.7 have been very kindly written by Prof. Anurag Sharma.

[5] Beyond the critical angle, the reflection is total, while slightly below the critical angle, the reflectivity is high (see Sec. 24.2).

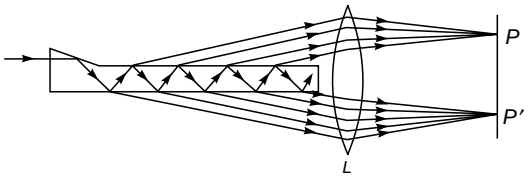


Fig. 16.13 The Lummer–Gehrcke plate.

We will not go into the details of the theory of the Lummer–Gehrcke plate, but we note two points:

1. Unlike in the Fabry–Perot interferometer, the space between the reflecting surfaces is a dispersive medium.
2. The number of reflections is also not very large as in the case of the Fabry–Perot interferometer; the number of reflections depends on the length of the plate and the angle θ (see Fig. 16.13). Thus, the resolving power of the instrument depends on the length of the plate.

Earlier, Lummer–Gehrcke plates were used in high-resolution spectroscopy. However, they have been replaced by the more flexible Fabry–Perot interferometer.

16.7 INTERFERENCE FILTERS

When a Fabry–Perot interferometer is illuminated by a monochromatic (uncollimated) beam, we get a spectrum consisting of different intensity maxima which satisfy the following relation:

$$2nh \cos \theta_r = m\lambda \qquad \textbf{(35)}$$

Now if a Fabry–Perot interferometer is illuminated with a collimated white light incident normally ($\theta_r \approx 0$), maxima of different orders are formed in the transmitted light corresponding to wavelengths given by

$$\lambda = \frac{2nh}{m} \qquad \textbf{(36)}$$

If h is large, a large number of maxima will be observed in the visible region; for example, about 23,000 maxima are observed if h = 1 cm. But if we go on reducing h, we reach a situation in which only one or two maxima are obtained in the visible region. For example, if n = 1.5 and $h = 6 \times 10^{-5}$ cm, there are only two maxima in the visible region, corresponding to λ = 6000 Å (m = 3) and λ = 4500 Å (m = 4). They are widely separated, and one of them can be masked so as to transmit only one wavelength. In this way, it is possible to filter a particular wavelength out of a white light beam. Such a structure is known as an interference filter.[6] Interference filters

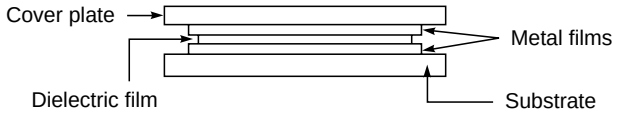


Fig. 16.14 The interference filter.

using this principle can be obtained by modern vacuum deposition techniques. A thin metallic film (usually of aluminum or silver) is deposited on a substrate (generally, a glass plate) by vacuum deposition techniques. Then a thin layer of a dielectric material such as cryolite ($3NaF{\cdot}AlF_3$) is deposited over this. This structure is again covered by another metallic film (see Fig. 16.14). To protect this film structure from any damage, another glass plate is placed over it. Thus a Fabry–Perot structure is formed between the two glass plates. By varying the thickness of the dielectric film, one can filter out any particular wavelength. However, the filtered light will have a finite width; i.e., it will have a narrow spectrum sharply peaked about one wavelength. The sharpness of the transmitted spectrum is determined by the resolving power of the formed Fabry–Perot structure, and hence by the reflectivity of the surfaces. The larger the reflectivity, the narrower is the transmitted spectrum. But it is not possible to increase the thickness of the metallic films indefinitely as absorption will reduce the intensity of the transmitted light. To overcome this difficulty, metallic films are replaced by all dielectric structures.

In an all-dielectric structure, layers of dielectric materials of appropriate refractive indices are deposited. It was shown in Chap. 15 how dielectric films can be used to enhance the reflectivity of a surface. If, on a glass plate, a $\lambda/4$ thick film of a dielectric material whose refractive index is more than that of glass is deposited, the reflectivity of the glass plate increases. The larger the difference between the refractive indices, the greater will be the reflectivity. The materials generally used in interference filters are titanium oxide (n = 2.8) or zinc sulfide (n = 2.3). To obtain interference filters, a $\lambda/4$ thick film of titanium oxide is deposited on a glass substrate. Then a thin layer of dielectric material with lower refractive index (such as cryolite or magnesium fluoride) is deposited. On this is again deposited a $\lambda/4$ thick layer of a material of higher refractive index. To increase the reflectivity, multilayer structures of alternate higher and lower refractive index materials are used. In this way, it is possible to achieve a reflectivity of more than 90% for any particular wavelength (see Sec. 15.6 for a more detailed account). Thus if the incident wave is polychromatic (like white light), the reflected light may have a high degree of monochromaticity.

[6] The Fabry–Perot structure also behaves as a resonator and supports the oscillation of what are known as modes.

Summary

- If a plane wave falls on a plane parallel film, then the beam undergoes multiple reflections at the two surfaces and a large number of beams of successively diminishing amplitude will emerge on both sides of the plate. These beams (on either side) interfere to produce an interference pattern at infinity. If the reflectivity R at each surface is close to unity, then the fringes so formed are much sharper than those formed by two-beam interference and, therefore, the interferometers involving multiple-beam interference have a high resolving power and hence find applications in-high resolution spectroscopy. The transmittivity of such a film is given by

$$T = \frac{1}{1 + F \sin^2 \delta/2}$$

where $F = 4R/(1 - R)^2$ is known as coefficient of Finesse and

$$\delta = \frac{4\pi n_2 h \cos\theta_2}{\lambda_0}$$

represents the phase difference (between two consecutive waves emanating from the film) due to the additional path traversed by the beam in the film; θ_2 is the angle of refraction inside the film (of refractive index n_2), h is the film thickness, and λ_0 is the free space wavelength. The transmittivity $T = 1$ when $\delta = 2m\pi$, $m = 1, 2, 3, \ldots$. For $R \approx 1$, the value of F is very large and the transmission resonances become very sharp. This is the principle used in the Fabry–Perot interferometer which is characterized by a high resolving power.

Problems

16.1 Calculate the resolving power of a Fabry–Perot interferometer made of reflecting surfaces of reflectivity 0.85 and separated by a distance 1 mm at $\lambda = 4880$ Å.

16.2 Calculate the minimum spacing between the plates of a Fabry–Perot interferometer which will resolve two lines with $\Delta\lambda = 0.1$ Å at $\lambda = 6000$ Å. Assume the reflectivity to be 0.8.

16.3 Consider a monochromatic beam of wavelength 6000 Å incident (from an extended source) on a Fabry–Perot etalon with $n_2 = 1$, $h = 1$ cm, and $F = 200$. Concentric rings are observed on the focal plane of a lens of focal length 20 cm.

(a) Calculate the reflectivity of each mirror.

(b) Calculate the radii of the first four bright rings. What will be the corresponding values of m?

(c) Calculate the angular width of each ring where the intensity falls by one-half and the corresponding FWHM (in mm) of each ring.

16.4 Consider now two wavelengths 6000 and 5999.9 Å incident on a Fabry–Perot etalon with the same parameters as given in Prob. 16.3. Calculate the radii of the first three bright rings corresponding to each wavelength. What will be the corresponding values of m? Will the lines be resolved?

16.5 Consider a monochromatic beam of wavelength 6000 Å incident normally on a scanning Fabry–Perot interferometer with $n_2 = 1$ and $F = 400$. The distance between the two mirrors is written as $h = h_0 + x$. Given $h_0 = 10$ cm:

(a) Calculate the first three values of x for which we will have unit transmittivity and the corresponding values of m.

(b) Also calculate the FWHM Δh for which the transmittivity will be one-half.

(c) What would be the value of Δh if F were 200?

[**Ans.:** (a) $x \approx 200$ nm (m = 333,334), 500 nm (m = 333,335); (b) $\Delta h \approx 9.5$ nm]

16.6 In continuation of Prob. 16.5, consider now two wavelengths λ_0 (= 6000 Å) and $\lambda_0 + \Delta\lambda$ incident normally on the Fabry–Perot interferometer with $n_2 = 1$, $F = 400$, and $h_0 = 10$ cm. What will be the value of $\Delta\lambda$ so that $T = ½$ occurs at the same value of h for both wavelengths?

16.7 Consider a laser beam incident normally on the Fabry–Perot interferometer as shown in Fig. 16.15.

(a) Assume $h_0 = 0.1m$, $c = 3 \times 10^8$ m/s, and $\nu = \nu_0 = 5 \times 10^{14}$ s^{-1}. Plot T as a function of x (-100 nm $< x <$ 400 nm) for F = 200 and F = 1000.

(b) Show that if $\nu = \nu_0 \pm p$ (1500 MHz), $p = 1, 2, \ldots,$) we will have the same T versus x curve; 1500 MHz is known as the free spectral range (FSR). What will be the corresponding values of δ?

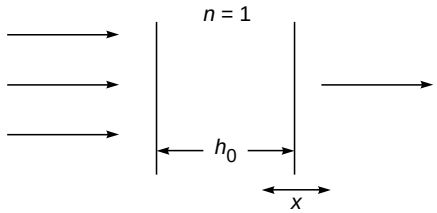


Fig. 16.15

REFERENCES AND SUGGESTED READINGS

1. R. Baierlein, *Newton to Einstein: The Trail of Light*, Cambridge University Press, 1992.
2. P. Baumeister and G. Pincus, "Optical Interference Coatings," *Scientific American*, Vol. 223, p. 59, December 1970.
3. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, 1975.
4. M. Cagnet, M. Francon, and S. Mallick, *Atlas of Optical Phenomena*, Springer-Verlag, Berlin, 1971.

5. R. W. Ditchburn, *Light*, Academic Press, London, 1976.
6. M. Francon, *Modern Applications of Physical Optics*, Interscience, New York, 1963.
7. M. Francon, *Optical Interferometry*, Academic Press, New York, 1966.
8. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, McGraw-Hill Book Co., New York, 1976.
9. C. Lin, "Optical Communications: Single Mode Optical Fiber Transmission Systems," *Optoelectronic Technology and Lightwave Communications Systems*, Ed. C. Lin, Van Nostrand Reinhold, New York, 1989.
10. W. H. Steel, *Interferometry*, Cambridge University Press, Cambridge, London, 1967.
11. S. Tolansky, *Multiple Beam Interferometry of Surfaces and Films*, Oxford University Press, London, 1948.
12. S. Tolansky, *An Introduction to Interferometry*, Longmans Green and Co., London, 1955.

Chapter Seventeen

COHERENCE

Light which is capable of interference is called 'coherent,' and it is evident that in order to yield many interference fringes, it must be very monochromatic. Coherence is conveniently measured by the path difference between two rays of the same source, by which they can differ while still giving observable interference contrast. This is called the coherence length. . . . Lord Rayleigh and Albert Michelson were the first to understand that it is a reciprocal measure of the spectroscopic line width. Michelson used it for ingenious methods of spectral analysis and for the measurement of the diameter of stars.

—Dennis Gabor in his Nobel Lecture on Holography, December 11, 1971

17.1 INTRODUCTION

In earlier chapters on interference we assumed that the displacement associated with a wave remained sinusoidal for all values of time. Thus the displacement (which we denote by E) was assumed to be given by

$$E = A \cos (kx - \omega t + \phi)$$

The above equation predicts that at any value of x, the displacement is sinusoidal for $-\infty < t < \infty$. For example, at $x = 0$ we have [see Fig. 17.1(a)]

$$E = A \cos (\omega t - \phi) \qquad -\infty < t < \infty \qquad \textbf{(1)}$$

Obviously this corresponds to an idealized situation because the radiation from an ordinary light source consists of finite size wave trains, a typical variation of which is shown in Fig. 17.1(b). Since we will be considering only light waves, the quantity E represents the electric field associated with the light wave. Now, in Fig. 17.1(b), τ_c represents the average duration of the wave trains; i.e., the electric field remains sinusoidal for times of the order of τ_c. Thus, at a given point, the electric fields at times t and $t + \Delta t$ will, in general, have a definite phase relationship if $\Delta t \ll \tau_c$ and will (almost) never have any phase relationship if $\Delta t \gg \tau_c$. The time duration τ_c is known as the coherence time of the source, and the field is said to remain coherent for times $\sim\tau_c$. The length of the wave train, given by

$$L = c\tau_c \qquad \textbf{(2)}$$

(where c is the speed of light in free space) is referred to as the coherence length. For example, for the neon line (λ = 6328 Å), $\tau_c \sim 10^{-10}$ s, and for the red cadmium line (λ = 6438 Å), $\tau_c \sim 10^{-9}$ s; the corresponding coherence lengths are 3 and 30 cm, respectively. The finite value of the coherence time τ_c could be due to many factors; for example, if a radiating atom undergoes collision with another atom, then the wave train undergoes an abrupt phase shift of the type shown in Fig. 17.1(b). The finite coherence time could also be due to the random motion of atoms or

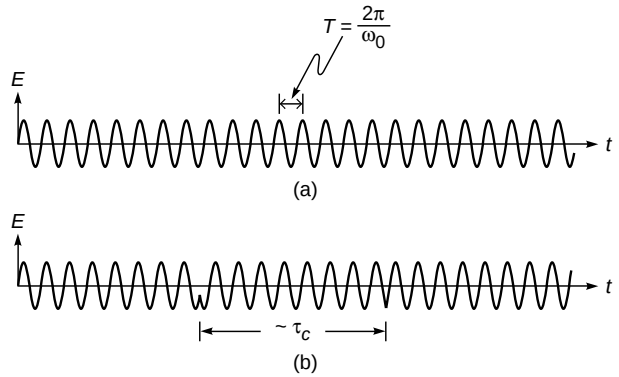


Fig. 17.1 (a) For a perfectly monochromatic beam, the displacement remains sinusoidal for $-\infty < t < +\infty$. (b) For an actual source, a definite phase relationship exists for times of the order of τ_c, which is known as the temporal coherence of the beam. For $\nu \sim 5 \times 10^{14}$ Hz and $\tau_c \sim 10^{-10}$ s, one has about 50,000 oscillations in time τ_c.

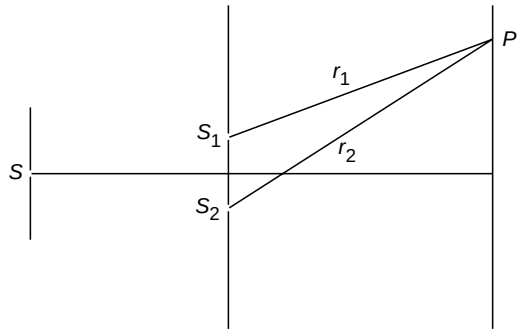


Fig. 17.2 Young's double-hole experiment. The interference pattern observed around point P at time t is due to the superposition of waves emanating from S_1 and S_2 at times $t - r_1/c$ and $t - r_2/c$, respectively; thus interference fringes of good contrast will be observed at P if $(r_2 - r_1)/c << \tau_c$.

due to the fact that an atom has a finite lifetime in the energy level from which it drops to the lower energy level while radiating.[1]

To understand the concept of coherence time (or of coherence length), we consider Young's double-hole experiment as shown in Fig. 17.2; the interference pattern produced by this experimental arrangement was discussed in considerable detail in Sec. 14.4. Now, the interference pattern observed around point P at time t is due to the superposition of waves emanating from S_1 and S_2 at times $t - r_1/c$, and $t - r_2/c$, respectively, where r_1 and r_2 are the distances S_1P, and S_2P, respectively. Obviously, if

$$\frac{r_2 - r_1}{c} \ll \tau_c$$

then the waves arriving at P from S_1 and S_2 will have a definite phase relationship, and an interference pattern of good contrast will be obtained. On the other hand, if the path difference $r_2 - r_1$ is large enough that

$$\frac{r_2 - r_1}{c} \gg \tau_c$$

then the waves arriving at P from S_1 and S_2 will have no fixed phase relationship, and no interference pattern will be observed. Thus the central fringe (for which $r_1 = r_2$) will, in general, have a good contrast, and as we move toward higher-order fringes, the contrast of the fringes will gradually become poorer. This point is discussed in greater detail in Sec. 17.7.

We next consider the Michelson interferometer experiment (see Sec. 15.10). A light beam falls on a beam splitter G (which is usually a partially silvered plate), and the waves reflected from mirrors M_1 and M_2 interfere (see Fig. 17.3). Let M_2'

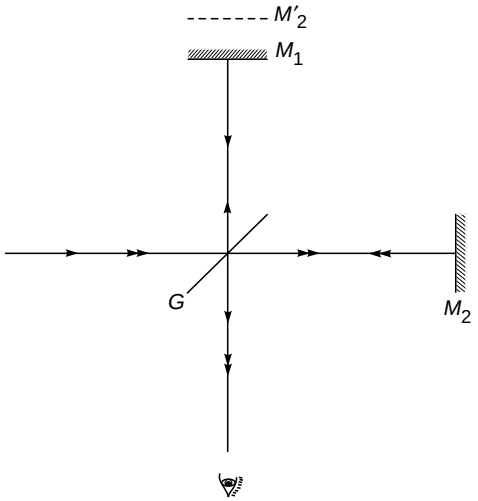


Fig. 17.3 The Michelson interferometer arrangement. G represents the beam splitter, and M_2' represents the image of M_2 as formed by G.

represent the image of mirror M_2 (formed by plate G) as seen by the eye. If distance M_1M_2' is denoted by d, then the beam which gets reflected by mirror M_2 travels an additional path equal to $2d$. Thus, the beam reflected from M_1 interferes with the beam reflected by M_2 which had originated $2d/c$ seconds earlier.

If the distance d is such that

$$\frac{2d}{c} << \tau_c$$

then a definite phase relationship exists between the two beams and well-defined interference fringes are observed. On the other hand, if

$$\frac{2d}{c} >> \tau_c$$

then, in general, there is no definite phase relationship between the two beams and no interference pattern is observed. There is no definite distance at which the interference pattern disappears; as the distance increases, the contrast of the fringes becomes gradually poorer and eventually the fringe system disappears. For the neon line ($\lambda = 6328$ Å), the disappearance occurs when the path difference is about a few centimeters giving $\tau_c \sim 10^{-10}$ s. On the other hand, for the red cadmium line ($\lambda = 6438$ Å), the coherence length is of the order of 30 cm, giving $\tau_c \sim 10^{-9}$ s.

The coherence time for a laser beam is usually much larger in comparison to ordinary light sources. Indeed, for a helium-neon laser, coherence times as large as 50 ms have been obtained (Ref. 9); this would imply a coherence length of 15,000 km! Commercially available helium-neon lasers have

[1] For more details, see Ref. 17.

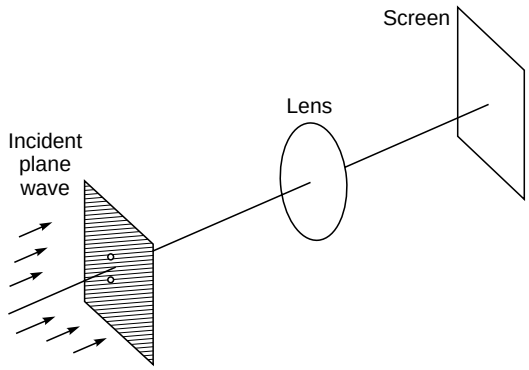


Fig. 17.4 A parallel beam of light is incident normally on a pair of circular holes, and the Fraunhofer diffraction pattern is observed on the focal plane of a convex lens.

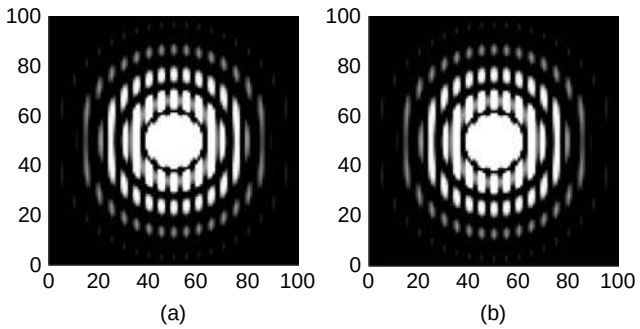


Fig. 17.5 (a) The interference pattern produced for the arrangement shown in Fig. 17.4 using a helium-neon laser beam. (b) The interference pattern produced by the same arrangement with 1 mm thick glass plate in front of one of the holes. (The above figures are computer-generated; the experimentally obtained photographs are very similar—see Ref. 16.)

$\tau_c \sim 50$ ns, implying coherence lengths of about 15 m. Thus using such a laser beam, high-contrast interference fringes can be obtained even for a path difference of a few meters.

To demonstrate the large coherence length of the laser beam, we consider an experimental arrangement shown in Fig. 17.4. A parallel beam of light is incident normally on a pair of circular holes. The Fraunhofer diffraction pattern is observed on the focal plane of a convex lens. We first use a helium-neon laser beam. The resulting interference pattern is shown in Fig. 17.5(a), which is simply the product of the Airy pattern and the interference pattern produced by two point sources (see Sec. 19.8). We next introduce a $\frac{1}{2}$ mm thick glass plate in front of one of the circular holes; there is almost no change in the interference pattern as can be seen from Fig. 17.5(b). Clearly, the extra path introduced by the plate [$= (\mu - 1)t$, see Sec. 14.10] is very small in comparison to the

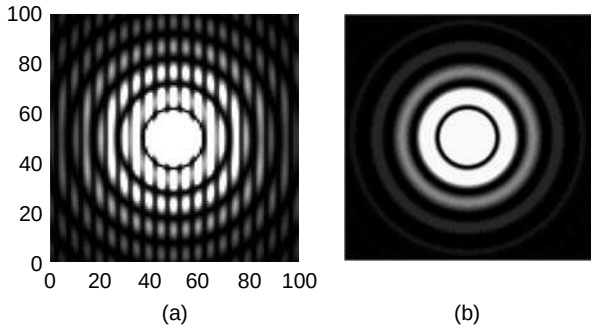


Fig. 17.6 (a) The interference pattern prodcued for the arrangement shown in Fig. 17.4 using a collimated mercury arc. (b) The interference pattern is washed out when 0.5 mm thick glass plate is introduced in front of one of the holes. (The above figures are computer-generated; the experimentally obtained photographs are very similar—see Ref. 16.)

coherence length associated with the laser beam. If we repeat the experiment with a collimated mercury arc beam, we find that with the introduction of the glass plate the interference pattern disappears (Fig. 17.6). This implies that the extra path length introduced by the glass plate is so large that there is no definite phase relationship between the waves arriving on the screen from the two circular apertures.

17.2 THE LINEWIDTH

In the Michelson interferometer experiment discussed in Sec. 17.5, the decrease in contrast of the fringes can also be interpreted as being due to the fact that the source is not emitting at a single frequency but over a narrow band of frequencies. When the path difference between the two interfering beams is zero or very small, the different wavelength components produce fringes superimposed on one another and the fringe contrast is good. On the other hand, when the path difference is increased, different wavelength components produce fringe patterns which are slightly displaced with respect to one another, and the fringe contrast becomes poorer. One can equally well say that the poor fringe visibility for a large optical path difference is due to the nonmonochromaticity of the light source.

The equivalence of the above two approaches can be easily understood if we consider the Michelson interferometer experiment using two closely spaced wavelengths λ_1 and λ_2. Indeed in Sec. 15.10 we showed that for two closely spaced wavelengths λ_1 and λ_2 (like the D_1 and D_2 lines of sodium),

the interference pattern will disappear if

$$\frac{2d}{\lambda_2} - \frac{2d}{\lambda_1} = \frac{1}{2} \tag{3}$$

where $2d$ represents the path difference between the two beams. Thus

$$2d = \frac{\lambda_1\lambda_2}{2(\lambda_1 - \lambda_2)} \simeq \frac{\lambda^2}{2(\lambda_1 - \lambda_2)} \tag{4}$$

Instead of two discrete wavelengths, if we assume that the beam consists of all wavelengths lying between λ and $\lambda + \Delta\lambda$, then the interference pattern produced by wavelengths λ and $\lambda + \frac{1}{2}\Delta\lambda$ will disappear if

$$2d = \frac{\lambda^2}{2\left(\frac{1}{2}\Delta\lambda\right)} = \frac{\lambda^2}{\Delta\lambda} \tag{5}$$

Further, for each wavelength lying between λ and $\lambda + \frac{1}{2}\Delta\lambda$, there will be a corresponding wavelength (lying between $\lambda + \frac{1}{2}\Delta\lambda$ and $\lambda + \Delta\lambda$) such that the minima of one fall on the maxima of the other, making the fringes disappear. Thus, for

$$2d \gtrsim \frac{\lambda^2}{\Delta\lambda} \tag{6}$$

the contrast of the interference fringes will be extremely poor. We may rewrite the above equation in the form

$$\Delta\lambda \gtrsim \frac{\lambda^2}{2d} \tag{7}$$

implying that if the contrast of the interference fringes becomes very poor when the path difference is $\sim d$, then the spectral width of the source will be $\sim \lambda^2/(2d)$.

Now, in Sec. 17.1 we observed that if the path difference exceeds the coherence length L, the fringes are not observed. From the above discussion it therefore follows that the spectral width of the source $\Delta\lambda$ is given by

$$\Delta\lambda \sim \frac{\lambda^2}{L} = \frac{\lambda^2}{c\tau_c} \tag{8}$$

Thus the temporal coherence τ_c of the beam is directly related to the spectral width $\Delta\lambda$. For example, for the red cadmium line, $\lambda = 6438$ Å, and $L \simeq 30$ cm ($\tau_c \simeq 10^{-9}$ s), giving

$$\Delta\lambda \sim \frac{\lambda^2}{c\tau_c} = \frac{(6.438 \times 10^{-5})^2}{3 \times 10^{10} \times 10^{-9}}$$

$$\sim 0.01 \text{ Å}$$

For the sodium line, $\lambda \simeq 5890$ Å, $L \simeq 3$ cm ($\tau_c \simeq 10^{-10}$ s), and $\Delta\lambda \sim 0.1$ Å. Further, since $\nu = c/\lambda$, the frequency spread $\Delta\nu$ of a line is

$$\Delta\nu \sim \frac{c}{\lambda^2}\Delta\lambda \sim \frac{c}{L} \tag{9}$$

where we have disregarded the sign. Since $\tau_c = L/c$, we obtain

$$\Delta\nu \sim \frac{1}{\tau_c} \tag{10}$$

Thus the frequency spread of a spectral line is of the order of the inverse of the coherence time. For example, for the yellow line of sodium ($\lambda = 5890$ Å),

$$\tau_c \sim 10^{-10} \text{ s} \quad \Rightarrow \quad \Delta\nu \sim 10^{10} \text{ Hz}$$

$$\nu = \frac{c}{\lambda} = \frac{3 \times 10^{10}}{5.89 \times 10^{-5}} \simeq 5 \times 10^{14} \text{ Hz}$$

we get

$$\frac{\Delta\nu}{\nu} \sim \frac{10^{10}}{5 \times 10^{14}} = 2 \times 10^{-5}$$

The quantity $\Delta\nu/\nu$ represents the monochromaticity (or the spectral purity) of the source, and one can see that even for an ordinary light source it is very small. For a commercially available laser beam, $\tau_c \sim 50$ ns, implying $\Delta\nu/\nu \sim 4 \times 10^{-8}$. The fact that the finite coherence time is directly related to the spectral width of the source can also be seen by using Fourier analysis; this is discussed in Sec. 17.6.

17.3 THE SPATIAL COHERENCE

Until now we have considered the coherence of two fields arriving at a particular point in space from a point source through two different optical paths. In this section we will discuss the coherence properties of the field associated with the finite dimension of the source.

We consider Young's double-hole experiment with point source S being equidistant from S_1 and S_2 [see Fig. 17.7(a)]. We assume S to be nearly monochromatic so that it produces interference fringes of good contrast on screen PP'. Point O on the screen is such that $S_1O = S_2O$. Clearly, point source S will produce an intensity maximum around point O. We next consider another similar source S' at a distance l from S. We assume that the waves from S and S' have no definite phase relationship. Thus the interference pattern observed on screen PP' will be a superposition of the intensity distributions of the interference patterns formed due to S and S' (see Sec. 14.6). If the separation l is slowly increased from zero, the contrast of the fringes on the screen PP' becomes poorer

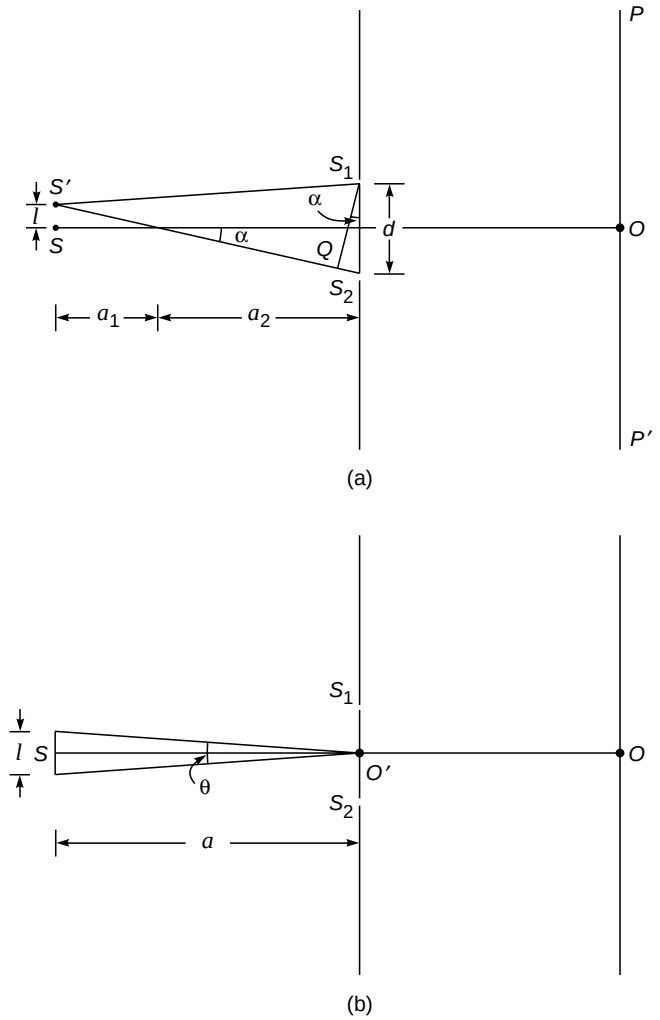


Fig. 17.7 (a) Young's double-hole interference experiment with two independent point sources S and S'. (b) The same experiment with an extended source.

because the interference pattern produced by S' is slightly shifted from that produced by S. Clearly, if

$$S'S_2 - S'S_1 = \frac{\lambda}{2} \qquad \textbf{(11)}$$

the minima of the interference pattern produced by S will fall on the maxima of the interference pattern produced by S′ and no fringe pattern will be observed. It can be easily seen that

$$S'S_2 = \left[a^2 + \left(\frac{d}{2} + l\right)^2\right]^{1/2} \simeq a + \frac{1}{2a}\left(\frac{d}{2} + l\right)^2$$

and

$$S'S_1 = \left[a^2 + \left(\frac{d}{2} - l\right)^2\right]^{1/2} \simeq a + \frac{1}{2a}\left(\frac{d}{2} - l\right)^2$$

where

$$a = a_1 + a_2$$

and we have assumed $a >> d, l$. Thus

$$S'S_2 - S'S_1 \simeq \frac{ld}{a}$$

Thus for the fringes to disappear, we must have

$$\frac{\lambda}{2} = S'S_2 - S'S_1 \simeq \frac{ld}{a}$$

or

$$l \simeq \frac{\lambda a}{2d}$$

Now, if we have an extended incoherent source whose linear dimension is $\sim \lambda a/d$, then for every point on the source, there is a point at a distance of $\lambda a/(2d)$ which produces fringes that are shifted by one-half of a fringe width. Therefore the interference pattern will not be observed. Thus for an extended incoherent source, interference fringes of good contrast will be observed only when

$$l << \frac{\lambda a}{d} \qquad \textbf{(12)}$$

Now, if θ is the angle subtended by the source at the slits [see Fig. 17.7(b)], then $\theta \simeq l/a$ and the above condition for obtaining fringes of good contrast takes the form

$$d << \frac{\lambda}{\theta} \qquad \textbf{(13)}$$

On the other hand, if

$$d \sim \frac{\lambda}{\theta} \qquad \textbf{(14)}$$

the fringes will be of very poor contrast. Indeed, a more rigorous diffraction theory tells us that the interference pattern disappears when (see, e.g., Sec. 5.5 of Ref. 7).

$$d = 1.22\frac{\lambda}{\theta}, 2.25\frac{\lambda}{\theta}, 3.24\frac{\lambda}{\theta}, \ldots \qquad \textbf{(15)}$$

Thus as the separation of the pinholes is increased from zero, the interference fringes disappear when $d = 1.22\lambda/\theta$; if d is further increased, the fringes reappear with relatively poor contrast and they are washed out again when $d = 2.25\lambda/\theta$, and so on. The distance

$$l_w = \lambda/\theta \qquad \text{lateral coherence width} \qquad \textbf{(16)}$$

gives the distance over which the beam may be assumed to be spatially coherent and is referred to as the lateral coherence width.

Example 17.1 On the surface of the Earth, the Sun subtends an angle of about 32 minutes. Assume that sunlight falls normally on a double-hole arrangement of the type shown in Fig. 17.7 and

that there is a filter in front of S_1S_2 so that light corresponding to $\lambda \approx 5000$ Å is incident on S_1S_2. What should be the separation between S_1 and S_2 so that fringes of good contrast are observed on the screen?

Solution:

$$\theta \approx 32' = \frac{32\pi}{180 \times 60} \text{ rad} \simeq 0.01 \text{ rad}$$

Thus the lateral coherence length

$$l_w \approx \frac{5 \times 10^{-5}}{10^{-2}} = 0.005 \text{ cm}$$

Therefore if the pinholes are separated by a distance which is small compared to 0.005 cm, interference fringes of good contrast should be observed.

17.4 MICHELSON STELLAR INTERFEROMETER

Using the concept of spatial coherence, Michelson developed an ingenious method for determining the angular diameter of stars. The method is based on the result that for a distant circular source, the interference fringes will disappear if the distance between pinholes S_1 and S_2 (see Fig. 17.8) is given by [see Eq. (15)]

$$d = 1.22 \frac{\lambda}{\theta} \qquad \textbf{(17)}$$

where θ is the angle subtended by the circular source as shown in Fig. 17.8. For a star whose angular diameter is 10^{-7} rad, the distance d for which the fringes will disappear is

$$d \sim \frac{1.22 \times 5 \times 10^{-5}}{10^{-7}} \simeq 600 \text{ cm}$$

where we have assumed $\lambda \simeq 5000$ Å. Obviously, for such a large value of d, the fringe width will become extremely small. Further, one has to use a big lens, which not only is difficult to make, but only a small portion of which will be used. To overcome this difficulty, Michelson used two movable mirrors M_1 and M_2 as shown in Fig. 17.9, and thus he effectively got a large value of d. The apparatus is known as Michelson's stellar interferometer. In a typical experiment the first disappearance occurred when the distance M_1M_2 was about 24 ft, which gave

$$\theta \simeq \frac{1.22 \times 5 \times 10^{-5}}{24 \times 12 \times 2.54} \text{ rad} \simeq 0.02 \text{ second}$$

for the angular diameter of the star. This star is known as Arctures. From the known distance of the star, one can estimate that the diameter of the star is about 27 times that of the Sun.

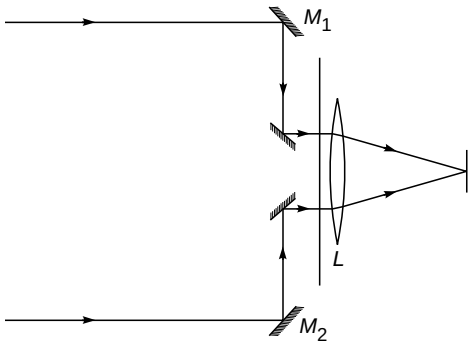


Fig. 17.9 Michelson's stellar interferometer.

Note that a laser beam is spatially coherent across the entire beam. Thus, if a laser beam is allowed to fall directly on a double-slit arrangement (see Fig. 17.10), then as long as the beam falls on both the slits, a clear interference pattern is observed on the screen. This shows that the laser beam is spatially coherent across the entire wave front.

Figure 17.11 shows the interference pattern obtained by Nelson and Collins (Ref. 14) by placing a pair of slits of width 7.5 μm separated by a distance 54.1 μm on the end of the ruby rod in a ruby laser. The interference pattern agrees with

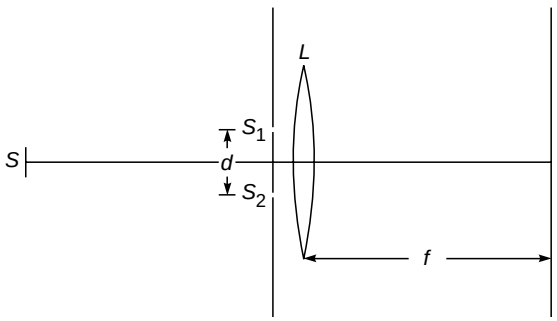


Fig. 17.8 S is a source of certain spatial extent; S_1 and S_2 are two slits separated by a distance d which can be varied. The fringes are observed on the focal plane of lens L.

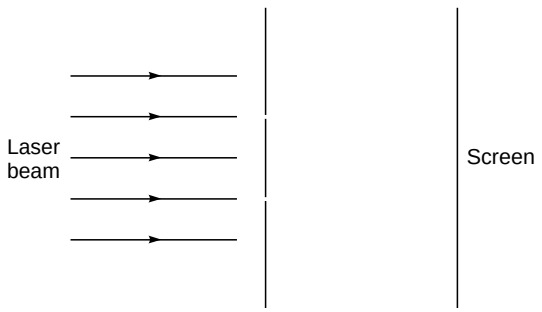


Fig. 17.10 If a laser beam falls on a double-slit arrangement, interference fringes are observed on the screen. This shows that the laser beam is spatially coherent across the entire wave front.

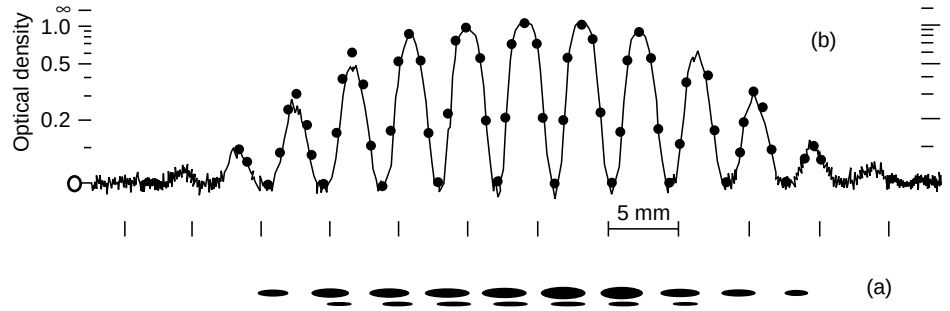


Fig. 17.11 The double-slit interference pattern obtained by placing a pair of slits each 7.5 μm wide and separated by a distance of 54.1 μm across the diameter of a ruby rod. (a) The actual interference pattern and (b) a densitometer trace of the interference pattern. The dots correspond to a theoretical calculation assuming that a plane wave strikes the pair of slits. (After Ref. 14. Photograph courtesy: Dr. D. F. Nelson.)

the theoretical calculation to within 20%. To show that the spatial coherence is indeed due to laser action, they showed that below threshold (of the laser) no regular interference pattern was observed; only a uniform darkening of the photographic plate was obtained.

17.5 OPTICAL BEATS

When two tuning forks, one having a frequency of 256 Hz and the other a frequency of 260 Hz, are made to vibrate at the same time, we hear a frequency of about 258 Hz whose intensity varies from zero to maximum and back with a frequency of 4 Hz. This phenomenon is known as *beats*. It can be easily understood by considering the superposition of two waves having frequencies ω and ω + Δω:

$$y_1 = a \sin(\omega t + \phi_1)$$
$$y_2 = a \sin[(\omega + \Delta\omega)t + \phi_2] \qquad \textbf{(18)}$$

where we are assuming (for the sake of simplicity) that both the waves have the same amplitude. The resultant displacement is given by

$$\begin{aligned} y &= y_1 + y_2 \\ &= 2a \sin\left[\left(\omega + \tfrac{1}{2}\Delta\omega\right)t + \frac{1}{2}(\phi_1 + \phi_2)\right] \\ &\quad \times \cos\left[\frac{1}{2}(\Delta\omega)t + \frac{1}{2}(\phi_2 - \phi_1)\right] \\ &= 2a \sin\left[\left(\omega + \frac{1}{2}\Delta\omega\right)t\right] \sin\left(\frac{1}{2}\,\Delta\omega\, t\right) \end{aligned} \qquad \textbf{(19)}$$

where we have assumed, without any loss of generality, $\phi_1 = \pi/2 = -\phi_2$. Figure 17.12(a) and (b) shows the time variation of the terms

$$\sin\left(\omega + \frac{1}{2}\Delta\omega\right)t \qquad \text{and} \qquad \sin\left(\frac{1}{2}\Delta\omega\right)t$$

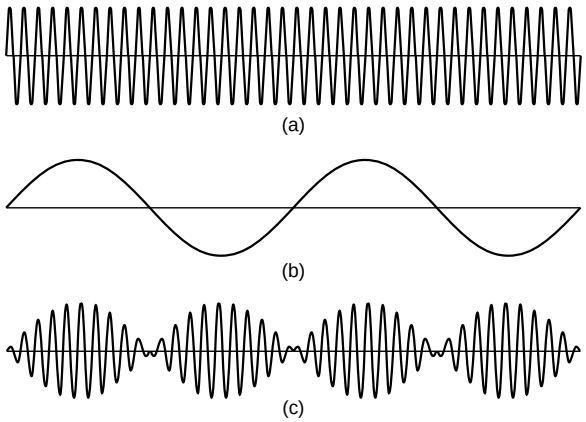


Fig. 17.12 (a), (b) Typical time variation of $\sin\left[\left(\omega + \frac{1}{2}\Delta\omega\right)t\right]$ and $\sin\left(\frac{1}{2}\Delta\omega\, t\right)$, respectively; (c) time dependence of the product.

respectively. In Fig. 17.12(c), we have plotted their product which represents the resultant displacement. Notice that although the envelope has a frequency of $\Delta\omega/4\pi\left(=\frac{1}{2}\Delta\nu\right)$ [see Fig. 17.12(b)], the intensity repeats itself after every $1/\Delta\nu$ seconds. This waxing and waning of sound is known as *beats*.

The beat phenomenon can be easily understood by observing the Moiré fringes obtained by the overlapping of two patterns of slightly different spatial frequency (see Fig. 17.13). Whenever the dark line of one of the patterns falls on the bright region of the other, the two waves can be considered to be "out of phase" and we have a broad dark region which appears periodically.

In a similar manner, one can consider the phenomenon of optical beats. For example, let us consider the superposition of two fields E_1 and E_2 having frequencies ω and ω + Δω:

$$E_1 = E_{01} \sin(\omega t + \phi_1) \qquad \textbf{(20)}$$

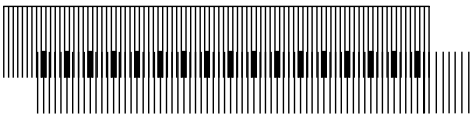


Fig. 17.13 The Moiré pattern produced by the overlapping of two patterns of parallel lines (of slightly different spatial periods) show the beating phenomenon [after Ref. 1].

and

$$E_2 = E_{02} \sin [(\omega + \Delta\omega)t + \phi_2] \qquad \textbf{(21)}$$

If we assume that both fields are linearly polarized in the same direction, then to calculate the resultant field, we simply algebraically add E_1 and E_2. Thus the resultant is

$$\begin{aligned} E &= E_1 + E_2 \\ &= E_{01} \sin (\omega t + \phi_1) + E_{02} \sin [(\omega + \Delta\omega)t + \phi_2] \end{aligned}$$

Now

$$\begin{aligned} E^2(t) = {} & E_{01}^2 \sin^2 (\omega t + \phi_1) + E_{02}^2 \sin^2 [(\omega + \Delta\omega)t + \phi_2] \\ & + E_{01}E_{02}[-\cos (2\omega t + \Delta\omega t + \phi_1 + \phi_2) \\ & + \cos (\Delta\omega t + \phi_2 - \phi_1)] \end{aligned} \qquad \textbf{(22)}$$

For optical frequencies, $\omega \approx 10^{15}$ Hz and therefore the first three terms will vary with extreme rapidity and a detector (such as the eye or the photodetector) will observe a time average of the quantity. Now, the time average of the quantity $F(t)$ over a duration of $2T$ is defined through the following equation:

$$\langle F(t) \rangle = \frac{1}{2T} \int_{-T}^{+T} F(t)\, dt \qquad \textbf{(23)}$$

Thus

$$\begin{aligned} &\langle E_{01}^2 \sin^2 (\omega t + \phi_1) \rangle \\ &\quad = E_{01}^2 \frac{1}{2T} \int_{-T}^{+T} \sin^2 (\omega t + \phi_1)\, dt \\ &\quad = E_{01}^2 \left\{ \frac{1}{2} - \frac{1}{2\omega T} \left[\sin 2(\omega t + \phi_1) \right]_{-T}^{+T} \right\} \\ &\quad = \frac{1}{2} E_{01}^2 \left(1 - \frac{1}{2\omega T} \sin 2\omega T \cos 2\phi_1 \right) \end{aligned} \qquad \textbf{(24)}$$

For averaging times $T >> 1/\omega$, the second term inside the curly braces will be extremely small and hence can be neglected. Thus we may write

$$\langle E_{01}^2 \sin^2 (\omega t + \phi_1) \rangle \simeq \tfrac{1}{2} E_{01}^2 \qquad \textbf{(25)}$$

For example, the eye would respond to changes in times of the order of 0.05 s. Thus $T \sim 0.05$ s and since $\omega \simeq 10^{15}$ Hz, we have

$$\frac{1}{\omega T} \simeq 2 \times 10^{-14}$$

which is an extremely small quantity in comparison to unity. It is for this reason that the eye does not see any intensity variations. Even for a fast photodetector with response times $\sim 10^{-9}$ s, $1/(\omega T) \sim 10^{-6}$ which can also be neglected.

Returning to Eq. (22), if we carry out an averaging over times which are long compared to $2\pi/\omega$ but short compared to $2\pi/\Delta\omega$, then we obtain

$$\begin{aligned} \langle E^2(t) \rangle = {} & \tfrac{1}{2} E_{01}^2 + \tfrac{1}{2} E_{02}^2 \\ & + E_{01} E_{02} \cos [(\Delta\omega)t + \phi_2 - \phi_1] \end{aligned} \qquad \textbf{(26)}$$

For example, if $\Delta\omega \simeq 10^7$ Hz and the photodetector resolution is about 10^{-9} s, then the detector will record only the average values of the first three terms on the RHS of Eq. (22); however, it will be able to record the time variation of the last term. This is what is shown in the above equation, leading to the familiar phenomenon of beats.

As an example, we consider the beating of the D_1 and D_2 lines of sodium for which

$$\lambda_1 = 5890 \text{ Å } (\Rightarrow \omega_1 \simeq 3.2003 \times 10^{15} \text{ Hz})$$
$$\lambda_2 = 5896 \text{ Å } (\Rightarrow \omega_2 \simeq 3.1970 \times 10^{15} \text{ Hz})$$

Thus $\Delta\omega \simeq 3.3 \times 10^{12}$ Hz

To observe the beating, the detector should have a response time much smaller than $1/\Delta\omega$; thus the photodetector response time should be $\lesssim 10^{-13}$ s which is a practical impossibility. Therefore, to observe the beats, we must decrease the value of $\Delta\omega$. Indeed the first experiment on optical beats was carried out by Forrester et al. (Ref. 6) in which they used two closely spaced frequencies by splitting a spectral line using a magnetic field (this splitting is known as the Zeeman effect). The weaker the magnetic field, the smaller is the value of $\Delta\omega$. In the experiment of Forrester and his coworkers, $\Delta\nu$ was of the order of 10^{10} Hz, and they were able to observe optical beats.

Obviously, for the beats to occur very slowly (so that we may use photodetectors of much longer response times), $\Delta\omega$ should be made even smaller—but then we may have the coherence problem. In the above analysis we assumed the phases ϕ_1 and ϕ_2 to remain constant in time. Now for an incoherent source, ϕ_1 and ϕ_2 will randomly change in times $\sim 10^{-9}$ s; thus if the detector response time is $\gtrsim 10^{-8}$ s, we will observe the average of the cos $[(\Delta\omega)t + \phi_2 - \phi_1]$ term in Eq. (26). Obviously, the average value of the cosine term is zero, and we will have

$$\langle E^2(t) \rangle = \frac{1}{2} E_{01}^2 + \frac{1}{2} E_{02}^2$$

implying that the resultant intensity will be just the sum of the independent intensities:

$$I = I_1 + I_2 \qquad \textbf{(27)}$$

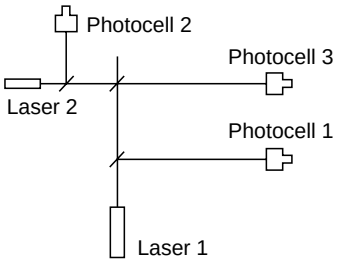


Fig. 17.14 The experimental arrangement of Lipsett and Mandel (Ref. 11) to observe optical beats using two laser beams.

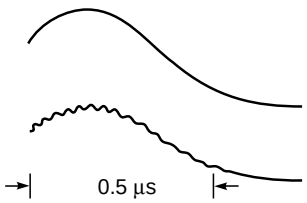


Fig. 17.15 Oscilloscope trace of the sum of the intensities of the laser beams (upper curve) and the intensity of the superposed laser beam (lower curve) (Ref. 11).

With the advent of laser beams, the beating experiments have become much easier; a typical arrangement (which resembles a Michelson interferometer) is shown in Fig. 17.14. A typical beat note of the experiment of Lipsett and Mandel (Ref. 11) is shown in Fig. 17.15. It was observed that the beat note changed in frequency from about 33 to approximately 21 MHz in about 0.7 μs. The coherence time is ~0.5 μs which is consistent with the duration of the spike.

We conclude this section by quoting Feynman: "With the availability of laser sources, someone will be able to demonstrate two sources shining on a wall, in which the beats are so slow that one can see the wall get bright and dark."

17.6 COHERENCE TIME AND LINEWIDTH VIA FOURIER ANALYSIS

That the frequency spread of a line is of the order of the inverse of the coherence time [see Eq. (10)] can also be shown by Fourier analysis. As an example, we consider a sinusoidal displacement of duration τ_c. Thus we may write

$$\psi(x = 0, t) = \begin{cases} ae^{i\omega_0 t} & |t| < \frac{1}{2}\tau_c \\ 0 & |t| > \frac{1}{2}\tau_c \end{cases} \qquad \textbf{(28)}$$

We will assume that τ_c is long enough that the disturbance consists of many oscillations. For example, for a 2 ns pulse corresponding to $\nu_0 \simeq 5 \times 10^{14}$ Hz, the number of oscillations will be $5 \times 10^{14} \times 2 \times 10^{-9} = 10^6$; i.e., the pulse will consist of about 1 million oscillations!

Now, while discussing the Fourier transform theory (see Secs. 8.4 and 9.5), we showed that for a time-dependent function $f(t)$, if we define

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \qquad \textbf{(29)}$$

then

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\omega) e^{i\omega t} d\omega \qquad \textbf{(30)}$$

Replacing $f(t)$ by $\psi(x = 0, t)$, we have

$$\psi(x = 0, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} A(\omega) e^{i\omega t} d\omega \qquad \textbf{(31)}$$

The RHS represents a superposition of plane waves with $A(\omega)$ representing the amplitude[2] of the plane wave

[2] Notice that the integral appearing on the RHS of Eq. (30) is over negative values of ω also. However, the displacement (or the electric field) is the real part of ψ which is given by (omitting the $\sqrt{2\pi}$ factor)

$$E = \text{Re}\,[\psi(x = 0, t)] = \text{Re}\left(\int_{-\infty}^{\infty} |A(\omega)| e^{i(\omega t + \phi)}\, d\omega \right)$$

$$= \int_{-\infty}^{\infty} |A(\omega)| \cos\,(\omega t + \phi)\, d\omega = \int_{0}^{\infty} |A(\omega)| \cos\,(\omega t + \phi)\, d\omega + \int_{0}^{\infty} |A(-\omega)| \cos[\omega t - \phi(-\omega)]\, d\omega$$

where we have used the relation $A(\omega) = |A(\omega)|\, e^{i\phi}$. The above equation can always be written in the form

$$\int_{0}^{\infty} C(\omega) \cos[\omega t + \theta(\omega)]\, d\omega$$

Thus the amplitudes associated with the negative frequencies contribute essentially to the corresponding positive frequencies.

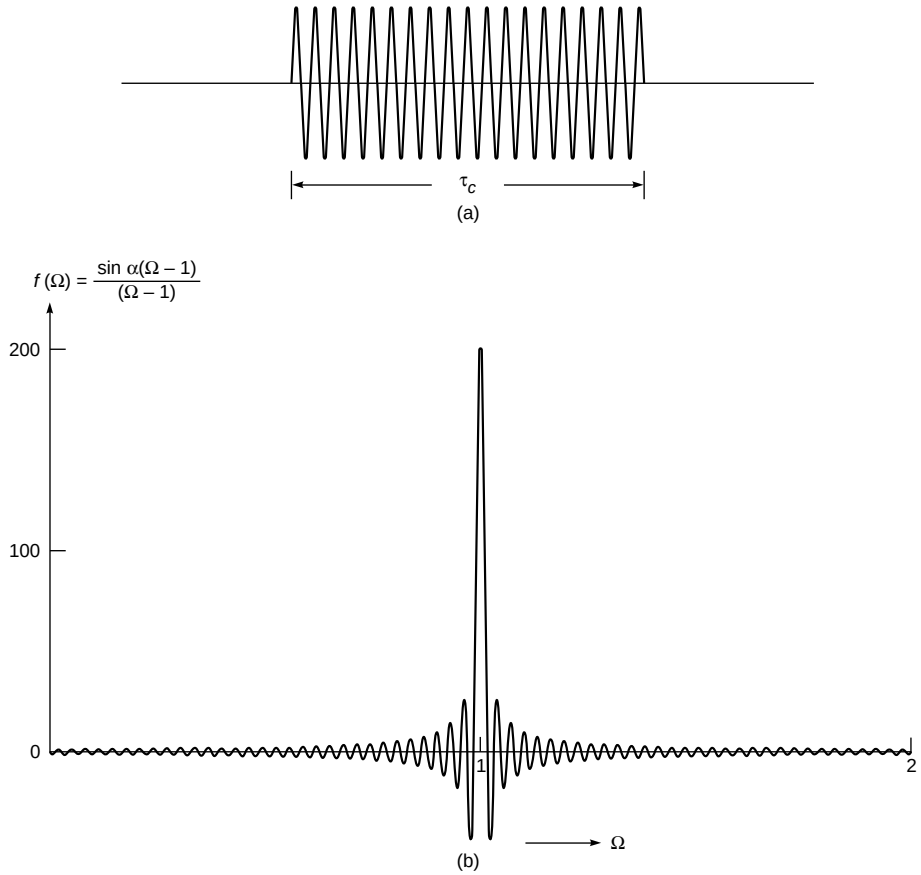


Fig. 17.16 (a) A sinusoidal displacement of duration τ_c. (b) The variation of the function $[\sin(\Omega - 1)\alpha]/(\Omega - 1)$ as a function of Ω for $\alpha = 200$. Notice that the function is sharply peaked around $\Omega = 1$.

corresponding to the frequency ω. Equation (31) tells us that $\psi(x = 0, t)$ is the Fourier transform of $A(\omega)$, and therefore using the inverse Fourier transform [see Eq. (29)], we get

$$A(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \psi(x = 0, t) e^{-i\omega t}\, dt$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\frac{1}{2}\tau_c}^{+\frac{1}{2}\tau_c} a e^{i(\omega_0 - \omega)t}\, dt$$

$$= \left(\frac{2}{\pi}\right)^{1/2} a \left\{\frac{\sin\left[\frac{1}{2}(\omega - \omega_0)\tau_c\right]}{\omega - \omega_0}\right\}$$

$$= \left(\frac{2}{\pi}\right)^{1/2} \frac{a}{\omega_0} \left\{\frac{\sin[\alpha(\Omega - 1)]}{\Omega - 1}\right\} \qquad \textbf{(32)}$$

where $\Omega \equiv \omega / \omega_0$ and $\alpha = \frac{1}{2}\omega_0 \tau_c$. In Fig. 17.16 we have plotted the function

$$\frac{\sin[\alpha(\Omega - 1)]}{\Omega - 1} \qquad \textbf{(33)}$$

as a function of Ω for $\alpha = 200$. One can see that the function is sharply peaked at $\Omega = 1$ (where it has a value equal to α) and that the first zero on either side occurs at $\Omega = 1 \pm \pi/\alpha$. For larger values of α the function will become more sharply peaked; the width of the peak is given by

$$\Delta\Omega \left(= \frac{\Delta\omega}{\omega_0}\right) \sim \frac{\pi}{\alpha} \qquad \textbf{(34)}$$

or

$$\Delta\omega \sim \frac{\pi\omega_0}{\alpha} \sim \frac{2\pi}{\tau_c}$$

Thus

$$\Delta\nu \sim \frac{1}{\tau_c} \qquad \textbf{(35)}$$

consistent with Eq. (10). The above analysis shows that a wave having a coherence time $\sim\tau_c$ is essentially a superposition of harmonic waves having frequencies in the region $\nu_0 - \frac{1}{2}\Delta\nu \lesssim \nu \lesssim \nu_0 + \frac{1}{2}\Delta\nu$, where $\Delta\nu \sim 1/\tau_c$.

The condition expressed by Eq. (35) is quite general in the sense that it is valid for a pulse of arbitrary shape. For example, for a Gaussian pulse having a duration $\sim\tau_c$, the corresponding frequency spread will again be given by Eq. (35) (see Example 10.4).

17.7 COMPLEX DEGREE OF COHERENCE AND FRINGE VISIBILITY IN YOUNG'S DOUBLE-HOLE EXPERIMENT

In this section we will introduce the complex degree of coherence and will show how it can be related to the contrast of the fringes in Young's double-hole interference experiment. We refer to Fig. 17.2. Let $\Psi_1(P, t)$ and $\Psi_2(P, t)$ represent the complex fields at point P due to the waves emanating from S_1 and S_2, respectively. The resultant displacement is given by

$$\Psi = \Psi_1(P, t) + \Psi_2(P, t) \tag{36}$$

Now, the intensity at point P will be proportional to $|\Psi|^2$ which is given by

$$\begin{aligned} |\Psi|^2 &= \Psi_1^*\Psi_1 + \Psi_2^*\Psi_2 + \Psi_1^*\Psi_2 + \Psi_1\Psi_2^* \\ &= |\Psi_1|^2 + |\Psi_2|^2 + 2\,\mathrm{Re}(\Psi_1^*\Psi_2) \end{aligned}$$

Since Ψ_1 and Ψ_2 vary with extreme rapidity, we can observe only the average values of $|\Psi_1|^2$ and $|\Psi_2|^2$. Thus, if we write

$$I_1 = \langle|\Psi_1(P, t)|^2\rangle$$

and

$$I_2 = \langle|\Psi_2(P, t)|^2\rangle$$

then

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\ \mathrm{Re}\,\gamma_{12} \tag{37}$$

where

$$\gamma_{12} = \frac{\langle\Psi_1^*(P, t)\Psi_2(P, t)\rangle}{[\langle|\Psi_1(P, t)|^2\rangle\langle|\Psi_2(P, t)|^2\rangle]^{1/2}} \tag{38}$$

is known as the complex degree of coherence and $\langle\ldots\rangle$ denotes the time average of the quantity inside the triangular brackets [see Eq. (23)]. The field $\Psi_1(P, t)$ is due to the waves emanating from point S_1 at $t - r_1/c$, where $r_1 = S_1P$. Thus, $\Psi_1(P, t)$ will be proportional to $\Psi(S_1, t - r_1/c)$ where $\Psi(S_1, t)$ denotes the field at S_1 at time t. Similarly $\Psi_2(P, t)$ will be proportional to $(S_2, t - r_2/c)$. Thus

$$\gamma_{12} = \frac{\langle\Psi^*(S_1, t - r_1/c)\Psi(S_2, t - r_2/c)\rangle}{[\langle|\Psi(S_1, t - r_1/c)|^2\rangle\langle|\Psi\left(S_2, t - r_2/c\right)|^2\rangle]^{1/2}}$$

Since the overall intensity distribution in the fringe pattern does not change with time, we may write

$$\gamma_{12} = \frac{\langle\Psi^*(S_1, t + \tau)\Psi(S_2, t)\rangle}{[\langle|\Psi^*(S_1, t)|^2\rangle\langle|\Psi(S_2, t)|^2\rangle]^{1/2}} \tag{39}$$

where $\tau = (r_2 - r_1)/c$. To discuss the effect of temporal coherence, we assume S, S_1, and S_2 to be of negligible spatial dimensions. Further, if S_1, and S_2 are equidistant from S, then we may assume that

$$\Psi(S_1, t) = \Psi(S_2, t) = \Psi(t) \tag{40}$$

Thus, for such a case,

$$\gamma_{12}(\tau) = \frac{\langle\Psi^*(t + \tau)\Psi(t)\rangle}{\langle|\Psi(t)|^2\rangle} \tag{41}$$

Now, for an actual field we may write

$$\Psi(t) = A(t)e^{-i[\omega t + \phi(t)]} \tag{42}$$

where $A(t)$ and $\phi(t)$ are slowly varying real functions of time. For a perfectly monochromatic beam (i.e., infinite coherence time) $A(t)$ and $\phi(t)$ are constants so that

$$\Psi^*(t + \tau)\Psi(t) = A^2e^{i\omega\tau}$$

Consequently

$$\gamma_{12}(\tau) = e^{i\omega\tau} \tag{43}$$

Thus, for such a case

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\omega\tau \tag{44}$$

and the visibility V, which is defined by

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{45}$$

is given by

$$V = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2} \tag{46}$$

For $I_1 = I_2$ we have $V = 1$, implying that, for a perfectly monochromatic beam, the contrast of the fringes is perfect. On the other hand, for an ordinary light source having $\tau_c \sim 10^{-10}$ s, the functions $A(t)$ and $\phi(t)$ can be assumed to be constants in times $\lesssim 10^{-10}$ s. Thus, if $\tau_c \gtrsim 10^{-10}$ s, $\Psi(t + \tau)$

will have no phase relationship with $\Psi(t)$ and the time average $\langle \Psi^*(t+\tau)\Psi(t)\rangle$ will be zero. Thus, if the path difference $S_2P \sim S_1P$ is such that

$$\frac{S_2P \sim S_1P}{c} \gtrsim \tau_c \qquad \textbf{(47)}$$

the fringe pattern will not be observed.

In general, we may write

$$\gamma_{12} = |\gamma_{12}| e^{i(\omega\tau+\beta)} \qquad \textbf{(48)}$$

where $|\gamma_{12}|$ and β may be assumed to be constants around the observation point. This gives us

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\,|\gamma_{12}|\cos\alpha \qquad \textbf{(49)}$$

where $\alpha = \omega\tau + \beta$. Thus

$$I_{\max} = I_1 + I_2 + 2\sqrt{I_1 I_2}\,|\gamma_{12}| \qquad \textbf{(50)}$$

and

$$I_{\min} = I_1 + I_2 - 2\sqrt{I_1 I_2}\,|\gamma_{12}| \qquad \textbf{(51)}$$

Hence the visibility becomes

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}|\gamma_{12}| \qquad \textbf{(52)}$$

Thus the visibility (or the contrast) of the fringes is a direct measure of $|\gamma_{12}|$. If $I_1 = I_2$, then $V = |\gamma_{12}|$. In the present case, since S, S_1, and S_2 have been assumed to be points, $|\gamma_{12}|$ depends only on the temporal coherence of the beam. For $\tau << \tau_c$, $|\gamma_{12}|$ is very close to unity and the contrast of the fringes will be very good; for $\tau >> \tau_c$, $|\gamma_{12}|$ will be close to zero and the contrast will be extremely poor.

Note from Eq. (43) that for a perfectly monochromatic beam $|\gamma_{12}| = 1$ and $\alpha = \omega\tau = \omega(S_2P \sim S_1P)/c$. In general, it can be shown that $0 < |\gamma_{12}| < 1$; $|\gamma_{12}| = 0$ implies complete incoherence and $|\gamma_{12}| = 1$ implies complete coherence. In practice, if $|\gamma_{12}| > 0.88$, the light is said to be "almost coherent." Further, since

$$\langle \Psi^*(t+\tau)\Psi(t)\rangle = e^{i\omega\tau}\langle A(t+\tau)A(t)e^{i[\phi(t+\tau)-\phi(t)]}\rangle$$

and for a nearly monochromatic source $A(t)$ and $\phi(t)$ are already slowly varying functions of time, the quantity inside the angular brackets (on the RHS of the above equation) will not vary rapidly with τ. Thus, we may write

$$\gamma_{12} = |\gamma_{12}| e^{i\beta} e^{i\omega\tau} \qquad \textbf{(53)}$$

where both $|\gamma_{12}|$ and β are slowly varying functions of

$$\tau = \frac{S_2P \sim S_1P}{c} \qquad \textbf{(54)}$$

For a more detailed theory of spatial and temporal coherence, see Refs 2, 3, 7, and 20.

17.8 FOURIER TRANSFORM SPECTROSCOPY[3]

In Sec. 17.7, we have showed that the contrast in an interference pattern depends on the relative magnitudes of the optical path difference Δ, vis à vis the coherence length of the source L_c (= $c\tau_c$). For a given source, the contrast varies as the optical path difference Δ is varied, beginning from an extremely good contrast for $\Delta << L_c$ to a very poor contrast for $\Delta >> L_c$. Indeed Fizeau in 1862 interpreted the periodic variation in contrast in Newton's rings, under illumination with a sodium lamp as the lens is moved up, as being due to the presence of two lines separated by 6 Å (see Example 15.4). Michelson in the years 1890 to 1900 performed various experiments with a number of spectral lines. Using the Michelson interferometer, he measured visibility as a function of optical path difference; and using a mechanical device he himself had built, he could obtain the spectra. It is the purpose of this section to show that from a knowledge of variation of intensity with optical path difference one can obtain the source spectral distribution by a Fourier transformation.

The use of the Michelson interferometer for spectroscopy was revived in the 1950s for application, especially for the relatively complex spectra in the infrared region.

We will derive expressions for the variation of visibility with optical path difference for a source having a certain spectral distribution, and we will show that from the interference pattern one can obtain the spectral intensity distribution of the given source.

17.8.1 Principle of Fourier Transform Spectroscopy

Figure 17.17 shows the arrangement used in a Fourier transform spectrometer. Light from the given source is collimated

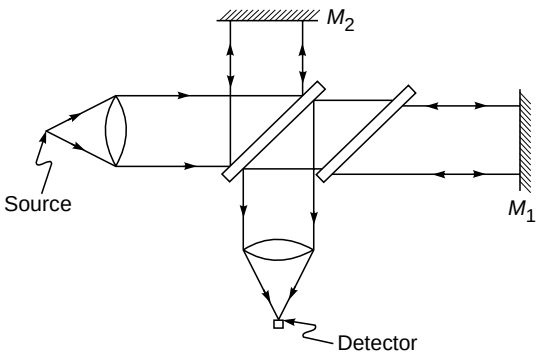


Fig. 17.17 The arrangement used in a Fourier transform spectrometer.

[3] This section was kindly written by Prof. K. Thyagarajan.

and enters the Michelson interferometer, and in the transmitted arm we measure the intensity at the focus of the lens as a function of the path difference Δ. Now, if a monochromatic beam of intensity I_0 is split into two beams (each of intensity $\frac{1}{2} I_0$) and are made to interfere, then the resultant intensity is given by

$$I = I_0(1 + \cos \delta) \tag{55}$$

where

$$\delta = \frac{2\pi}{\lambda}\Delta = \frac{2\pi\nu}{c}\Delta \tag{56}$$

represents the phase difference between the interfering beams, and in writing Eq. (55), we used Eq. (30) of Chap. 14 with

$$I_1 = I_2 = \tfrac{1}{2} I_0$$

Thus if $I(\nu)\, d\nu$ represents the intensity emitted by the source between ν and $\nu + d\nu$, then the intensity at O lying between ν and $\nu + d\nu$ is given by

$$I_t(\nu)\, d\nu = I(\nu)\, d\nu \left(1 + \cos\frac{2\pi\nu\Delta}{c}\right) \tag{57}$$

Hence, the total intensity at O corresponding to a path difference Δ is

$$I_t(\Delta) = \int_0^\infty I_t(\nu)\, d\nu$$

$$= \int_0^\infty I(\nu)\, d\nu + \int_0^\infty I(\nu) \cos\frac{2\pi\nu\Delta}{c} d\nu \tag{58}$$

The quantity

$$I_T = \int_0^\infty I(\nu)\, d\nu = \frac{1}{2} I_t(0) \tag{59}$$

represents the total intensity of the source. We define normalized transmission as

$$\gamma(\Delta) = \frac{I_t(\Delta) - I_T}{I_T}$$

$$= \frac{1}{I_T}\int_0^\infty I(\nu) \cos\frac{2\pi\nu\Delta}{c} d\nu \tag{60}$$

It is the quantity $I_t(\Delta)$ which is measured as a function of Δ from which $\gamma(\Delta)$ is evaluated. We first consider some examples giving explicit expressions for $I_t(\Delta)$ and $\gamma(\Delta)$ for some specific cases.

a. Monochromatic Source For a monochromatic source of intensity I_0 emitting at a frequency ν_0, we have

$$I(\nu)\, d\nu = I_0\, \delta(\nu - \nu_0)\, d\nu \tag{61}$$

where $\delta(\nu - \nu_0)$ represents the Dirac delta function. Hence

$$\gamma(\Delta) = I_0 \frac{\int_0^\infty \delta(\nu - \nu_0) \cos(2\pi\nu\Delta/c)\, d\nu}{\int_0^\infty \delta(\nu - \nu_0)\, d\nu}$$

$$= \cos\frac{2\pi\nu_0\Delta}{c} \tag{62}$$

and

$$I_t(\Delta) = I_0 \left(1 + \cos\frac{2\pi\nu_0\Delta}{c}\right) \tag{63}$$

Hence $I_t(\Delta)$ and γ vary sinusoidally for all values of path difference Δ [see Fig. 17.18(a) and (b)], implying that the coherence length of the source is infinite.

b. Source Emitting Two Monochromatic Lines We now consider a source emitting two monochromatic lines at frequencies ν_1 and ν_2, each characterized by an intensity $\frac{1}{2} I_0$. Thus

$$I(\nu)\, d\nu = \frac{1}{2} I_0[\delta(\nu - \nu_1) + \delta(\nu - \nu_2)] \tag{64}$$

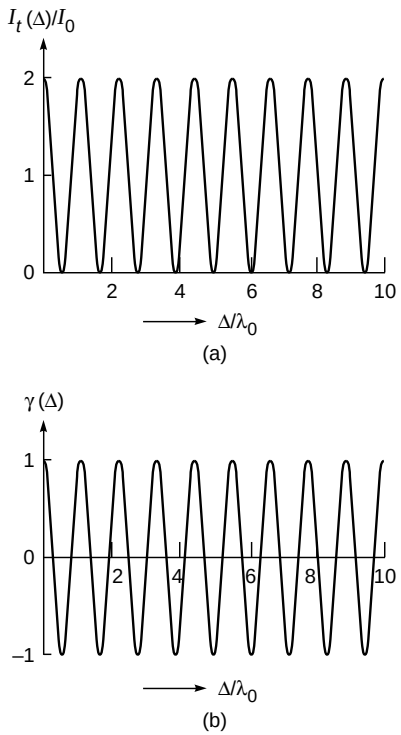


Fig. 17.18 (a) The variation of the total intensity at O as a function of the path difference Δ for a monochromatic source. (b) The corresponding cosinusoidal variation of $\gamma(\Delta)$ with Δ.

and

$$\gamma(\Delta) = \frac{1}{2}\left[\int_0^\infty \delta(\nu-\nu_1)\cos\frac{2\pi\nu\Delta}{c}d\nu + \int_0^\infty \delta(\nu-\nu_2)\cos\frac{2\pi\nu\Delta}{c}d\nu\right]$$

$$= \frac{1}{2}\left(\cos\frac{2\pi\nu_1\Delta}{c} + \cos\frac{2\pi\nu_2\Delta}{c}\right)$$

$$= \cos\left(2\pi\frac{\nu_1+\nu_2}{2c}\Delta\right) \times \cos\left(2\pi\frac{\nu_1-\nu_2}{2c}\Delta\right) \quad \textbf{(65)}$$

and

$$I_t(\Delta) = I_0\left[1+\cos\left(2\pi\frac{\nu_1-\nu_2}{2c}\Delta\right) \times \cos\left(2\pi\frac{\nu_1+\nu_2}{2c}\Delta\right)\right] \quad \textbf{(66)}$$

Such a variation of $I_t(\Delta)$ and $\gamma(\Delta)$ with Δ is shown in Fig. 17.19. From Eq. (65) we note that $\gamma(\Delta)$ corresponds to an amplitude-modulated sinusoidal variation. The sinusoidal variation has a period

$$p = \frac{2c}{\nu_1+\nu_2} = \frac{2\lambda_1\lambda_2}{\lambda_1+\lambda_2} \simeq \lambda_0 \quad \textbf{(67)}$$

where $\lambda_0 (\simeq \lambda_1 \simeq \lambda_2)$ is the average wavelength. The modulation amplitude has zeros at Δ values given by

$$2\pi\frac{\nu_1-\nu_2}{2c}\Delta = \left(m+\frac{1}{2}\right)\pi$$

or

$$\Delta = \left(m+\frac{1}{2}\right)\frac{c}{\nu_1-\nu_2} \quad \textbf{(68)}$$

Hence the minimum path difference at which the visibility vanishes is given by

$$\Delta_m = \frac{c}{2(\nu_1-\nu_2)} = \frac{c}{2\,\delta\nu} \quad \textbf{(69)}$$

which corresponds to the coherence length of the source. Expressing $\delta\nu$ in terms of $\delta\lambda$, we have

$$L_c = \Delta_m = \frac{\lambda^2}{2\delta\lambda} \quad \textbf{(70)}$$

consistent with Eq. (2).

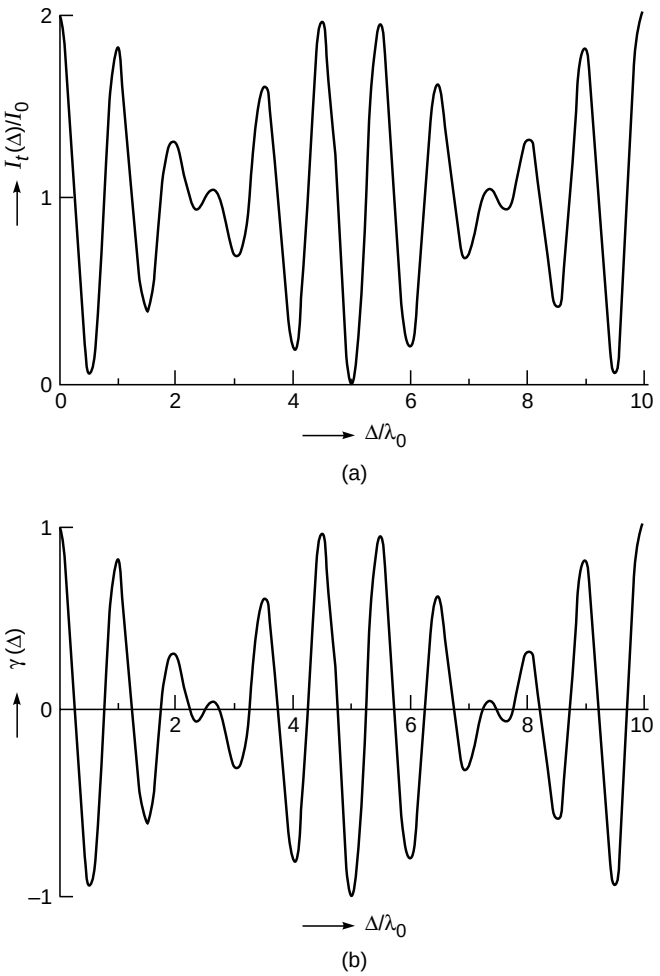


Fig. 17.19 (a) The variation of the total intensity at O as a function of the path difference Δ for a source emitting two monochromatic lines. (b) The corresponding variation of $\gamma(\Delta)$ with Δ.

The difference in path difference between two consecutive positions of the disappearance of the fringes is $c/\delta\nu = \lambda^2/\delta\lambda$. As a simple consequence of this, we may consider Newton's rings experiment with a sodium lamp. If we assume that the sodium lamp emits two discrete wavelengths λ_1 and λ_2, then as we raise the convex lens above the glass plate, we should have a periodic appearance of fringes as we discussed in Example 15.3.

17.8.2 Inversion to Recover $I(\nu)$ from $\gamma(\Delta)$

In an actual experiment, we measure $I_t(\Delta)$ and I_T. Thus Eq. (60) has to be inverted to obtain the source spectral distribution $I(\nu)$ from the measured $\gamma(\Delta)$. To do this, we just

multiply Eq. (60) by $\cos 2\pi\nu'\Delta/c$ and integrate over Δ. Thus

$$\int_0^\infty \gamma(\Delta)\cos\frac{2\pi\nu'\Delta}{c}d\Delta$$

$$= \frac{1}{I_T}\int_0^\infty d\Delta\int_0^\infty d\nu\, I(\nu)\cos\frac{2\pi\nu\Delta}{c}\cos\frac{2\pi\nu'\Delta}{c}$$

$$= \frac{1}{I_T}\int_0^\infty d\nu\, I(\nu)\int_0^\infty \cos\frac{2\pi\nu\Delta}{c}\cos\frac{2\pi\nu'\Delta}{c}d\Delta$$

Now,

$$\int_0^\infty \cos\frac{2\pi\nu\Delta}{c}\cos\frac{2\pi\nu'\Delta}{c}d\Delta$$

$$= \frac{1}{2}\int_{-\infty}^{+\infty}\cos\frac{2\pi\nu\Delta}{c}\cos\frac{2\pi\nu'\Delta}{c}d\Delta$$

since the integrand is an even function of Δ. Writing the two cosine terms in terms of exponentials and using

$$\int_{-\infty}^{+\infty} e^{\pm 2\pi i(\nu-\nu')\Delta/c}\,d\Delta = \delta\left(\frac{\nu-\nu'}{c}\right) = c\delta(\nu-\nu') \qquad \textbf{(71)}$$

and

$$\int_{-\infty}^{+\infty} e^{\pm 2\pi i(\nu+\nu')\Delta/c}\,d\Delta = 0 \qquad \textbf{(72)}$$

(since ν and ν' are positive), we obtain

$$\int_0^\infty \gamma(\Delta)\cos\frac{2\pi\nu'\Delta}{c}d\Delta = \frac{c}{4I_T}\int_0^\infty \delta(\nu-\nu')\,I(\nu)\,d\nu$$

$$= \frac{c}{4I_T}I(\nu') \qquad \textbf{(73)}$$

Hence
$$I(\nu) = \frac{4I_T}{c}\int_0^\infty \gamma(\Delta)\cos\frac{2\pi\nu\Delta}{c}d\Delta \qquad \textbf{(74)}$$

Thus one can obtain the source spectral distribution $I(\nu)$ from the measured $\gamma(\Delta)$ just by a cosine transformation. Such an inversion from $\gamma(\Delta)$ to $I(\nu)$ is usually performed using a computer.

17.8.3 Resolution

From Eq. (74), it follows that to obtain $I(\nu)$ one must measure $\gamma(\Delta)$ for all values of path difference Δ lying between 0 and ∞. Since in an actual experiment there is a maximum limit to path differences that can be introduced, this maximum path difference determines the resolution obtainable in the estimated $I(\nu)$. To estimate the resolution, we consider a perfectly monochromatic beam of frequency ν_0 incident on the interferometer. We have seen that for such a case $\gamma(\Delta)$ varies with Δ as given by Eq. (62). Now in the experiment if Δ_m is the maximum path difference measured, then $\gamma(\Delta)$ is

$$\gamma(\Delta) = \begin{cases}\cos\dfrac{2\pi\nu_0\Delta}{c} & 0<\Delta<\Delta_m\\ 0 & \text{otherwise}\end{cases} \qquad \textbf{(75)}$$

Hence using Eq. (74), we have

$$I(\nu) = \frac{4I_T}{c}\int_0^{\Delta_m}\cos\left(\frac{2\pi\nu_0\Delta}{c}\right)\cos\left(\frac{2\pi\nu\Delta}{c}\right)d\Delta$$

$$= \frac{2I_T}{c}\int_0^{\Delta_m}\left[\cos\frac{2\pi(\nu+\nu_0)\Delta}{c}+\cos\frac{2\pi(\nu-\nu_0)\Delta}{c}\right]d\Delta$$

$$= \frac{2I_T}{c}\left[\frac{\sin\dfrac{2\pi(\nu+\nu_0)\Delta_m}{c}}{\dfrac{2\pi}{c}(\nu+\nu_0)}+\frac{\sin\dfrac{2\pi(\nu-\nu_0)\Delta_m}{c}}{\dfrac{2\pi}{c}(\nu-\nu_0)}\right]$$

Since ν and ν_0 are both positive and much much greater than c/Δ, the first term in the RHS within brackets is negligible and we obtain

$$I(\nu) \simeq \frac{2I_T}{c}\left[\frac{\sin\dfrac{2\pi(\nu-\nu_0)\Delta_m}{c}}{\dfrac{2\pi}{c}(\nu-\nu_0)}\right] \qquad \textbf{(76)}$$

The above estimated source spectrum is similar to that shown in Fig. 17.16. The spectrum is peaked at ν_0, and the first zero appears at

$$\frac{2\pi(\nu-\nu_0)}{c}\Delta_m = \pm\pi$$

or

$$\nu = \nu_0 \pm \frac{c}{2\Delta_m} \qquad \textbf{(77)}$$

Thus although the incident beam is monochromatic, the inversion process gives us a finite spectral width due to a finite value of Δ_m.

If the incident source contains two frequencies, then we may use the Rayleigh criterion and define the minimum resolvable frequency separation to be the frequency width

from the peak to the first zero in $I(\nu)$. Hence

$$\delta\nu = \frac{c}{2\Delta_m} \tag{78}$$

Hence, the larger the maximum path difference Δ_m over which γ is measured, the higher will be the resolution.

As an example, if $\Delta_m = 5$ cm, then

$$\delta\nu = \frac{3\times 10^{10}}{2\times 5} = 3\,\text{GHz}$$

At $\lambda = 1$ μm, this corresponds to $\delta\lambda = 0.1$ Å.

Example 17.2 We consider a quasi-monochromatic source characterized by a Gaussian spectral distribution given by

$$I(\nu) = \frac{1}{\sqrt{\pi}\,\delta\nu} I_0 e^{-(\nu-\nu_0)^2/(\delta\nu)^2}$$

$$= \frac{I_0}{\sqrt{\pi}}\tau e^{-(\nu-\nu_0)^2\tau^2} \tag{79}$$

Here $\delta\nu = 1/\tau$ characterizes the width of the spectrum since $I(\nu)$ drops to $1/e$ of the value at $\nu = \nu_0$ at $\nu = \nu_0 \pm \delta\nu$. For a quasi-monochromatic source $\delta\nu/\nu_0 << 1$. Thus

$$I_T = \int_0^\infty I(\nu)\,d\nu$$

$$= \frac{I_0\tau}{\sqrt{\pi}}\int_0^\infty e^{-(\nu-\nu_0)^2\tau^2} d\nu$$

$$\simeq \frac{I_0}{\sqrt{\pi}}\tau\int_{-\infty}^{+\infty} e^{-(\nu-\nu_0)^2\tau^2} d\nu \tag{80}$$

where in the last step we used the condition $1/\tau = \delta\nu << \nu_0$. If we now use the integral

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2+\beta x}\,dx = \left(\frac{\pi}{\alpha}\right)^{1/2}\exp\left(\frac{\beta^2}{4\alpha}\right) \qquad \text{Re}\,\alpha > 0 \tag{81}$$

we obtain

$$I_T = I_0 \tag{82}$$

Now,

$$\int_0^\infty I(\nu)\cos\frac{2\pi\nu\Delta}{c}\,d\nu$$

$$= \frac{I_0\tau}{\sqrt{\pi}}\int_0^\infty e^{-(\nu-\nu_0)^2\tau^2}\cos\frac{2\pi\nu\Delta}{c}\,d\nu$$

$$\simeq \frac{\tau}{\sqrt{\pi}} I_0 \int_{-\infty}^{+\infty} e^{-(\nu-\nu_0)^2\tau^2}\cos\frac{2\pi\nu\Delta}{c}\,d\nu$$

$$= \frac{\tau}{\sqrt{\pi}} I_0\,\text{Re}\int_{-\infty}^{+\infty} e^{-(\nu-\nu_0)^2\tau^2} e^{i2\pi\nu\Delta/c}\,d\nu$$

$$= \frac{\tau}{\sqrt{\pi}} I_0\,\text{Re}\,e^{2\pi i\nu_0\Delta/c}\int_{-\infty}^{+\infty} e^{-\xi^2\tau^2} e^{i2\pi\xi\Delta/c}\,d\xi \qquad \xi = \nu - \nu_0$$

$$= I_0\,\text{Re}\left[e^{2\pi i\nu_0\Delta/c}\,\exp\left(-\frac{\pi^2\Delta^2}{c^2\tau^2}\right)\right]$$

where we have used Eq. (81) with $\alpha = \tau$ and $\beta = i2\pi\Delta/c$. Thus,

$$\int_0^\infty I(\nu)\cos\frac{2\pi\nu\Delta}{c}\,d\nu = I_0\exp\left(-\frac{\pi^2\Delta^2}{c^2\tau^2}\right)\cos\frac{2\pi\nu_0\Delta}{c} \tag{83}$$

Hence,

$$\gamma(\Delta) = \exp\left(-\frac{\pi^2\Delta^2}{c^2\tau^2}\right)\cos\frac{2\pi\nu_0\Delta}{c} \tag{84}$$

Figure 17.20 shows the source spectral distribution as well as the variation of $\gamma(\Delta)$ with Δ. Notice that in this case for path differences $\Delta << c/\delta\nu$, $\gamma(\Delta) \simeq \cos(2\pi\nu_0\Delta/c)$ much like that for a

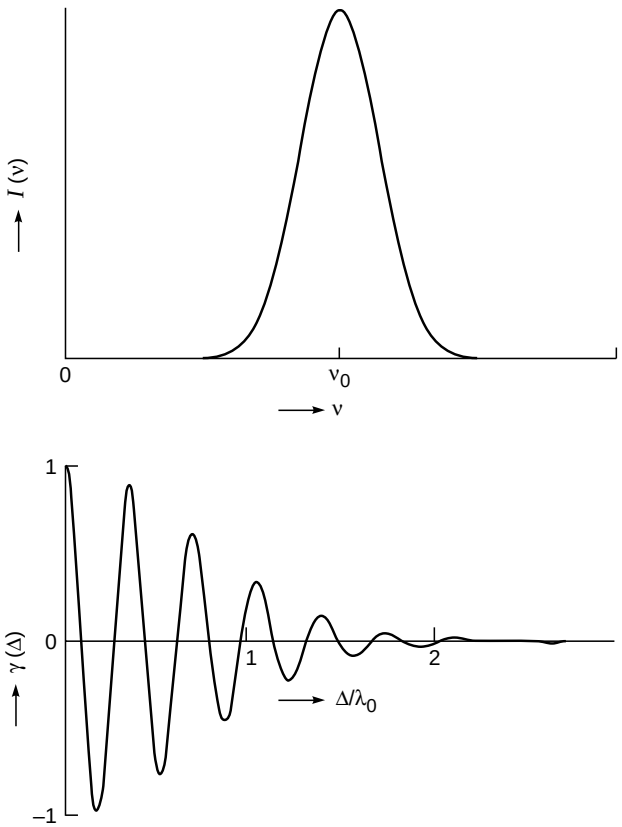


Fig. 17.20 Spectral distribution and the variation of $\gamma(\Delta)$ with Δ for a source characterized by Eq. (79).

monochromatic source. But as the path difference increases, the modulation amplitude of $\gamma(\Delta)$ is reduced. For good contrast, one must have

$$\Delta << c\tau = \frac{c}{\delta\nu} \tag{85}$$

We may thus define the coherence length as

$$L_c = c\tau = \frac{c}{\delta\nu} \tag{86}$$

consistent with Eq. (2).

Example 17.3 Consider a quasi-monochromatic source characterized by a spectral distribution

$$I(\nu) = \begin{cases} \dfrac{1}{\delta\nu} I_0 & \nu_0 - \frac{1}{2}\delta\nu < \nu < \nu_0 + \frac{1}{2}\delta\nu \\ 0 & \text{otherwise} \end{cases} \tag{87}$$

Calculate $\gamma(\Delta)$ and show that again for path differences $\Delta >> c/\delta\nu$, the contrast will be very poor.

Solution:

$$I_T = \frac{1}{\delta\nu} I_0 \int_{\nu_0 - \frac{1}{2}\delta\nu}^{\nu_0 + \frac{1}{2}\delta\nu} d\nu = I_0 \tag{88}$$

$$\int_0^\infty I(\nu)\cos\frac{2\pi\nu\Delta}{c} d\nu = \frac{I_0}{\delta\nu}\cos\frac{2\pi\nu_0\Delta}{c}\frac{\sin \pi\delta\nu\Delta/c}{\pi\Delta/c}$$

Thus

$$\gamma(\Delta) = \frac{1}{I_T}\int_0^\infty I(\nu)\cos\frac{2\pi\nu\Delta}{c} d\nu$$

$$= \frac{\sin(\pi\delta\nu\Delta/c)}{\pi\delta\nu\Delta/c}\cos\frac{2\pi\nu_0\Delta}{c} \tag{89}$$

For $\Delta << c/\delta\nu$, $\gamma(\Delta) \simeq \cos(2\pi\nu_0\Delta/c)$ and the contrast vanishes for

$$\Delta = \frac{c}{\delta\nu} \tag{90}$$

which represents the coherence length. Plot $\gamma(\Delta)$ as a function of Δ, and notice that unlike in the earlier example, in this case $\gamma(\Delta)$ does not monotonically reduce to zero.

For more details on Fourier transform spectroscopy, see Refs. 10, 12, and 18.

Summary

- The coherence time τ_c represents the average duration of the wave trains; i.e., the electric field remains sinusoidal for times of the order of τ_c.
- The length of the wave train, given by

  $$L_c = c\tau_c$$

  (where c is the speed of the light in free space) is referred to as the coherence length. For example, for the red cadmium line (λ = 6438 Å), $\tau_c \sim 10^{-9}$ s; the corresponding coherence length is ~30 cm.
- The lateral coherence width l_w of an extended incoherent source represents the distance over which the beam may be assumed to be spatially coherent; it is given by

  $$l_w \approx \frac{\lambda}{\theta}$$

  where θ is the angle subtended by the source at the point of observation.
- Using the concept of spatial coherence, Michelson developed an ingenious method for determining the angular diameter of stars. The method is based on the result that for a distant circular source, the interference fringes (formed by two pinholes) will disappear if the distance between the two pinholes is given by

  $$d = 1.22\frac{\lambda}{\theta}$$

  where θ is the angle subtended by the circular source.
- Using two laser beams, it is possible to observe optical beats.
- In the two-beam interference pattern, the contrast of the interference fringes varies as the optical path difference Δ is varied, beginning from an extremely good contrast for $\Delta << L_c$ to a very poor contrast for $\Delta >> L_c$.
- Indeed from a knowledge of variation of intensity with optical path difference, one can obtain the source spectral distribution by a Fourier transformation.

Problems

17.1 The orange krypton line (λ = 6058 Å) has a coherence length of ~20 cm. Calculate the line width and the frequency stability.

[**Ans:** ~0.018 Å, ~3×10^{-6}]

17.2 Laser line widths as low as 20 Hz have been obtained. Calculate the coherence length and the frequency stability. Assume λ = 6328 Å.

17.3 In Sec. 17.4 we mentioned that the lateral coherence width of a circular source is $1.22\lambda/\theta$. It can be shown that for good coherence (i.e., for a visibility of 0.88 or better), the coherence width should be $\leq 0.3\lambda/\theta$. Assuming that the angular diameter of the Sun is about 30 minutes, calculate the distance between two pinholes which would produce a clear interference pattern.

[**Ans:** ~ 0.02 mm]

17.4 Calculate the distance at which a source of diameter 1 mm should be kept from a screen so that two points separated by

a distance of 0.5 mm may be said to be coherent. Assume $\lambda = 6 \times 10^{-5}$ cm.

17.5 In a Michelson interferometer experiment, it is found that for a source *S*, as one of the mirrors is moved away from the equal path length position by a distance of about 5 cm, the fringes disappear. What is the coherence time of the radiation emerging from the source?

17.6 If we perform Young's double-hole experiment using white light, then only a few colored fringes are visible. Assuming that the visible spectrum extends from 4000 to 7000 Å, explain this phenomenon qualitatively on the basis of coherence length.

17.7 Using the stellar interferometer, Michelson observed for the star Betelgeuse that the fringes disappear when the distance between the movable mirrors is 25 in. Assuming $\lambda \simeq 6 \times 10^{-5}$ cm, calculate the angular diameter of the star.

17.8 Consider Young's double-hole experiment as shown in Fig. 17.2. The distance $SS_1 \simeq 1$ m and $S_1S_2 \approx 0.5$ mm. Calculate the angular diameter of the hole *S* which will produce a good interference pattern on the screen. Assume $\lambda = 6000$ Å.

17.9 Assume a Gaussian pulse of the form

$$\Psi(x = 0,\, t) = E_0 \exp\left(-\frac{t^2}{2\tau^2}\right) e^{i\omega_0 t}$$

Show that the Fourier transform is given by

$$A(\omega) = E_0 \tau \exp\left[-\frac{1}{2}(\omega - \omega_0)^2 \tau^2\right]$$

You will have to use the following integral:

$$\int_{-\infty}^{+\infty} \exp(-\alpha x^2 + \beta x)\, dx = \left(\frac{\pi}{\alpha}\right)^{1/2} \exp\left(\frac{\beta^2}{4\alpha}\right) \qquad \alpha > 0$$

Show that the temporal coherence is $\sim\tau$. Assume $\tau >> 1/\omega_0$, plot the Fourier transform $A(\omega)$ (as a function of ω) and interpret it physically. Show that the frequency spread $\Delta\omega \sim 1/\tau$.

17.10 In Prob. 17.9, assume $\lambda_0 = 6 \times 10^{-5}$ cm and $\tau \sim 10^{-9}$ s. Calculate the frequency components predominantly present in the pulse, and compare than with the case corresponding to $\tau \sim 10^{-6}$ s.

REFERENCES AND SUGGESTED READINGS

1. D. E. Bailey and M. J. Welch, "Moiré Fringes," *Proceedings of the Conference and Workshop on the Teaching of Optics*, Eds. G. I. Opat, D. Booth, A. P. Mazzolini, and G. Smith, University of Melbourne, 1989.
2. J. Beran and G. B. Parrent, *Theory of Partial Coherence*, Prentice-Hall, Englewood Cliffs, N.J., 1964.
3. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, 1975.
4. M. Francon, *Diffraction: Coherence in Optics*, Pergamon Press, Oxford, 1966.
5. A. T. Forrester, "On Coherence Properties of Light Waves," *American Journal of Physics*, Vol. 24, p. 192, 1956.
6. A. T. Forrester, R. A. Gudmundsen, and P. O. Johnson, "Photoelectric Mixing of Incoherent Light," *Physical Review*, Vol. 99, No. 6, p. 1891, 1955.
7. A. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978. Reprinted by Macmillan India, New Delhi.
8. E. Hecht and A. Zajac, *Optics*, Addison-Wesley, Reading, Mass., 1974.
9. T. S. Jaseja, A. Javan, and C. H. Townes, "Frequency Stability of He–Ne Masers and Measurement of Length," *Physical Review Letters*, Vol. 10, p. 165, 1963.
10. M. V. Klein, *Optics*, Wiley, New York, 1970.
11. M. S. Lipsett and L. Mandel, "Coherence Time Measurement of Light from Ruby Optical Masers," *Nature*, Vol. 199, p. 553, 1963.
12. G. F. Lothian, *Optics and Its Uses*, Van Nostrand Reinhold, New York, 1975.
13. H. F. Meiners, *Physics Demonstrations and Experiments*, Vol. 2, The Ronald Press Co., New York, 1970.
14. D. F. Nelson and R. J. Collins, "Spatial Coherence in the Output of an Optical Maser," *Journal of Applied Physics*, Vol. 32, p. 739, 1961.
15. A. E. Siegman, *Lasers*, Oxford University Press, 1986.
16. B. J. Thompson, *Journal Sociology Photography Institute Engineering*, Vol. 4, p. 7, 1965.
17. K. Thyagarajan and A. K. Ghatak, *Lasers: Theory and Applications*, Plenum Press, New York, 1981. Reprinted by Macmillan India, New Delhi.
18. G. A. Vanasse and H. Sakai, "Fourier Spectroscopy," in *Progress in Optics*, Vol. 6, Ed. E. Wolf, North-Holland Pub. Co., Amsterdam, 1967.
19. H. Weltin, "Light Beats," *American Journal of Physics*, Vol. 30, p. 653, 1962.
20. A. Sharma, A. K. Ghatak, and H. C. Kandpal, "Coherence," *Encyclopaedia of Modern Optics*, Eds. R. Guenther, A. Miller, L. Bayvel, and J. Midwinter, Elsevier, 2005.

PART 4
Diffraction

Chapters 18, 19, and 20 cover the very important area of diffraction and discuss the principle behind topics such as diffraction divergence of laser beams, resolving power of telescopes, laser focusing, spatial frequency filtering, and X-ray diffraction. Chapter 21 discusses holography, giving the underlying principle and many applications. Dennis Gabor received the 1971 Nobel Prize in Physics for discovering the principle of holography.

Chapter Eighteen

FRAUNHOFER DIFFRACTION: I

No one has ever been able to define the difference between interference and diffraction satisfactorily. It is just a question of usage, and there is no specific, important physical difference between them. The best we can do is, roughly speaking, is to say that when there are only a few sources, say two, interfering, then the result is usually called interference, but if there is a large number of them, it seems that the word diffraction is more often used.

—Richard Feynman, *Feynman Lectures on Physics*, Vol. 1

Important Milestones

1819	*Joseph Fraunhofer demonstrated the diffraction of light by gratings which were initially made by winding fine wires around parallel screws.*
1823	*Fraunhofer published his theory of diffraction.*
1835	*George Airy calculated the (Fraunhofer) diffraction pattern produced by a circular aperture.*

18.1 INTRODUCTION

Consider a plane wave incident on a long narrow slit of width b (see Fig. 18.1). According to geometrical optics, one expects region AB of screen SS' to be illuminated and the remaining portion (known as the geometrical shadow) to be absolutely dark. However, if the observations are made carefully, then one finds that if the width of the slit is not very large compared to the wavelength, then the light intensity in region AB is not uniform and there is also some intensity inside the geometrical shadow. Further, if the width of the slit is made smaller, larger amounts of energy reach the geometrical shadow. This spreading out of a wave when it passes through a narrow opening is usually referred to as diffraction, and the intensity distribution on the screen is known as the diffraction pattern. We will discuss the phenomenon of diffraction in this chapter and will show that the spreading out decreases with a decrease in wavelength. Indeed, since the light wavelengths are very small ($\lambda \sim 5 \times 10^{-5}$ cm), the effects due to diffraction are not readily observed.

Fig. 18.1 If a plane wave is incident on an aperture, then according to geometrical optics a sharp shadow will be cast in region AB of the screen.

Actually, there is not much of a difference between the phenomena of interference and diffraction; indeed, interference corresponds to the situation when we consider the superposition of waves coming out from a number of point sources, and diffraction corresponds to the situation when we consider waves coming out from an area source such as a circular or rectangular aperture or even a large number of rectangular apertures (such as the diffraction grating).

The diffraction phenomena are usually divided into two categories: Fresnel diffraction and Fraunhofer diffraction.

In the Fresnel class of diffraction the source of light and the screen are, in general, at a finite distance from the diffracting aperture [see Fig. 18.2(a)]. In the Fraunhofer class of diffraction, the source and the screen are at infinite distances from the aperture; this is easily achieved by placing the source on the focal plane of a convex lens and placing the screen on the focal plane of another convex lens [see Fig. 18.2(b)]. The two lenses effectively moved the source and the screen to infinity because the first lens makes the light beam parallel and the second lens effectively makes the screen receive a parallel beam of light. It turns out that it is much easier to calculate the

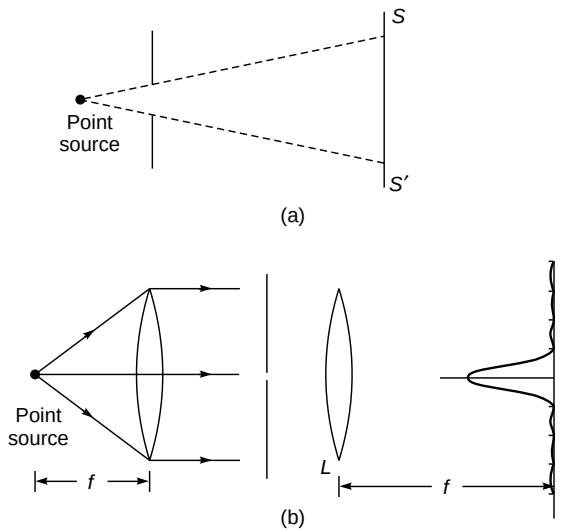


Fig. 18.2 (a) When either the source or the screen (or both) is at a finite distance from the aperture, the diffraction pattern corresponds to the Fresnel class. (b) In the Fraunhofer class both the source and the screen are at infinity.

intensity distribution of a Fraunhofer diffraction pattern, which we will do in this chapter. Further, the Fraunhofer diffraction pattern is not difficult to observe. All that one needs is an ordinary laboratory spectrometer; the collimator renders a parallel beam of light, and the telescope receives parallel beams of light on its focal plane. The diffracting aperture is placed on the prism table. In Chap. 20 we will study the Fresnel class of diffraction and will discuss the transition from the Fresnel region to the Fraunhofer region.

18.2 SINGLE-SLIT DIFFRACTION PATTERN

We will first study the Fraunhofer diffraction pattern produced by an infinitely long slit, of width b. A plane wave is assumed to fall normally on the slit, and we wish to calculate the intensity distribution on the focal plane of lens L [see Fig. 18.3(a)]. We assume that the slit consists of a large number of equally spaced point sources and that each point on

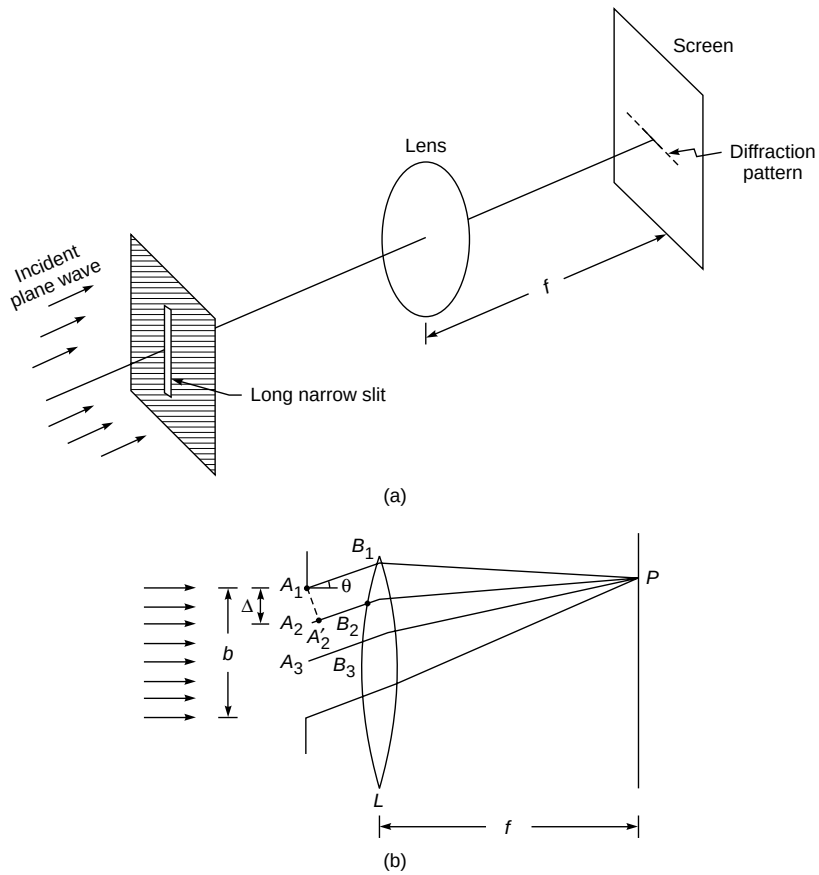


Fig. 18.3 (a) Diffraction of a plane wave incident normally on a long narrow slit of width b. Notice that the spreading occurs along the width of the slit. (b) To calculate the diffraction pattern, the slit is assumed to consist of a large number of equally spaced points.

the slit is a source of Huygens' secondary wavelets which interfere with the wavelets emanating from other points. Let the point sources be at A_1, A_2, A_3, . . ., and let the distance between two consecutive points be Δ [see Fig. 18.3(b)]. Thus, if the number of point sources is n, then

$$b = (n-1)\Delta \tag{1}$$

We will now calculate the resultant field produced by these n sources at point P, with P being an arbitrary point (on the focal plane of the lens) receiving parallel rays making an angle θ with the normal to the slit [see Fig. 18.3(b)]. Since the slit actually consists of a continuous distribution of sources, we will, in the final expression, let n go to infinity and Δ go to zero such that $n\Delta$ tends to b.

Now, at point P, the amplitudes of the disturbances reaching from $A_1, A_2, \ldots$ will be very nearly the same because point P is at a distance which is very large in comparison to b [see Fig. 18.3(b)]. However, because of even slightly different path lengths to point P, the field produced by A_1 will differ in phase from the field produced by A_2.

For an incident plane wave, points $A_1, A_2, \ldots$ are in phase, and, therefore, the additional path traversed by the disturbance emanating from point A_2 will be A_2A_2', where A_2' is the foot of the perpendicular drawn from A_1 on A_2B_2. This follows from the fact that the optical paths A_1B_1P and $A_2'B_2P$ are the same. If the diffracted rays make an angle θ with the normal to the slit, then the path difference is

$$A_2A_2' = \Delta \sin\theta$$

The corresponding phase difference ϕ is given by

$$\phi = \frac{2\pi}{\lambda}\Delta \sin\theta \tag{2}$$

Thus, if the field at point P due to the disturbance emanating from point A_1 is $a\cos\omega t$, then the field due to the disturbance emanating from A_2 is $a\cos(\omega t - \phi)$. Now the difference in the phases of the disturbance reaching from points A_2 and A_3 will also be ϕ, and thus the resultant field at point P is given by

$$E = a[\cos\omega t + \cos(\omega t - \phi) + \cdots + \cos[(\omega t - (n-1)\phi)] \tag{3}$$

where

$$\phi = \frac{2\pi}{\lambda}\Delta \sin\theta$$

Now, we showed in Sec. 11.7 that

$$\cos\omega t + \cos(\omega t - \phi) + \cdots + \cos[\omega t - (n-1)\phi] = \frac{\sin(n\phi/2)}{\sin(\phi/2)}\cos\left[\omega t - \frac{1}{2}(n-1)\phi\right] \tag{4}$$

Thus

$$E = E_0\cos\left[\omega t - \tfrac{1}{2}(n-1)\phi\right] \tag{5}$$

where the amplitude E_θ of the resultant field is given by[1]

$$E_\theta = a\frac{\sin(n\phi/2)}{\sin(\phi/2)} \tag{6}$$

In the limit of $n \to \infty$ and $\Delta \to 0$ in such a way that $n\Delta \to b$, we have

$$\frac{n\phi}{2} = \frac{\pi}{\lambda}n\Delta\sin\theta \to \frac{\pi}{\lambda}b\sin\theta$$

Further

$$\phi = \frac{2\pi}{\lambda}\Delta\sin\theta = \frac{2\pi}{\lambda}\frac{b\sin\theta}{n}$$

will tend to zero, and, so we may write

$$E_\theta \approx \frac{a\sin\left(n\phi/2\right)}{\phi/2} = na\,\frac{\sin\left(\pi b\sin\theta/\lambda\right)}{\left(\pi b\sin\theta/\lambda\right)} = A\,\frac{\sin\beta}{\beta} \tag{7}$$

where[2]

$$A = na$$

and

$$\beta = \frac{\pi b\sin\theta}{\lambda} \tag{8}$$

Thus

$$E = A\frac{\sin\beta}{\beta}\cos(\omega t - \beta) \tag{9}$$

The corresponding intensity distribution is given by

$$I = I_0\frac{\sin^2\beta}{\beta^2} \tag{10}$$

where I_0 represents the intensity at $\theta = 0$.

18.2.1 Positions of Maxima and Minima

The variation of the intensity with β is shown in Fig. 18.4(a). It is obvious from Eq. (10) that the intensity is zero when

$$\beta = m\pi \qquad m \neq 0 \tag{11}$$

[1] Equation (6) represents the amplitude distribution due to the interference of n point sources. Thus, for $n = 2$, the amplitude E_θ becomes $\cos(\phi/2)$, giving rise to $\cos^2(\phi/2)$ intensity distribution [cf. Eq. (13) of Chap. 14]. Notice that if we have a large number of equidistant sources oscillating in phase, then the propagation is only in certain directions where the displacements add in phase.

[2] Note that in the limit $n \to \infty$ and $a \to 0$ the product na tends to a finite limit.

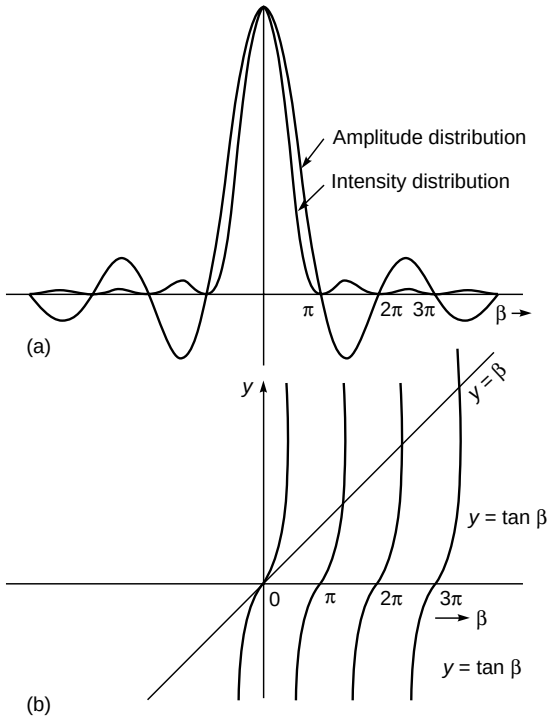


Fig. 18.4 (a) The intensity distribution corresponding to the single-slit Fraunhofer diffraction pattern. (b) Graphical method for determining the roots of the equation $\tan\beta = \beta$.

When $\beta = 0$, $(\sin\beta)/\beta = 1$ and $I = I_0$, which corresponds to the maximum of the intensity. Substituting the value of β, one obtains

$$b \sin\theta = m\lambda \qquad m = \pm 1, \pm 2, \pm 3, \ldots \text{ (minima)} \qquad \textbf{(12)}$$

as the conditions for minima. The first minimum occurs at $\theta = \pm \sin^{-1}(\lambda/b)$; the second minimum occurs at $\theta = \pm \sin^{-1}(2\lambda/b)$, etc. Since $\sin\theta$ cannot exceed unity, the maximum value of m is the integer which is less than (and closest to) b/λ.

The positions of minima can be directly obtained by simple qualitative arguments. Let us consider the case $m = 1$. The angle θ satisfies the equation

$$b \sin\theta = \lambda \qquad \textbf{(13)}$$

We divide the slit into two halves as shown in Fig. 18.5. Consider two points A and A' separated by a distance $b/2$. Clearly the path difference between the disturbances (reaching the point P) emanating from A and A' is $(b/2)\sin\theta$, which in this case is $\lambda/2$. The corresponding phase difference will be π, and the resultant disturbance will be zero. Similarly, the disturbance from point B will be canceled by the

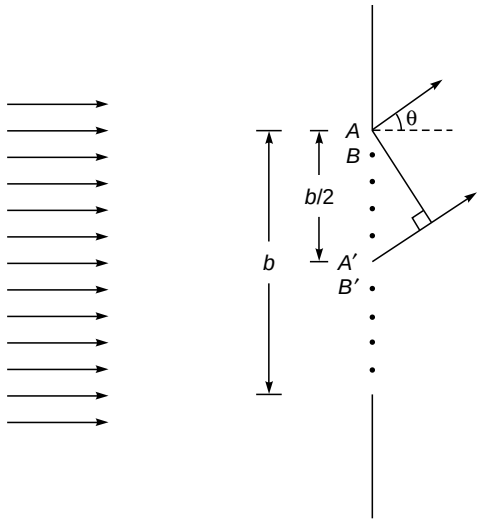


Fig. 18.5 The slit is divided into two halves for deriving the condition for the first minimum.

disturbance reaching from point B'. Thus the resultant disturbance due to the upper half of the slit will be canceled by the disturbances reaching from the lower half, and the resultant intensity will be zero. In a similar manner when

$$b \sin\theta = 2\lambda \qquad \textbf{(14)}$$

we divide the slit into four parts; the first and second quarters cancel each other, and the third and fourth quarters cancel each other. Similarly when $m = 3$, the slit is divided into six parts, and so on.

To determine the positions of maxima, we differentiate Eq. (10) with respect to β and set it equal to zero. Thus

$$\frac{dI}{d\beta} = I_0\left(\frac{2\sin\beta\cos\beta}{\beta^2} - \frac{2\sin^2\beta}{\beta^3}\right) = 0$$

or

$$\sin\beta(\beta - \tan\beta) = 0 \qquad \textbf{(15)}$$

The condition $\sin\beta = 0$, or $\beta = m\pi$ ($m \neq 0$), corresponds to minima. The conditions for maxima are roots of the transcendental equation

$$\tan\beta = \beta \qquad \text{(maxima)} \qquad \textbf{(16)}$$

The root $\beta = 0$ corresponds to the central maximum. The other roots can be found by determining the points of intersections of the curves $y = \beta$ and $y = \tan\beta$ [see Fig. 18.4(b)]. The intersections occur at $\beta = 1.43\pi$, $\beta = 2.46\pi$, etc., and are known as the first maximum, the second maximum, etc. Since

$$\left(\frac{\sin 1.43\pi}{1.43\pi}\right)^2$$

is about 0.0496, the intensity of the first maximum is about 4.96% of the central maximum. Similarly, the intensities of the second and third maxima are about 1.68% and 0.83% of the central maximum, respectively.

Example 18.1 A parallel beam of light is incident normally on a narrow slit of width 0.2 mm. The Fraunhofer diffraction pattern is observed on a screen which is placed at the focal plane of a convex lens whose focal length is 20 cm. Calculate the distance between the first two minima and the first two maxima on the screen. Assume that $\lambda = 5 \times 10^{-5}$ cm and that the lens is placed very close to the slit.

Solution:

$$\frac{\lambda}{b} = \frac{5\times10^{-5}}{2\times10^{-2}} = 2.5 \times 10^{-3}$$

Now, the conditions for diffraction minima are given by $\sin\theta = m\lambda/b$. We assume θ to be small (measured in radians) so that we may write $\sin\theta \approx \theta$ (an assumption which will be justified by subsequent calculations); thus, on substituting the value of λ/b, we get

$$\theta \simeq 2.5 \times 10^{-3} \text{ and } 5 \times 10^{-3} \text{ rad}$$

as the angles of diffraction corresponding to the first and second minima, respectively. Notice that since

$$\sin(2.5\times10^{-3}) = 2.4999973\times10^{-3}$$

the error in the approximation $\sin\theta \simeq \theta$ is about 1 part in 1 million! These minima will be separated by a distance $(5 \times 10^{-3} - 2.5 \times 10^{-3}) \times 20 = 0.05$ cm on the focal plane of the lens. Similarly, the first and second maxima occur at

$$\beta = 1.43\pi \quad \text{and} \quad 2.46\pi$$

respectively. Thus

$$b \sin\theta = 1.43\lambda \quad \text{and} \quad 2.46\lambda$$

or

$$\sin\theta = 1.43 \times 2.5 \times 10^{-3} \quad \text{and} \quad 2.46 \times 2.5 \times 10^{-3}$$

Consequently, the maxima will be separated by the distance given by

$$(2.46 - 1.43) \times 2.5 \times 10^{-3} \times 20 \simeq 0.05 \text{ cm}$$

Example 18.2 Consider, once again, a parallel beam of light ($\lambda = 5 \times 10^{-5}$ cm) to be incident normally on a long narrow slit of width 0.2 mm. A screen is placed at a distance of 3 m from the slit. Assuming that the screen is so far away that the diffraction is essentially of the Fraunhofer type, calculate the total width of the central maximum.

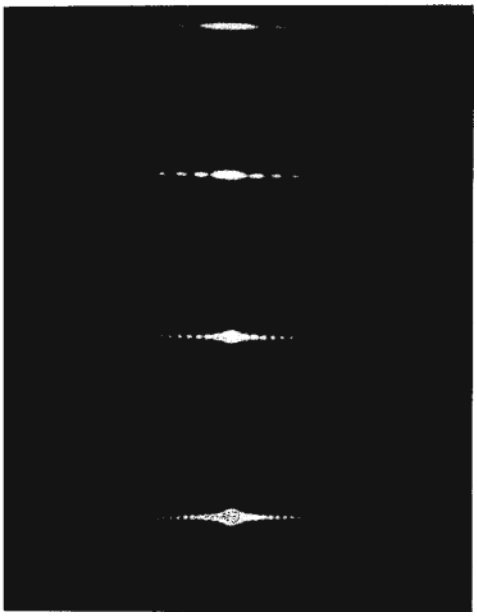


Fig. 18.6 The single-slit diffraction patterns corresponding to b = 0.0088, 0.0176, 0.035, and 0.070 cm, respectively. The wavelength of the light used is 6.328×10^{-5} cm (After Ref. 17; used with permission).

Solution: As in Example 18.1, the first minimum occurs at $\theta \simeq 2.5 \times 10^{-3}$ rad; thus the total width of the central maximum is approximately given by

$$2 \times 3 \times \tan(2.5 \times 10^{-3}) \simeq 0.015 \text{ m}$$

In Fig. 18.6, we have given the actual single-slit diffraction pattern (as seen on a screen) for the following values of slit widths: 8.8×10^{-3}, 1.76×10^{-2}, 3.5×10^{-2}, and 7.0×10^{-2} cm. The light wavelength used was 6328 Å = 6.328×10^{-5} cm. We note the following two points:

1. The spreading is only in the direction of the width of the slit. This is so because the lengths of the slits were very large compared to their widths.
2. The values of λ/b corresponding to the four slit widths are 7.191×10^{-3}, 3.595×10^{-3}, 1.808×10^{-3}, and 0.904×10^{-3}. Thus the diffraction angle at which the first minimum will occur is

$$\theta \simeq \sin\theta = 7.191 \times 10^{-3}, 3.595 \times 10^{-3}, 1.808 \times 10^{-3}, \text{ and } 0.904 \times 10^{-3}$$

where the angles are measured in radians.[3] The intensity distributions predicted by Eq. (10) are given in Fig. 18.7 for $b = 8.8 \times 10^{-3}$ cm and

[3] Figure 18.6 corresponds to the photographic film being 15 ft away from the slit. Thus it records the Fraunhofer pattern (see also Sec. 20.7); and for $b = 8.8 \times 10^{-3}$, 1.76×10^{-2}, 3.5×10^{-2}, and 7.0×10^{-2} cm, the first minima occur at distances of 3.288, 1.644, 0.827, and 0.413 cm, respectively, from the central maximum.

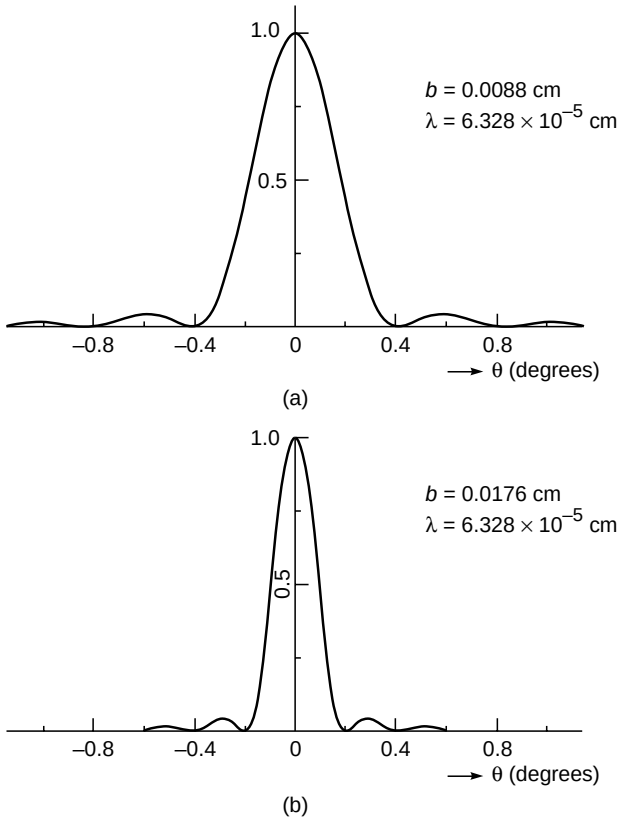


Fig. 18.7 The intensity distribution as calculated by using Eq. (10) for b = 0.0088 cm and 0.0176 cm ($\lambda = 6.328 \times 10^{-5}$ cm).

1.76×10^{-2} cm. For $b >> \lambda$, most of the energy (of the diffracted beam) is contained between the first two minima, i.e., for

$$-\frac{\lambda}{b} \lessapprox \theta \lessapprox \frac{\lambda}{b} \qquad \textbf{(17)}$$

(where θ is measured in radians). Thus the divergence angle (which contains most of the energy) is given by

$$\Delta\theta \sim \frac{\lambda}{b} \qquad \textbf{(18)}$$

For very small values of b, the light almost uniformly spreads out from the slit. Also in the limit of $\lambda \to 0$, $\Delta\theta \to 0$ and the diffraction effects are absent.

18.3 DIFFRACTION BY A CIRCULAR APERTURE

In Sec. 18.2 we showed that when a plane wave is incident on a long narrow slit (of width b), then the emergent wave spreads out (along the width of the slit) with angular divergence $\sim\lambda/b$. In a similar manner one can discuss the diffraction of a plane wave by a circular aperture. Figure 18.8 shows the arrangement for observing the diffraction pattern;

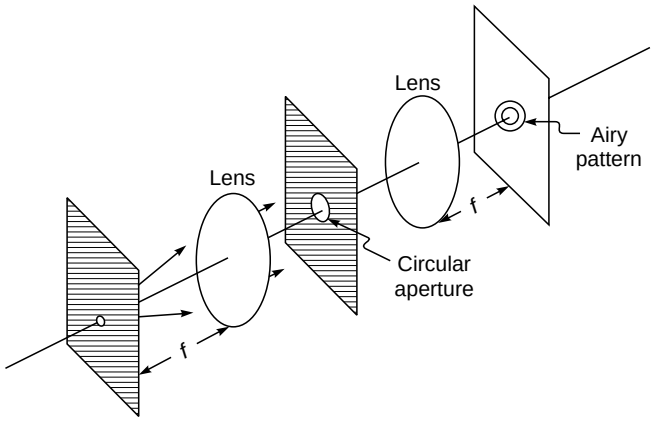


Fig. 18.8 Experimental arrangement for observing the Fraunhofer diffraction pattern by a circular aperture.

a plane wave is incident normally on the circular aperture, a lens whose diameter is much larger than that of the aperture of is placed close to the aperture, and the Fraunhofer diffraction pattern is observed on the focal plane of the lens. Because of the rotational symmetry of the system, the diffraction pattern will consist of concentric dark and bright rings; this diffraction pattern (as observed on the back focal plane of the lens) is known as the Airy pattern. In Fig. 18.9(a) and (b) we have shown the Airy patterns corresponding to the radius of the circular aperture of 0.5 and 0.25 mm, respectively. The detailed derivation of the diffraction pattern for a circular aperture is somewhat complicated (see Sec. 19.7); we give here the final result: The intensity distribution is given by

$$I = I_0 \left[\frac{2J_1(v)}{v}\right]^2 \qquad \textbf{(19)}$$

where

$$v = \frac{2\pi}{\lambda} a \sin\theta \qquad \textbf{(20)}$$

a being the radius of the circular aperture, λ the wavelength of light, and θ the angle of diffraction; I_0 is the intensity at $\theta = 0$ (which represents the central maximum) and $J_1(v)$ is known as the Bessel function of the first order. On the focal plane of the convex lens

$$v \approx \frac{2\pi}{\lambda} a \frac{(x^2 + y^2)^{1/2}}{f} \qquad \textbf{(21)}$$

where f is the focal length of the lens. For those not familiar with Bessel functions, the variation of $J_1(v)$ is somewhat like a damped sine curve (see Fig. 18.10), and although $J_1(0) = 0$, we have

$$\lim_{v \to 0} \frac{2J_1(v)}{v} = 1$$

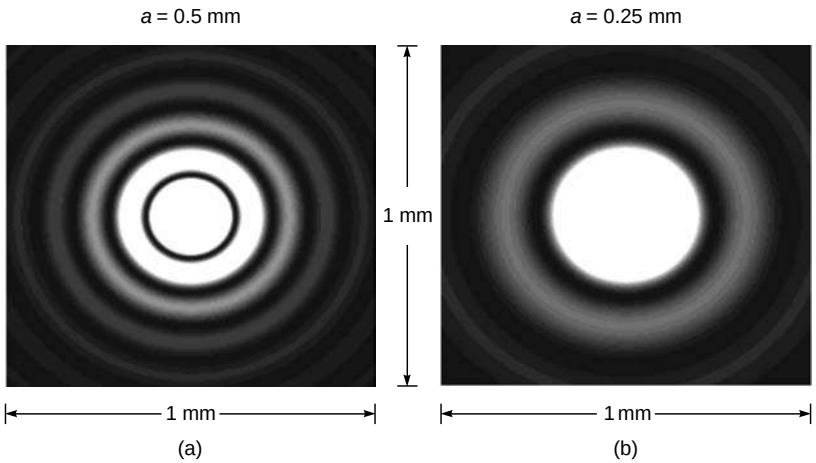


Fig. 18.9 Computer-generated Airy patterns; (a) and (b) correspond to a = 0.5 and 0.25 mm, respectively, at the focal plane of a lens of focal length 20 cm (λ = 0.5 μm).

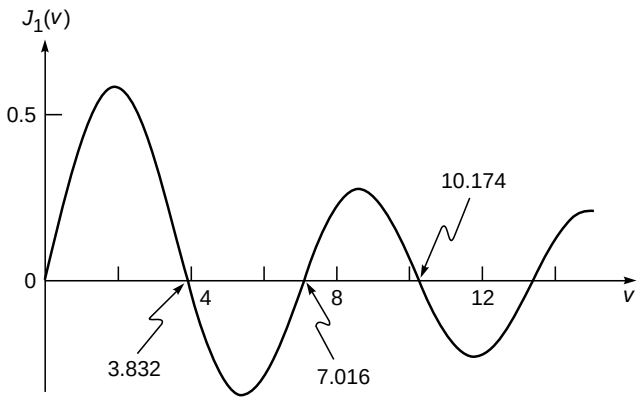


Fig. 18.10 The variation of $J_1(v)$ with v.

similar to the relation

$$\operatorname{Lim}_{x\to 0}\frac{\sin x}{x} = 1$$

Other zeros of $J_1(v)$ occur at

$$v = 3.832, 7.016, 10.174, \ldots$$

In Fig. 18.11 we have plotted the function

$$\left[\frac{2J_1(v)}{v}\right]^2$$

which represents the intensity distribution corresponding to the Airy pattern. Thus the successive dark rings in the Airy pattern (see Fig. 18.9) will correspond to

$$v = \frac{2\pi}{\lambda} a \sin\theta = 3.832, 7.016, 10.174, \ldots \quad \textbf{(22)}$$

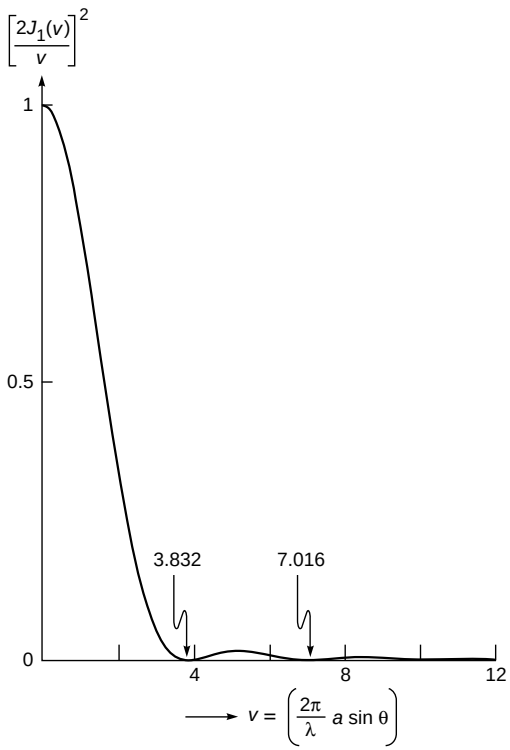


Fig. 18.11 The intensity variation associated with the Airy pattern.

or

$$\sin\theta = \frac{3.832\,\lambda}{2\pi a}, \frac{7.016\,\lambda}{2\pi a}, \ldots \quad \textbf{(23)}$$

If f represents the focal length of the convex lens, then

$$\text{Radii of dark rings} = f\tan\theta \approx \frac{3.832\,\lambda f}{2\pi a}, \frac{7.016\,\lambda f}{2\pi a}, \ldots \quad \textbf{(24)}$$

where we have assumed θ to be small so that tan θ ≈ sin θ. The Airy patterns shown in Fig. 18.9(a) and (b) correspond to a = 0.5 and 0.25 mm, respectively; both figures correspond to λ = 5000 Å and f = 20 cm. Thus

Radius of first dark ring ≈ 0.12 and 0.24 mm

corresponding to a = 0.5 and 0.25 mm, respectively. Detailed mathematical analysis shows that about 84% of the energy is contained within the first dark ring (see Sec. 19.7); thus the angular spread of the beam is approximately given by

$$\Delta\theta \approx \frac{0.61\lambda}{D} \approx \frac{\lambda}{D} \qquad \textbf{(25)}$$

where D (= 2a) represents the diameter of the aperture. Comparing Eqs. (18) and (25), we see that the angular divergence associated with the diffraction pattern can be written in the following general form:

$$\Delta\theta \sim \frac{\lambda}{\text{linear dimension of aperture}} \qquad \textbf{(26)}$$

An interesting application of the above phenomenon is shown in Fig. 18.12. A layperson would expect that to obtain greater directionality of sound waves, one should use a loudspeaker of small aperture as shown in Fig. 18.12(a); however, this will result in a greater diffraction divergence and only a small fraction of energy will reach the observer. On the other hand, if one uses a loudspeaker of larger diameter, greater directionality is achieved [see Fig. 18.12(b)].

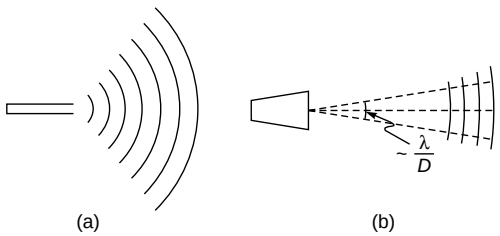


Fig. 18.12 The directionality of sound waves increases with an increase in the diameter of the speaker.

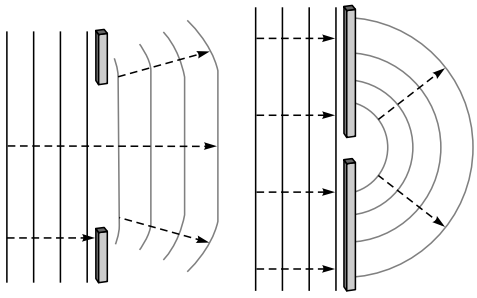


Fig. 18.13 If an obstacle with a small gap is placed in the tank, the ripples emerge in an almost semicircular pattern, the small gap acting almost as a point source. If the gap is large, however, the diffraction is much more limited. *Small*, in this context, means that the size of the obstacle is comparable to the wavelength of the ripples.

In Fig. 18.13 we have shown that if an obstacle with a small gap is placed in the tank, the ripples emerge in an almost semicircular pattern, the small gap acting almost as a point source. If the gap is large, however, the diffraction is much more limited. *Small*, in this context, means that the size of the obstacle is comparable to the wavelength of the ripples.

Example 18.3 Calculate the radii of the first two dark rings of the Fraunhofer diffraction pattern produced by a circular aperture of radius 0.02 cm at the focal plane of a convex lens of focal length 20 cm. Assume λ = 6×10^{-5} cm.

Solution: The first dark ring occurs at

$$\theta \approx \sin\theta = \frac{1.22 \times 6 \times 10^{-5}}{2 \times 0.02} \approx 1.8 \times 10^{-3} \text{ rad}$$

Thus the radius of the first dark ring is

$$\approx 20 \times 1.8 \times 10^{-3} = 3.6 \times 10^{-2} \text{ cm}$$

Similarly, the radius of the second dark ring is

$$\approx 20 \times \frac{7.016 \times 6 \times 10^{-5}}{2\pi \times 0.02} \approx 6.7 \times 10^{-2} \text{ cm}$$

18.4 DIRECTIONALITY OF LASER BEAMS

An ordinary source of light (such as a sodium lamp) radiates in all directions. On the other hand, the divergence of a laser beam is primarily due to diffraction effects. For most laser beams, the transverse amplitude distribution is approximately Gaussian; indeed just when the beam is leaving the laser (which we assume to be z = 0), the amplitude distribution can be assumed to be given by

$$A(x, y) = a \exp\left(-\frac{x^2 + y^2}{w_0^2}\right) \qquad \textbf{(27)}$$

where we have assumed that the phase front is plane at z = 0. From the above equation it follows that at a distance w_0 from the z axis, the amplitude falls by a factor 1/e (i.e., the intensity reduces by a factor 1/e^2). This quantity w_0 is called the *spot size* of the beam. In Sec. 20.5 we will show that as the beam propagates in the z direction, the intensity distribution is given by

$$I(x, y, z) = \frac{I_0}{1 + \gamma^2} \exp\left[-\frac{2(x^2 + y^2)}{w^2(z)}\right] \qquad \textbf{(28)}$$

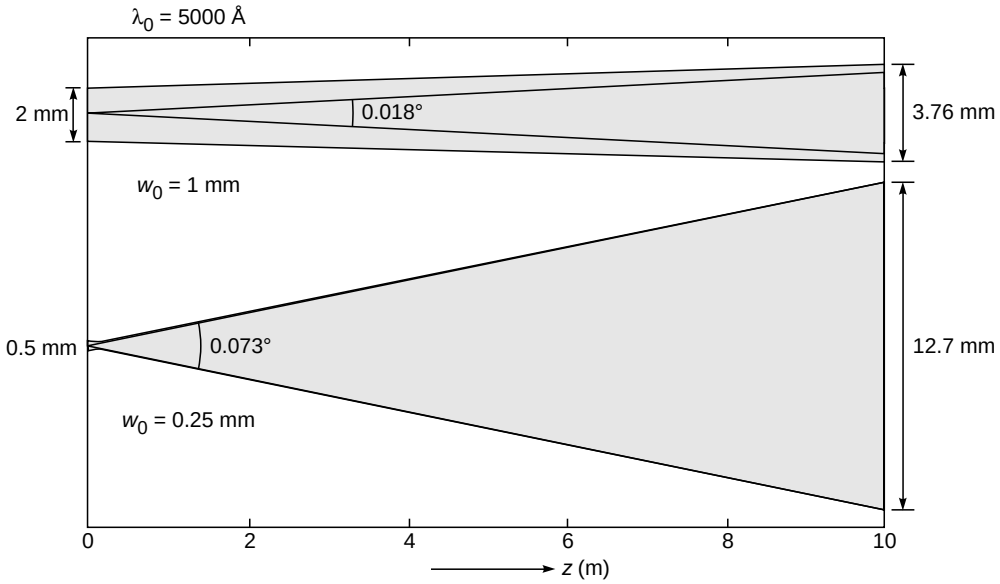


Fig. 18.14 Diffraction divergence of a Gaussian beam whose phase front is plane at $z = 0$. The figure shows the increase in the diffraction divergence as the initial spot size is decreased from 1 to 0.25 mm; the wavelength is assumed to be 5000 Å.

where

$$\gamma = \frac{\lambda z}{\pi w_0^2}$$

$$w(z) = w_0 (1 + \gamma^2)^{1/2}$$

$$= w_0 \left(1 + \frac{\lambda^2 z^2}{\pi^2 w_0^4}\right)^{1/2} \quad \textbf{(29)}$$

Thus the transverse intensity distribution remains Gaussian with the beam width increasing with z. For large values of z ($>> w_0^2/\lambda$), we obtain

$$w(z) \approx w_0 \frac{\lambda z}{\pi w_0^2} = \frac{\lambda z}{\pi w_0} \quad \textbf{(30)}$$

which shows that the width increases linearly with z. We define the diffraction angle as

$$\tan \theta = \frac{w(z)}{z} \approx \frac{\lambda}{\pi w_0} \quad \textbf{(31)}$$

showing that the rate of increase in the width is proportional to the wavelength and inversely proportional to the initial width of the beam; the above equation is consistent with Eq. (26). To get some numerical values, we assume $\lambda = 0.5\ \mu$m. Then for $w_0 = 1$ mm

$$2\theta \approx 0.018° \quad \text{and} \quad w \approx 1.88 \text{ mm} \quad \text{at } z = 10 \text{ m}$$

Similarly, for $w_0 = 0.25$ mm,

$$2\theta \approx 0.073° \quad \text{and} \quad w \approx 6.35 \text{ mm} \quad \text{at } z = 2 \text{ m}$$

(see Fig. 18.14). Notice that θ increases with a decrease in w_0 (the smaller the size of the aperture, the greater the diffraction). From Eq. (31) we find the following:

1. For a given value of λ_0, θ increases with a decrease in the value of w_0, implying that the smaller the initial spot size of the beam, the greater the diffraction divergence.
2. For a given value of w_0 the value of θ (and hence the diffraction divergence) decreases with decrease in the value of λ_0. Indeed, as $\lambda_0 \to 0$, there is no spreading of the beam, and we have what is known as the geometrical optics limit.

In Fig. 18.15 we have shown a laser beam propagating through air. Notice the nondivergence of the beam. We give a more

Fig. 18.15 A laser beam. Notice the nondivergence of the beam. A color photograph appears in the insert at the back of the book.

detailed discussion in Sec. 20.5. From Eq. (27) one can readily show that

$$\int_{-\infty}^{+\infty}\int I(x, y, z)\, dx\, dy = \frac{\pi w_0^2}{2} I_0$$

which is independent of z. This is to be expected, as the total energy crossing the entire xy plane will not change with z.

Example 18.4 The output from a single-mode fiber operating at the He-Ne laser wavelength (λ_0 = 0.6328 μm) is approximately Gaussian with w_0 = 5 μm. Thus, the corresponding divergence is

$$\theta \approx \tan^{-1}\frac{\lambda_0}{\pi w_0} \approx 2.3°$$

Thus, if a screen is placed at a distance of about 50 cm from the fiber, the radius of the beam is about 2 cm.

A beam is said to be *diffraction-limited* if it diverges only due to diffraction. Usually laser beams are diffraction-limited. On the other hand, if we have a tiny filament at the focal plane of a lens, the beam will diverge primarily due to the finite size of the filament (see Fig. 18.16). The angular spread of the beam is given by (see Fig. 18.16)

$$\Delta\theta \approx \frac{l}{f} \qquad \textbf{(32)}$$

where l is the length of the filament and f is the focal length of the lens. If the linear dimension of the filament is about 2 mm (placed on the focal plane of a convex lens of focal length 10 cm), then the angular divergence of the beam (due to the finite size of the filament) is approximately given by

$$\Delta\theta \approx \frac{2\text{ mm}}{100\text{ mm}} = 0.02\text{ rad}$$

If the diameter of the aperture of the lens is 5 cm, then the angular divergence due to diffraction is

$$\Delta\theta \approx \frac{\lambda}{D} \approx \frac{5\times10^{-5}\text{ cm}}{5\text{ cm}} = 0.00001\text{ rad}$$

which is much much smaller than the angular divergence of the beam due to the finite size of the filament. Only if the size of the filament is smaller than 10^{-3} mm will the beam divergence be determined by diffraction. Thus for most practical sources, the beam divergence is due to the finite size of the filament rather than to diffraction.

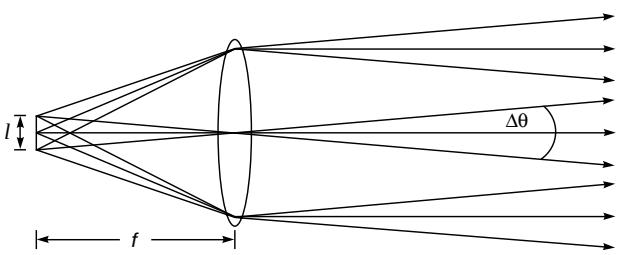


Fig. 18.16 A filament placed at the focal plane of a convex lens.

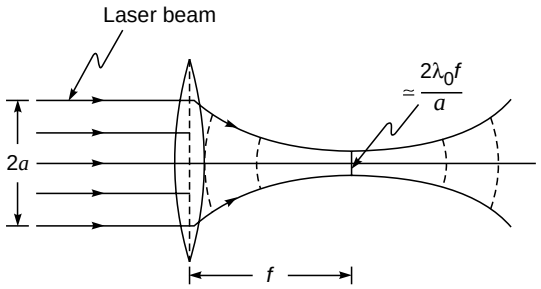


Fig. 18.17 If a truncated plane wave (of diameter 2a) is incident on an aberrationless lens of focal length f, then the wave emerging from the lens will get focused to the spot of radius $\approx \lambda_0 f/a$; the area of the focused spot size is $\approx \pi(\lambda_0 f/a)^2$.

18.4.1 Focusing of Laser Beams

As mentioned earlier, laser beams are usually diffraction-limited. If such a diffraction-limited beam is allowed to fall on a convex lens, then

$$\text{Radius of focused spot} \approx \frac{\lambda_0 f}{a} \qquad \textbf{(33)}$$

(see Fig. 18.17). In Eq. (33), f represents the focal length of the lens and a represents the beam radius or the radius of the aperture of the lens (whichever is smaller). Thus

$$\text{Area of focused spot } A_m \approx \pi\left(\frac{\lambda_0 f}{a}\right)^2$$

We illustrate the effects of this focusing through some examples.

Example 18.5 We consider a 2 mW laser beam ($\lambda_0 \approx 6 \times 10^{-5}$ cm) incident on the eye whose focal length is given by $f \approx$ 2.5 cm. If the pupil diameter (= 2a) is taken to be 2 mm, then

$$\text{Area of focused spot } A = \pi\left(\frac{\lambda_0 f}{a}\right)^2 \approx 7\times10^{-6}\text{ cm}^2$$

On the retina, the intensity will be approximately given by

$$I \approx \frac{P}{A} \approx \frac{2\times10^{-3}\text{ W}}{7\times10^{-10}\text{ m}^2} \approx 3\times10^{6}\text{ W m}^{-2}$$

Such high intensities will damage the retina! *So never look into a (seemingly innocent) low-power laser beam.*

Example 18.6 We next consider a 3 MW laser beam ($\lambda_0 \approx 6 \times 10^{-5}$ cm and beam width 2$a \approx$ 1 cm) incident on a lens of focal length 5 cm, then

$$\text{Area of focused spot } A = \pi\left(\frac{\lambda_0 f}{a}\right)^2 \approx 10^{-6}\text{ cm}^2 = 10^{-10}\text{ m}^2$$

On the focal plane of the lens, the intensity will be approximately given by

$$I \approx \frac{P}{A} \approx \frac{3\times10^{6}\text{ W}}{10^{-10}\text{ m}^2} \approx 3\times10^{16}\text{ W m}^{-2}$$

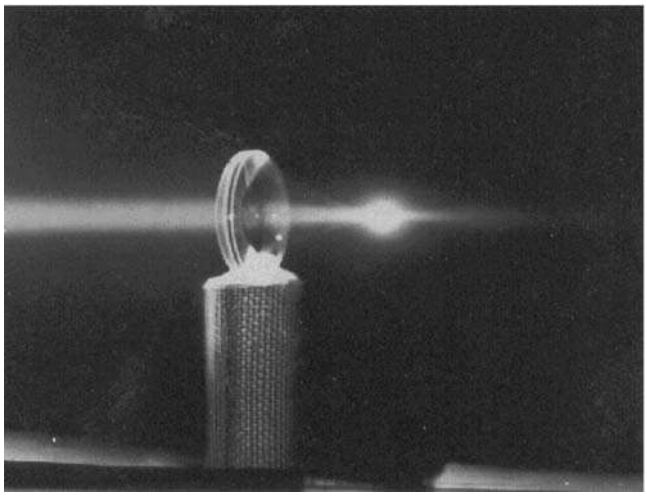


Fig. 18.18 Focusing of a 3 MW peak power pulsed ruby laser beam. At the focus, the electric field strengths are of the order of 10^9 V m^{-1} which results in the creation of a spark in the air, (Photograph courtesy Dr. R. W. Terhune).

Now, the intensity of the beam is related to the electric field amplitude E_0 through the relation [see Eq. (78) of Chap. 23]

$$I = \tfrac{1}{2}\,\varepsilon_0 c E_0^2 \qquad \textbf{(34)}$$

where $\varepsilon_0 \approx 8.854 \times 10^{-12}$ MKS units represents the dielectric permittivity of free space and $c \approx 3 \times 10^8$ m s^{-1} represents the speed of light in free space. Substituting $I \approx 3 \times 10^{16}$ W m^{-2} in Eq. (34), we readily get

$$E_0 \approx 5 \times 10^9 \text{ V m}^{-1}$$

Such high electric fields result in the creation of spark in air (see Fig. 18.18). Thus laser beams (because of their high directionality) can be focused to extremely small regions, producing very high intensities. Such high intensities lead to numerous industrial applications of the laser such as welding, hole drilling, and cutting materials (see, e.g., Ref. 5).

In the following two examples, we will calculate the intensities (at the retina of *the eye*) when we directly view a 500 W bulb or the Sun. *Caution:* Never look into the Sun; the retina will be damaged not only because of high intensities but also because of the large ultraviolet content of the sunlight.

Example 18.7 We consider a 6 cm diameter incandescent source (such as a 500 W bulb) at a distance of about 5 m from the eye (see Fig. 18.19). We assume the pupil diameter to be about 2 mm. Thus

Area of pupil of eye $\approx \pi(1 \times 1)$ mm$^2 \approx 3 \times 10^{-6}$ m^2

$$\text{Power entering eye} \approx (500 \text{ W})\frac{\pi r^2}{4\pi R^2} \approx 5 \times 10^{-6} \text{ W}$$

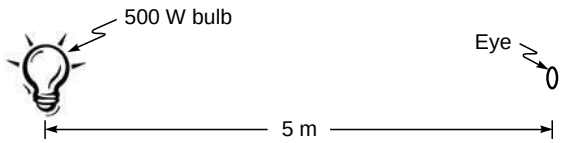


Fig. 18.19 A 500 W bulb at a distance of about 5 m from the eye.

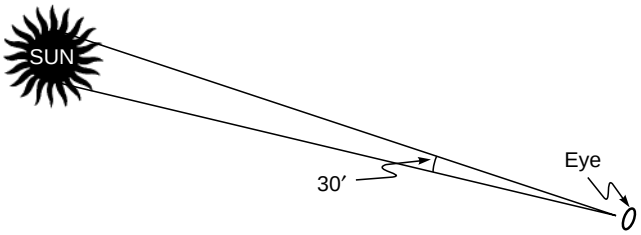


Fig. 18.20 If we look directly at the Sun, intensities as high as 130 kW m^{-2} are produced; this can damage the retina of the eye!

Radius of image = radius of source × demagnification

$$\approx \frac{0.03}{5} \times 0.025 \approx 1.5 \times 10^{-4} \text{ m}$$

where we have assumed the image to be formed at a distance of about 2.5 cm from the pupil of the eye. Thus,

$$\text{Power density in image} = \frac{5 \times 10^{-6} \text{ W}}{\pi(1.5 \times 10^{-4})^2 \text{ m}^2} \approx 70 \text{ W/m}^2$$

Example 18.8 We next calculate the intensity at the retina if we are directly looking at the Sun (see Fig. 18.20). Now

Intensity of solar energy on Earth ≈ 1.35 kW m^{-2}

Thus

Energy entering eye $\approx 1.35 \times 10^3 \times \pi \times 10^{-6} \approx 4$ mW

The Sun subtends about 0.5° on the Earth. Thus

Diameter of image of Sun

$$\approx 0.5 \times \frac{\pi}{180} \times 25 \approx 0.2 \text{ mm}$$

$$= 2 \times 10^{-4} \text{ m}$$

and

$$\text{Power density in image} \approx \frac{4 \times 10^{-3} \text{ W}}{\pi \times (1 \times 10^{-4})^2 \text{ m}^2}$$

$$\approx 130 \text{ kW m}^{-2}$$

To summarize, a 2 mW diffraction-limited laser beam incident on the eye can produce an intensity of about 10^6 W m^{-2} at the retina—this would certainly damage the retina. Thus, whereas it is quite safe to look at a 500 W bulb, it is very dangerous to

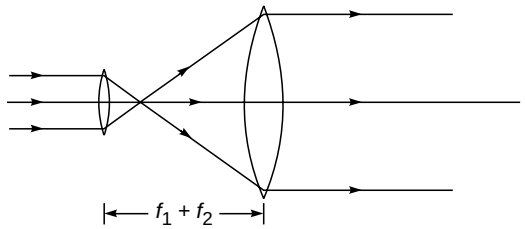


Fig. 18.21 Two convex lenses separated by a distance equal to the sum of their focal lengths act as a beam expander.

look directly into a 2 mW laser beam. Indeed, because a laser beam can be focused to very narrow areas, it has found important applications in areas such as eye surgery and welding.

From the above discussion it immediately follows that the greater the radius of the beam, the smaller the size of the focused spot and hence the greater the intensity at the focused spot. Indeed, one may use a beam expander (see Fig. 18.21) to produce a beam of greater size and hence a smaller focused spot size. However, after the focused spot, the beam would have a greater divergence and would therefore expand within a very short distance. One usually defines a *depth of focus* as the distance over which the intensity of the beam (on the axis) decreases by a certain factor of the value at the focal point. Thus a small focused spot leads to a small depth of focus. The intensity distribution at the focal plane of the lens is given by Eq. (19) where the parameter v is given by Eq. (21). On the other hand, the intensity along the axis is given by

$$I = I_0 \left[\frac{\sin (w/4)}{w/4} \right]^2 \tag{35}$$

where

$$w = \frac{2\pi}{\lambda} \left(\frac{a}{f} \right)^2 z \tag{36}$$

and $z = 0$ represents the focal plane.[4]

The intensity would drop by about 20% at

$$z \approx \pm 0.5\lambda \, (f/a)^2 \tag{37}$$

which is usually referred to as the depth of the focus or focal tolerance. Notice that larger the value of a, the smaller the focal tolerance. For $\lambda \approx 6 \times 10^{-5}$ cm, $f \approx 10$ cm, and $a \approx 1$ cm, the focal tolerance is about 3×10^{-3} cm.

18.5 LIMIT OF RESOLUTION

Consider two point sources, such as stars (so that we can consider plane waves entering the aperture) being focused by a telescope objective of diameter D (see Fig. 18.22). As discussed in Sec. 18.4, the system can be thought of as being equivalent to a circular aperture of diameter D, followed by a converging lens of focal length f, as shown in Fig. 18.8. As such, each point source will produce its Airy pattern as schematically shown in Fig. 18.22. The diameters of the Airy rings will be determined by the diameter of the objective, its focal length, and the wavelength of light (see Example 18.3).

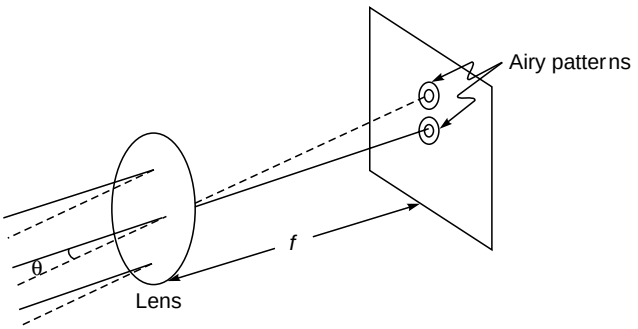


Fig. 18.22 The image of two distant objects on the focal plane of a convex lens. If the diffraction patterns are well separated, they are said to be resolved.

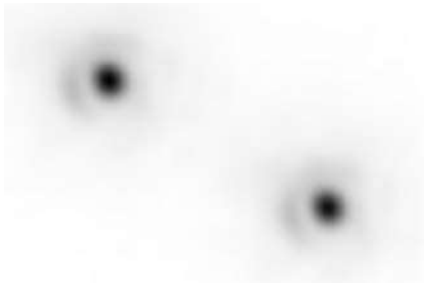


Fig. 18.23 Image of the binary star Zeta Bootis by a 2.56 m telescope aperture; the Airy disk around each of the stars can be seen [Image by Bob Tubbs and collaborators, used with permission from Dr. Tubbs].

In Fig. 18.22 the Airy patterns are shown to be quite far away from each other, and therefore, the two objects are said to be well resolved. Since the radius of the first ring is $1.22\lambda \, f/D$, the Airy patterns will overlap more for smaller values of D; hence for better resolution one requires a larger diameter of the objective. It is for this reason that a telescope is usually characterized by the diameter of the objective; for example, a 40 in. telescope implies that the diameter of the objective is 40 in. In Fig. 18.23 we have shown the image of the binary star Zeta Bootis by a 2.56 m telescope aperture; the Airy disk around each of the stars can be seen.

[4] The derivation of the formulas has been given at many places; see, e.g., Sec. 6.5 of Ref. 6.

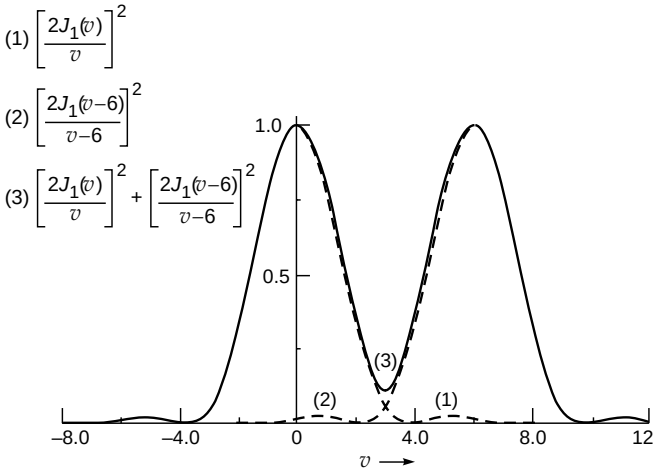


Fig. 18.24 The dashed curves correspond to the intensity distribution produced independently by two distant point objects having an angular separation of $6\lambda/(\pi D)$. The resultant intensity distribution (shown as a solid curve) has two well defined peaks and the objects are well resolved.

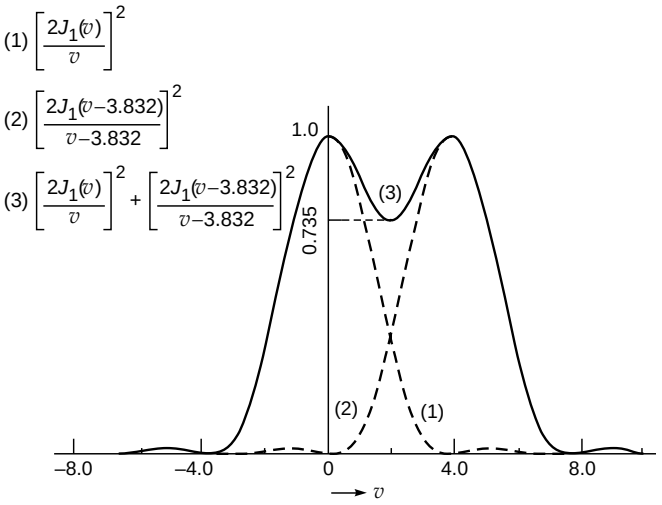


Fig. 18.25 The dashed curves correspond to the intensity distribution produced independently by two distant point objects having an angular separation of $1.22\lambda/D$ and according to the Rayleigh criterion, the objects are just resolved.

In Figs. 18.24, 18.25, and 18.26, we have plotted the independent intensity distributions and their resultant produced by two distant objects for various angular separations; in each case we have assumed that the two sources produce the same intensity at their respective central spots. Obviously, the resultant intensity distributions are quite complicated (see Fig. 18.27); what we have plotted in Figs. 18.24, 18.25,

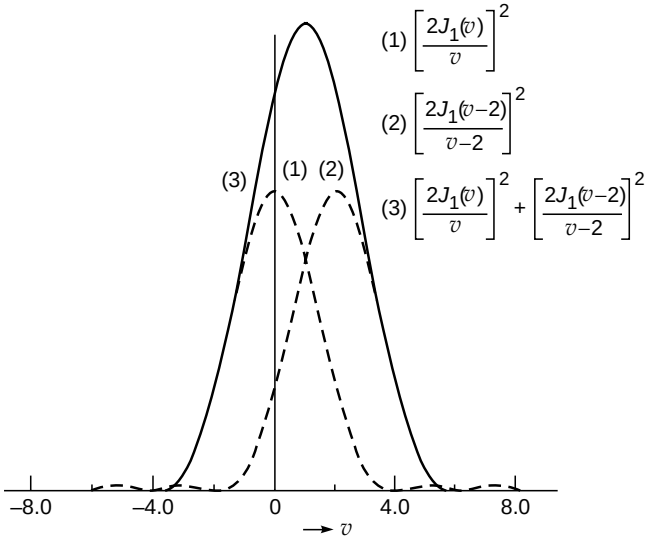


Fig. 18.26 The dashed curves correspond to the intensity distribution produced independently by two distant point objects having an angular separation of $2\lambda/(\pi D)$. The resultant intensity distribution (shown as a solid curve) has only one peak, and hence the objects are unresolved.

and 18.26 are the intensity distributions on the line joining the two centers of the Airy patterns; since the point sources are independent sources, their intensity distributions (Airy patterns) will add. If we choose this line as the x axis, then the parameter v in Figs. 18.24, 18.25, and 18.26 is given by

$$v = \frac{2\pi a}{\lambda f} x \qquad \textbf{(38)}$$

Now, the intensity distributions given in Fig. 18.24 correspond to two distant point objects having an angular separation of $6\lambda/\pi D$, and as can be seen, the two images are clearly resolved. Figure 18.26 corresponds to

$$\Delta\theta \simeq \frac{2}{\pi}\frac{\lambda}{D} \qquad \textbf{(39)}$$

and as can be seen, the resultant intensity distribution has only one peak and therefore the two points cannot be resolved at all. Finally, if the angular separation of the two objects is $1.22\lambda/D$, then the central spot of one pattern falls on the first minimum of the second and the objects are said to be just resolved. This criterion of limit of resolution is called the Rayleigh criterion of resolution, and the intensity distribution corresponding to this is plotted in Fig. 18.25. The actual diffraction patterns are shown in Fig. 18.27.

To get a numerical appreciation of the above results, we consider a telescope objective whose diameter and focal length are 5 and 30 cm, respectively. Assuming the light wavelength to be

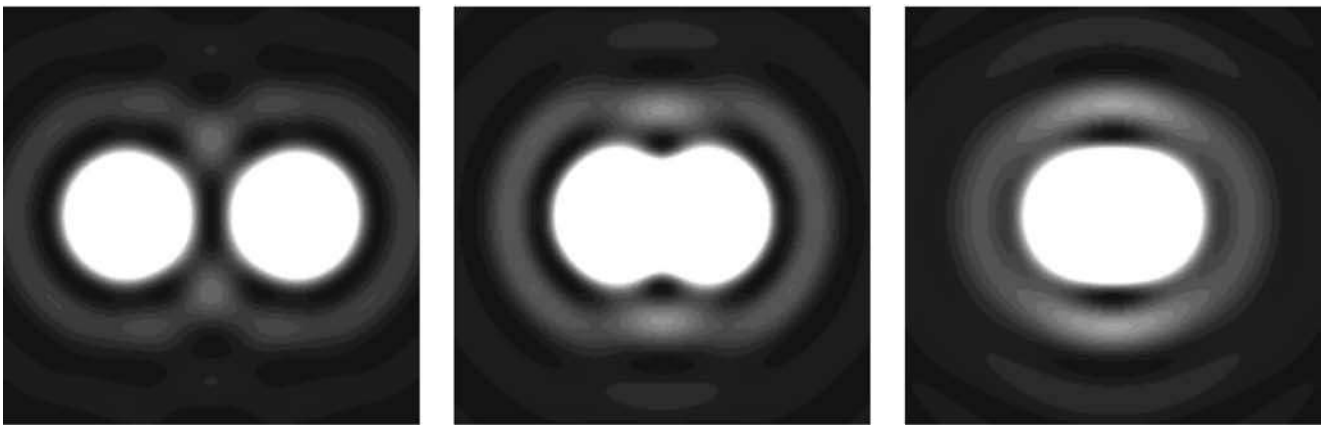


Fig. 18.27 Computer-generated intensity distributions corresponding to two point sources when they are (a) well resolved, (b) just resolved, and (c) unresolved.

6×10^{-5} cm, one finds that the minimum angular separation of two distant objects which can just be resolved will be

$$\frac{1.22\lambda}{D} = \frac{1.22 \times 6 \times 10^{-5}}{5} \simeq 1.5 \times 10^{-5} \text{ rad}$$

Further, the radius of the first dark ring (of the Airy pattern) will be

$$\frac{1.22\lambda}{D} \times \text{focal length} = \frac{1.22 \times 6 \times 10^{-5}}{5} \times 30$$

$$\simeq 4.5 \times 10^{-4} \text{ cm}$$

It is immediately obvious that the larger the diameter of the objective, the better its resolving power. For example, the diameter of the largest telescope objective is about 80 in., and the corresponding angular separation of the objects that it can resolve is $\simeq$ 0.07 second of arc. This very low limit of resolution is never achieved in ground-based telescopes due to the turbulence of the atmosphere. However, a larger aperture still provides a larger light-gathering power and hence the ability to see deeper in space.

If we assume that the angular resolution of the human eye is primarily due to diffraction effects, then it will be given by

$$\Delta\theta \sim \frac{\lambda}{D} \approx \frac{6 \times 10^{-5}}{2 \times 10^{-1}} = 3 \times 10^{-4} \text{ rad} \qquad \textbf{(40)}$$

where we have assumed the pupil diameter to be 2 mm. Thus, at a distance of 20 m, the eye should be able to resolve two points which are separated by a distance

$$3 \times 10^{-4} \times 20 = 6 \times 10^{-3} \text{ m} = 6 \text{ mm}$$

One can indeed verify that this result is qualitatively valid by finding the distance at which the millimeter scale will become blurred.

In the above discussion we have assumed that the two object points produce identical (but displaced) Airy patterns. If that is not the case, then the two central maxima will have different intensities; accordingly one has to set up a modified criterion for the limit of resolution such that the two maxima stand out.

18.5.1 Resolving Power of a Microscope

We next consider the resolving power of a microscope objective of diameter D as shown in Fig. 18.28. Let P and Q represent two closely spaced self-luminous point objects which are to be viewed through the microscope. Assuming the absence of any geometrical aberrations, rays emanating from points P and Q will produce spherical wave fronts (after refraction through the lens) which will form Airy patterns around their paraxial image points P' and Q'. For points P and Q to be just resolved, point Q' should lie on the first dark ring surrounding point P', and therefore we must have

$$\sin\alpha' \approx \frac{1.22\lambda}{D} = \frac{1.22\lambda_0}{n'D} \qquad \textbf{(41)}$$

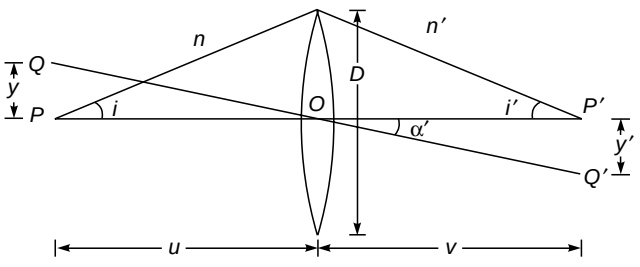


Fig. 18.28 The resolving power of a microscope objective.

where n and n' represent the refractive indices of the object and image spaces, and λ_0 and λ ($= \lambda_0/n'$) represent the wavelength of light in free space and in the medium of refractive index n', respectively. The angle α' is defined in Fig. 18.28, and we have

$$\sin \alpha' \approx \frac{y'}{OP'} = \frac{y' \tan i'}{D/2} \approx \frac{y' \sin i'}{D/2} \qquad \textbf{(42)}$$

where we have assumed $\sin i' \approx \tan i'$; this is justified since the image distance OP' is large compared to D. Using Eqs. (41) and (42), we get

$$y' \approx \frac{0.61\,\lambda_0}{n' \sin i'}$$

If we now use the sine law $n'y' \sin i' = ny \sin i$ [see Eq. (39) of Chap. 4], we get

$$y \approx \frac{0.61\lambda_0}{n \sin i} \qquad \textbf{(43)}$$

which represents the smallest distance that the microscope can resolve. The quantity $n \sin i$ is the numerical aperture of the optical system, and the resolving power increases with an increase in the numerical aperture. For this reason in some microscopes the space between the object and the objective is filled with an oil—and they are referred to as *oil immersion objectives*. Equation (43) also tells us that the resolving power increases with a decrease in λ. As such, one often uses blue light (or even ultraviolet light) for the illumination of the object. For example, in an electron microscope the de Broglie wavelength of electrons accelerated to 100 keV is about 0.03×10^{-8} cm, and therefore such a microscope has a very high resolving power.

In the above analysis, we have assumed that the two object points are self-luminous so that the intensities can be added. However, in actual practice, the objects are illuminated by the same source, and therefore, in general, there is some phase relationship between the waves emanating from the two object points. For such a case the intensities will not be strictly additive (see Sec. 14.6); nevertheless Eq. (43) will give the correct order for the limit of resolution.

18.6 TWO-SLIT FRAUNHOFER DIFFRACTION PATTERN

In Sec. 18.3 we studied the Fraunhofer diffraction pattern produced by a slit of width b and found that the intensity distribution consisted of maxima and minima. In this section we will study the Fraunhofer diffraction pattern produced by two parallel slits (each of width b) separated by a distance d. We will find that the resultant intensity distribution is a product of the single-slit diffraction pattern and the interference pattern produced by two point sources separated by a distance d.

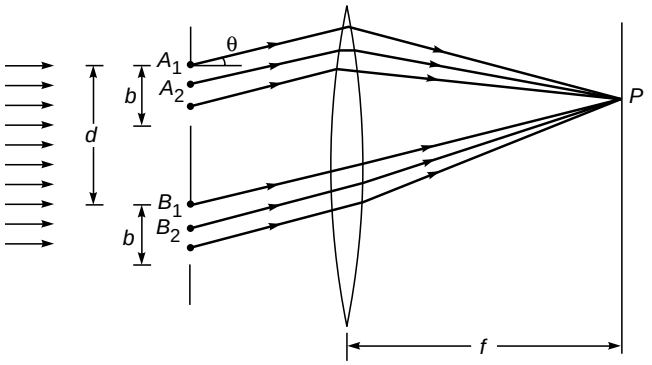


Fig. 18.29 Fraunhofer diffraction of a plane wave incident normally on a double slit.

To calculate the diffraction pattern, we use a method similar to that used for the case of a single slit. We assume that the slits consist of a large number of equally spaced point sources and that each point on the slit is a source of Huygens' secondary wavelets. Let the point sources be at $A_1, A_2, A_3, \ldots$ (in the first slit) and at $B_1, B_2, B_3, \ldots$ (in the second slit) (see Fig. 18.29). As before, we assume that the distance between two consecutive points in either of the slits is Δ. If the diffracted rays make an angle θ with the normal to the plane of the slits, then the path difference between the disturbances reaching point P from two consecutive points in a slit will be $\Delta \sin \theta$. The field produced by the first slit at point P will, therefore, be given by [see Eq. (9)]

$$E_1 = A \frac{\sin \beta}{\beta} \cos(\omega t - \beta)$$

Similarly, the second slit will produce a field

$$E_2 = A \frac{\sin \beta}{\beta} \cos(\omega t - \beta - \Phi_1)$$

at point P, where

$$\Phi_1 = \frac{2\pi}{\lambda} d \sin\theta$$

represents the phase difference between the disturbances (reaching point P) from two corresponding points on the slits; by corresponding points we imply pairs of points such as (A_1, B_1), (A_2, B_2), ... which are separated by a distance d. Hence the resultant field will be

$$E = E_1 + E_2$$
$$= A \frac{\sin \beta}{\beta} [\cos(\omega t - \beta) + \cos(\omega t - \beta - \Phi_1)]$$

which represents the interference of two waves, each of amplitude A $\sin\beta/\beta$ and differing in phase by Φ_1. The above equation can be rewritten in the form

$$E = 2A\frac{\sin\beta}{\beta}\cos\gamma\cos\left(\omega t - \beta - \frac{1}{2}\Phi_1\right)$$

where

$$\gamma = \frac{\Phi_1}{2} = \frac{\pi}{\lambda} d\sin\theta \qquad \textbf{(44)}$$

The intensity distribution will be of the form

$$I = 4I_0\frac{\sin^2\beta}{\beta^2}\cos^2\gamma \qquad \textbf{(45)}$$

where I_0 $(\sin^2\beta)/\beta^2$ represents the intensity distribution produced by one of the slits. As can be seen, the intensity distribution is a product of two terms; the first term $(\sin^2\beta)/\beta^2$ represents the diffraction pattern produced by a single slit of width b, and the second term $\cos^2\gamma$ represents the interference pattern produced by two point sources separated by a distance d. Indeed, if the slit widths are very small [so that there is almost no variation of the $(\sin^2\beta)/\beta^2$ term with θ], then one simply obtains Young's interference pattern (see Sec. 14.6).

In Fig. 18.30, we have shown the two-slit diffraction patterns corresponding to d = 0, 0.0176, 0.035, and 0.070 cm with b = 0.0088 cm and $\lambda = 6.328\times10^{-5}$ cm. The intensity distributions as predicted by Eq. (45) are shown in Figs. 18.31 and 18.32.

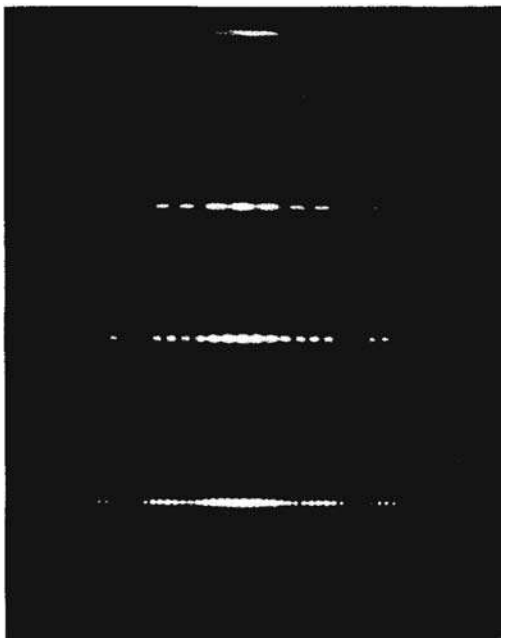


Fig. 18.30 The double-slit Fraunhofer diffraction pattern corresponding to b = 0.0088 cm and $\lambda = 6.328\times10^{-5}$ cm. The values of d are 0, 0.0176, 0.035, and 0.070 cm, respectively [After Ref. 17; used with permission].

18.6.1 Positions of Maxima and Minima

Equation (45) tells us that the intensity is zero wherever

$$\beta = \pi, 2\pi, 3\pi, \ldots$$

or when

$$\gamma = \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \ldots$$

The corresponding angles of diffraction are given by the following equations:

$$\left.\begin{aligned} b\sin\theta &= m\lambda \qquad m = 1, 2, 3, \ldots \\ d\sin\theta &= \left(n + \tfrac{1}{2}\right)\lambda \qquad n = 1, 2, 3, \ldots \end{aligned}\right\} \qquad \textbf{(46)}$$

The interference maxima occur when

$$\gamma = 0, \pi, 2\pi, \ldots$$

or when

$$d\sin\theta = 0, \lambda, 2\lambda, 3\lambda, \ldots \qquad \textbf{(47)}$$

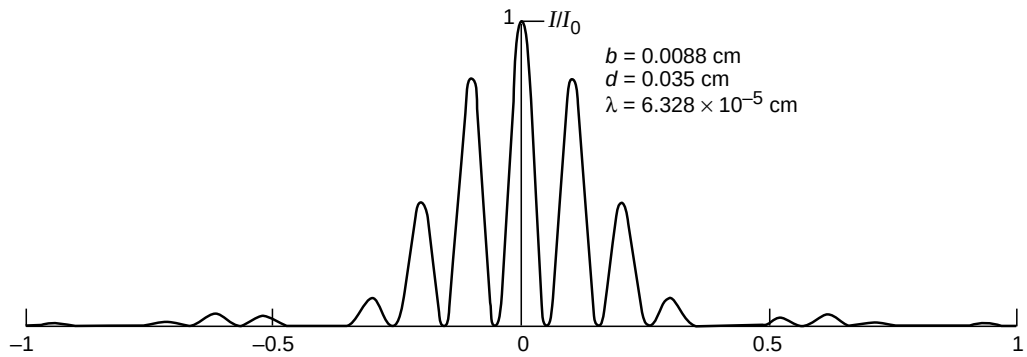


Fig. 18.31 The double-slit intensity distribution as predicted by Eq. (45) corresponding to b = 0.0088 cm, $\lambda = 6.328 \times 10^{-5}$ cm, and d = 0.035 cm.

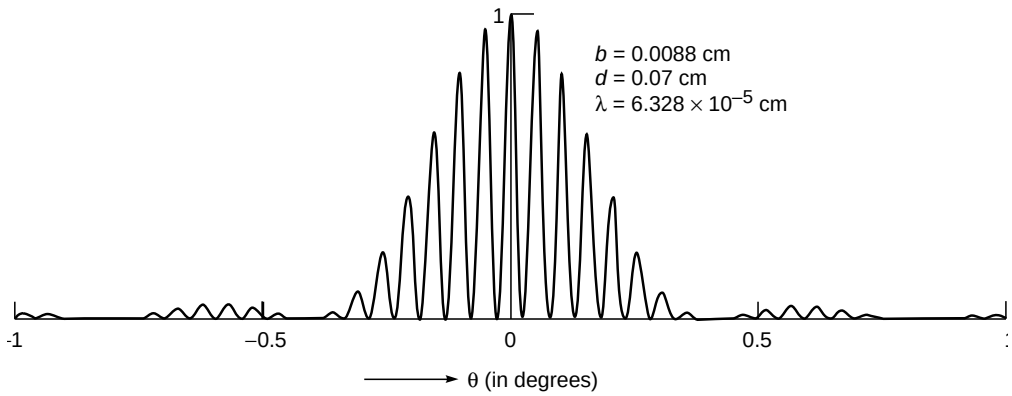


Fig. 18.32 The double-slit intensity distribution as predicted by Eq. (45) corresponding to b = 0.0088 cm, λ = 6.328 x 10^{-5} cm, and d = 0.07 cm.

The actual positions of the maxima will approximately occur at the above angles provided the variation of the diffraction term is not too rapid. Further, a maximum may not occur at all if θ corresponds to a diffraction minimum, i.e., if $b \sin\theta = \lambda, 2\lambda, 3\lambda, \ldots$. These are usually referred to as missing orders. For example, in Fig. 18.31 we can see that for b = 0.0088 cm, the interference maxima are extremely weak around $\theta \simeq 0.41°$; this is so because at

$$\theta = \sin^{-1}\left(\frac{\lambda}{b}\right)$$

$$= \sin^{-1}\left(\frac{6.328\times10^{-5}}{8.8\times10^{-3}}\right) = \sin^{-1}(7.19\times10^{-3})$$

$$\simeq 0.00719 \text{ rad}$$

$$\simeq 0.412°$$

the first minimum of the diffraction term occurs.

Example 18.9 Consider the case when $b = 8.8 \times 10^{-3}$ cm, $d = 7.0 \times 10^{-2}$ cm, and $\lambda = 6.328 \times 10^{-5}$ cm (see Fig. 18.32). How many interference minima will occur between the two diffraction minima on either side of the central maximum? In the experimental arrangement corresponding to Fig. 18.30 the screen was placed at a distance of 15 ft. Calculate the fringe width.

Solution: The interference minima will occur when Eq. (46) is satisfied, i.e., when

$$\sin\theta = \left(n+\frac{1}{2}\right)\frac{\lambda}{d} = 0.904\times10^{-3}\left(n+\frac{1}{2}\right)$$

$$n = 0, 1, 2, \ldots$$

$$= 0.452\times10^{-3}, 1.356\times10^{-3}, 2.260\times10^{-3},$$
$$3.164\times10^{-3}, 4.068\times10^{-3}, 4.972\times10^{-3},$$
$$5.876\times10^{-3}, 6.780\times10^{-3}$$

Thus there will be 16 minima between the two first-order diffraction minima.

The angular separation between two interference maxima is approximately given by [see Eq. (47)]

$$\Delta\theta \simeq \frac{\lambda}{d} = 0.904\times10^{-4}$$

Thus the fringe width is

$$15 \times 12 \times 2.54 \times 0.904 \times 10^{-4} \simeq 0.0413 \text{ cm}$$

18.7 *N*-SLIT FRAUNHOFER DIFFRACTION PATTERN

We next consider the diffraction pattern produced by N parallel slits, each of width b; the distance between two consecutive slits is assumed to be d.

As before, we assume that each slit consists of n equally spaced point sources with spacing Δ (see Fig. 18.33). Thus the field at an arbitrary point P will essentially be a sum of N terms:

$$E = A\frac{\sin\beta}{\beta}\cos(\omega t-\beta) + A\frac{\sin\beta}{\beta}\cos(\omega t-\beta-\Phi_1)$$

$$+\cdots+A\frac{\sin\beta}{\beta}\cos[\omega t-\beta-(N-1)\Phi_1] \qquad \textbf{(48)}$$

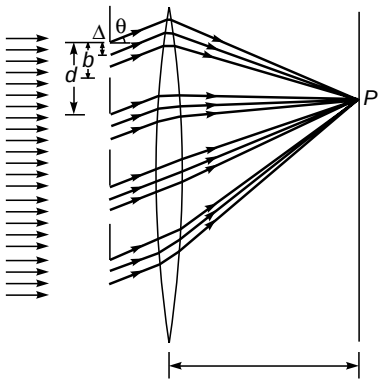


Fig. 18.33 Fraunhofer diffraction of a plane wave incident normally on a multiple slit.

where the first term represents the amplitude produced by the first slit, the second term by the second slit, etc. and the various symbols have the same meaning as in Sec. 18.5. Rewriting Eq. (48), we get

$$E = \frac{A \sin \beta}{\beta} \{\cos(\omega t - \beta) + \cos(\omega t - \beta + \Phi_1) + \cdots + \cos[\omega t - \beta - (N-1)\Phi_1]\}$$

$$= \frac{A \sin \beta}{\beta} \frac{\sin N\gamma}{\sin \gamma} \cos\left[\omega t - \beta - \frac{1}{2}(N-1)\Phi_1\right] \qquad \textbf{(49)}$$

where

$$\gamma = \frac{\Phi_1}{2} = \frac{\pi}{\lambda} d \sin \theta$$

The corresponding intensity distribution will be

$$I = I_0 \frac{\sin^2 \beta}{\beta^2} \frac{\sin^2 N\gamma}{\sin^2 \gamma} \qquad \textbf{(50)}$$

where $I_0 (\sin^2\beta)/\beta^2$ represents the intensity distribution produced by a single slit. As can be seen, the intensity distribution is a product of two terms; the first term $(\sin^2 \beta)/\beta^2$ represents the diffraction pattern produced by a single slit, and the second term $(\sin^2 N\gamma)/\sin^2 \gamma$ represents the interference pattern produced by N equally spaced point sources. For $N = 1$, Eq. (50) reduces to the single-slit diffraction pattern [see Eq. (10)] and for $N = 2$, to the double-slit diffraction pattern [see Eq. (45)]. In Fig. 18.34 we have given a plot of the function

$$\frac{\sin^2 N\gamma}{\sin^2 \gamma}$$

as a function of γ for $N = 5$ and $N = 11$. One can immediately see that as the value of N becomes very large, the above function becomes very sharply peaked at $\gamma = 0, \pi, 2\pi, \ldots$. Between the two peaks, the function vanishes when

$$\gamma = \frac{p\pi}{N} \qquad p = \pm 1, \pm 2, \ldots \qquad \text{but} \qquad p \neq 0, \pm N, \pm 2N$$

which are referred to as secondary minima.

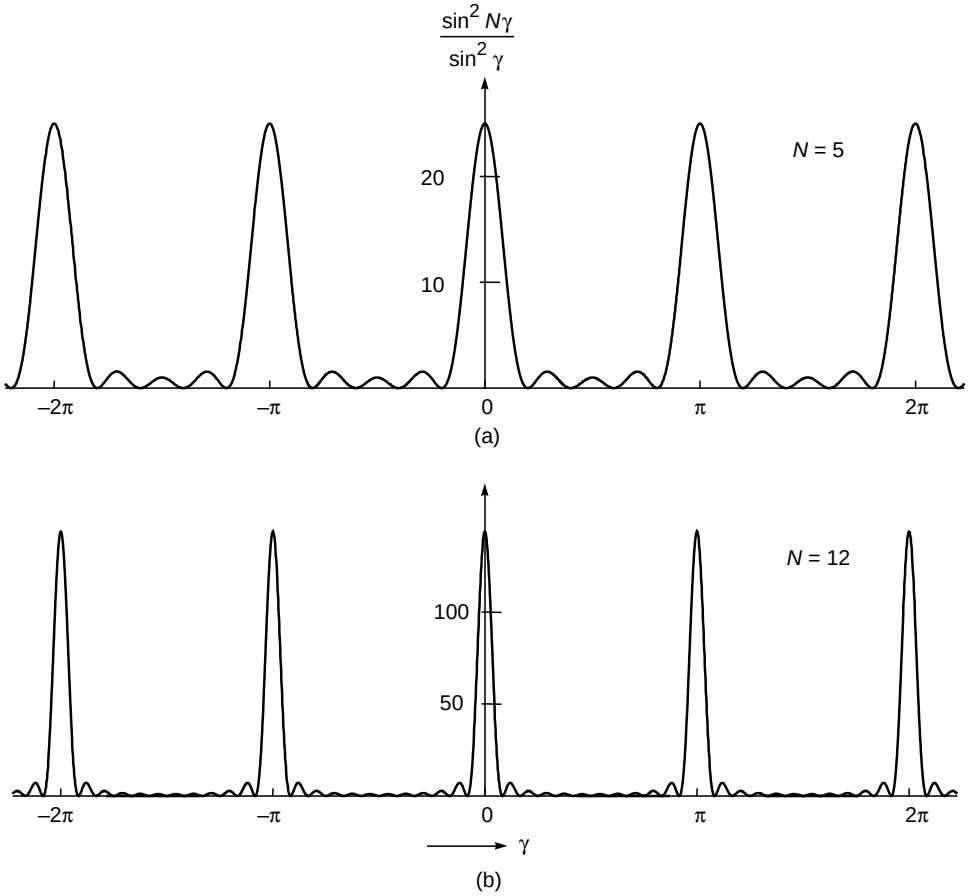


Fig. 18.34 The variation of the function $(\sin^2 N\gamma)/\sin^2\gamma$ with γ for $N = 5$ and 12. As N becomes larger, the function becomes more and more sharply peaked at $\gamma = 0, \pm\pi, \pm 2\pi, \pm 3\pi, \ldots$.

18.7.1 Positions of Maxima and Minima

When the value of N is very large, one obtains intense maxima at $\gamma \simeq m\pi$, i.e., when

$$d \sin \theta = m\lambda \qquad m = 0, 1, 2, \ldots \tag{51}$$

This can be easily seen by noting that

$$\lim_{\gamma \to m\pi} \frac{\sin N\gamma}{\sin \gamma} = \lim_{\gamma \to m\pi} \frac{N \cos N\gamma}{\cos \gamma} = \pm N$$

thus, the resultant amplitude and the corresponding intensity distributions are given by

$$E = N \frac{A \sin \beta}{\beta} \tag{52}$$

and

$$I = N^2 I_0 \frac{\sin^2 \beta}{\beta^2} \tag{53}$$

where

$$\beta = \frac{\pi b \sin \theta}{\lambda} = \frac{\pi b}{\lambda} \frac{m\lambda}{d} = \frac{\pi b m}{d} \tag{54}$$

Such maxima are known as principal maxima. Physically, at these maxima the fields produced by each of the slits are in phase, and therefore, they add and the resultant field is N times the field produced by each of the slits. Consequently, the intensity has a large value unless $(\sin^2)\beta/\beta^2$ itself is very small. Since $|\sin \theta| \leq 1$, m cannot be greater than d/λ [see Eq. (51)]; thus, there will only be a finite number of principal maxima.

From Eq. (50) it can be easily seen that the intensity is zero when either

$$b \sin \theta = n\lambda \qquad n = 1, 2, 3, \ldots \tag{55}$$

or

$$N\gamma = p\pi \qquad p \neq N, 2N, \ldots \tag{56}$$

Equation (55) gives us the minima corresponding to the single-slit diffraction pattern. The angles of diffraction corresponding to Eq. (56) are

$$d \sin \theta = \frac{\lambda}{N}, \frac{2\lambda}{N}, \ldots, \frac{(N-1)\lambda}{N}, \frac{(N+1)\lambda}{N}, \frac{(N+2)\lambda}{N}, \ldots, \frac{(2N-1)\lambda}{N}, \frac{(2N+1)\lambda}{N}, \frac{(2N+2)\lambda}{N}, \ldots \tag{57}$$

Thus, between two principal maxima we have $N - 1$ minima. Between two such consecutive minima the intensity has to have a maximum; these maxima are known as secondary maxima. Typical diffraction patterns for $N = 1$, 2, 3, and 4 are shown in Fig. 18.35, and the intensity distribution as predicted by Eq. (50) for $N = 4$ is shown in Fig. 18.36. When N is very large, the principal maxima will be much more intense in comparison to the secondary maxima. We mention here two points:

1. A particular principal maximum may be absent if it corresponds to the angle which also determines the minimum of the single-slit diffraction pattern. This will happen when

$$d \sin \theta = m\lambda \tag{58}$$

and

$$b \sin \theta = \lambda, 2\lambda, 3\lambda, \ldots \tag{59}$$

are satisfied simultaneously, and it is usually referred to as a missing order. Even when Eq. (59) does not hold exactly (i.e., if $b \sin \theta$ is close to an integral multiple of λ), the intensity of the corresponding principal maximum will be very weak (see, for example, Fig. 18.36 around $\theta \approx 0.8°$).

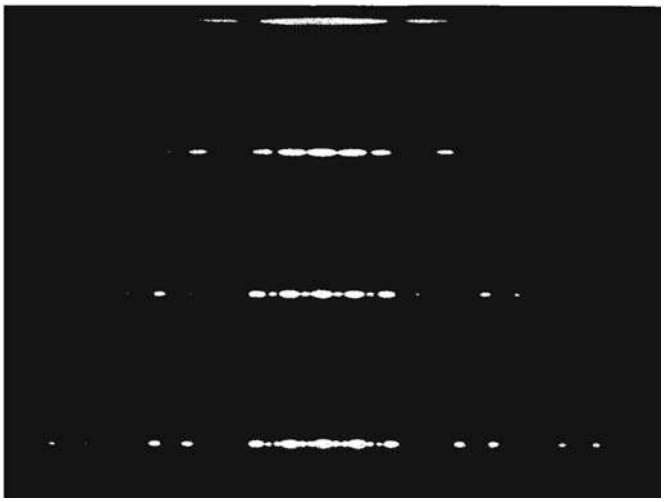


Fig. 18.35 The multiple-slit Fraunhofer diffraction patterns corresponding to b = 0.0044 cm, d = 0.0132 cm, and $\lambda = 6.328 \times 10^{-5}$ cm. The number of slits is 1, 2, 3, and 4 respectively (After Ref. 17; used with permission).

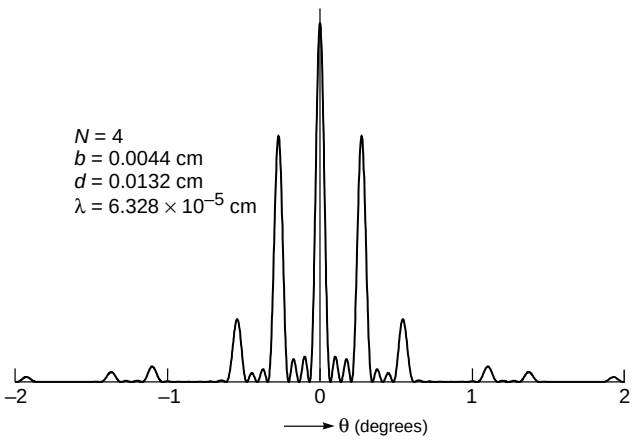


Fig. 18.36 The intensity distribution corresponding to the four-slit Fraunhofer diffraction pattern as predicted by Eq. (50) corresponding to b = 0.0044 cm, d = 0.0132 cm, and $\lambda = 6.328 \times 10^{-5}$ cm. The principal maxima occur at $\theta \approx 0.275°$, 0.55°, 0.82°, 1.1°, ... Notice the (almost) absent third order.

2. In addition to the minima predicted by Eq. (56), we will have the diffraction minima [see Eq. (55)]; however, when N is very large, the number of such minima is very small.

18.7.2 Width of the Principal Maxima

We have shown above that in the diffraction pattern produced by N slits, the mth-order principal maximum occurs at

$$d \sin \theta_m = m\lambda \qquad m = 0, 1, 2, \ldots \tag{60}$$

Further, the minima occur at the angles given by Eq. (57). If $\theta_m + \Delta\theta_{1m}$ and $\theta_m - \Delta\theta_{2m}$ represent the angles of diffraction corresponding to the first minimum on either side of the principal maximum, then $\frac{1}{2}(\Delta\theta_{1m} + \Delta\theta_{2m})$ is known as the angular half width of the mth-order principal maximum. For a large value of N, $\Delta\theta_{1m} \simeq \Delta\theta_{2m}$ which we write as $\Delta\theta_m$. Clearly,

$$d \sin (\theta_m \pm \Delta\theta_m) = m\lambda \pm \frac{\lambda}{N} \tag{61}$$

But

$$\begin{aligned} \sin (\theta_m \pm \Delta\theta_m) &= \sin \theta_m \cos \Delta\theta_m \pm \cos \theta_m \sin \Delta\theta_m \\ &\simeq \sin \theta_m \pm \Delta\theta_m \cos \theta_m \end{aligned} \tag{62}$$

Thus Eq. (61) gives

$$\Delta\theta_m \simeq \frac{\lambda}{Nd \cos \theta_m} \tag{63}$$

which shows that the principal maximum becomes sharper as N increases.

18.8 THE DIFFRACTION GRATING

In Sec. 18.6 we discussed the diffraction pattern produced by a system of parallel equidistant slits. An arrangement which essentially consists of a large number of equidistant slits is known as a diffraction grating; the corresponding diffraction pattern is known as the grating spectrum. Since the exact positions of the principal maxima in the diffraction pattern depend on the wavelength, the principal maxima corresponding to different spectral lines (associated with a source) will correspond to different angles of diffraction. Thus the grating spectrum provides us with an easily obtainable experimental setup for determination of wavelengths. From Eq. (63) we see that for narrow principal maxima (i.e., sharper spectral lines), a large value of N is required. A good-quality grating, therefore, requires a large number of slits (typically about 15,000 per inch). This is achieved by ruling grooves with a diamond point on an optically transparent sheet of material; the grooves act as opaque spaces. After each groove is ruled, the machine lifts the diamond point and moves the sheet forward for the ruling of the next groove. Since the distance between two consecutive grooves is extremely small, the movement of the sheet is obtained with the help of the rotation of a screw which drives the carriage carrying it. Further, one of the important requirements of a good-quality grating is that the lines be as equally spaced as possible; consequently, the pitch of the screw must be constant, and it was not until the manufacture of a nearly perfect screw (which was achieved by Rowland in 1882) that the problem of construction of gratings was successfully solved. Rowland's arrangement gave 14,438 lines per inch, corresponding to $d = 2.54/14{,}438 = 1.759 \times 10^{-4}$ cm. For such a grating, for $\lambda = 6 \times 10^{-5}$ cm, the maximum value of m would be 2, and, therefore, only the first two orders of the spectrum will be observed. However, for $\lambda = 5 \times 10^{-5}$ cm, the third-order spectrum will also be visible.

Commercial gratings are produced by taking the cast of an actual grating on a transparent film like that of cellulose acetate. An appropriate strength solution of cellulose acetate is poured on the ruled surface and allowed to dry to form a strong thin film, detachable from the parent grating. These impressions of a grating are preserved by mounting the film between two glass sheets. Nowadays gratings are also produced holographically, where one records the interference pattern between two plane or spherical waves (see Example 14.5). In contrast to ruled gratings, holographic gratings have a much larger number of lines per centimeter.

18.8.1 The Grating Spectrum

In Sec. 18.6 we showed that the positions of the principal maxima are given by

$$d \sin \theta = m\lambda \qquad m = 0, 1, 2, \ldots . \tag{64}$$

This relation, which is also called the grating equation, can be used to study the dependence of the angle of diffraction θ on the wavelength λ. The zeroth-order principal maximum occurs at $\theta = 0$ irrespective of the wavelength. Thus, if we are using a polychromatic source (e.g., white light), then the central maximum will be of the same color as the source itself. However, for $m \neq 0$, the angles of diffraction are different for different wavelengths, and therefore, various spectral components appear at different positions. Thus by measuring the angles of diffraction for various colors one can (knowing the value of m) determine the values of the wavelengths. The intensity is maximum for the zeroth-order spectrum (where no dispersion occurs), and it falls off as the value of m increases.

If we differentiate Eq. (64), we obtain

$$\frac{\Delta\theta}{\Delta\lambda} = \frac{m}{d\cos\theta} \qquad \textbf{(65)}$$

From this result we can deduce the following conclusions:

1. Assuming θ to be very small (i.e., $\cos\theta \simeq 1$), we can see that the angle $\Delta\theta$ is directly proportional to the order of spectrum m for a given $\Delta\lambda$, so that for a given m, $\Delta\theta/\Delta\lambda$ is a constant. Such a spectrum is known as a normal spectrum, and in this the difference in angle for two spectral lines is directly proportional to the difference in wavelengths. However, for large θ, it can be easily shown that the dispersion is greater at the red end of the spectrum.
2. Equation (65) tells us that $\Delta\theta$ is inversely proportional to d, and therefore the smaller the grating element, the larger the angular dispersion.

Figures 18.37 and 18.38 show schematic diagrams of the experimental arrangement for studying the grating spectrum of a polychromatic source. In Fig. 18.37 we have shown a small hole placed at the focal plane of lens L_1. A parallel beam of white light emerging from L_1 falls on the grating, and the diffraction pattern is observed on the focal plane of lens L_2. If instead of a hole we have a slit at the focal plane of L_1 (see Fig. 18.38) — as is indeed the case in a typical laboratory set up — we will have parallel beams propagating in different directions, and in the focal plane of the lens L_2 we will have a band spectrum as shown in Fig. 18.38.

Lens L_2 is the objective of a telescope, and the diffraction pattern is viewed through an eyepiece. The angles of diffraction for various orders of the grating spectrum can be measured, and knowing the value of d, one can calculate the wavelength of different spectral lines.

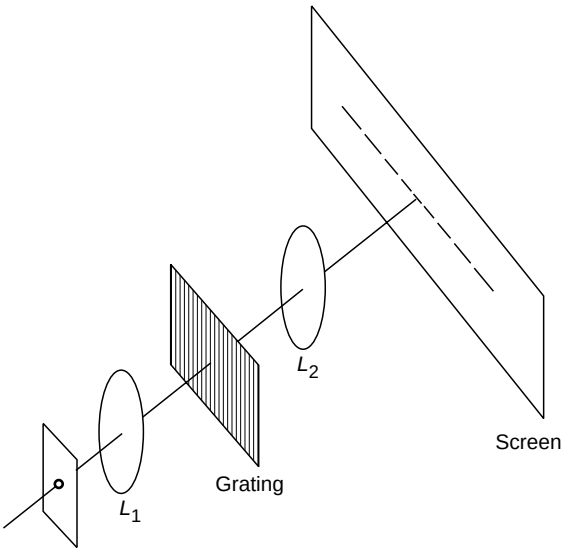


Fig. 18.37 Fraunhofer diffraction of a plane wave incident normally on a grating.

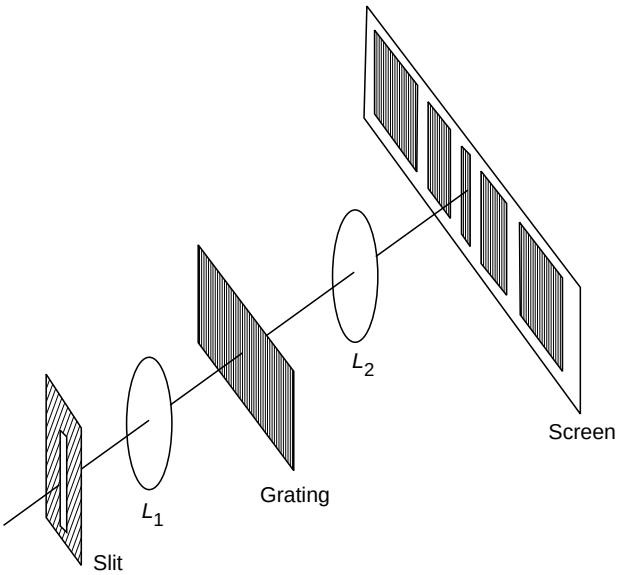


Fig. 18.38 If instead of a point source we have a slit in the focal plane of L_1, then we will obtain bands on the focal plane of L_2.

Example 18.10 Consider a diffraction grating with 15,000 lines per inch. (a) Show that if we use a white light source, the second- and third-order spectra overlap. (b) What will be the angular separation of the D_1 and D_2 lines of sodium in the second-order spectra?

Solution: (a) The grating element is

$$d = \frac{2.54}{15{,}000} = 1.69 \times 10^{-4} \text{ cm}$$

Let θ_{mv} and θ_{mr} represent the angles of diffraction for the mth-order spectrum corresponding to the violet and red colors, respectively. Thus

$$\theta_{2v} = \sin^{-1}\left(\frac{2\times 4\times 10^{-5}}{1.69\times 10^{-4}}\right) \approx \sin^{-1} 0.473 \approx 28.2°$$

$$\theta_{2r} = \sin^{-1}\left(\frac{2\times 7\times 10^{-5}}{1.69\times 10^{-4}}\right) \approx \sin^{-1} 0.828 \approx 55.90°$$

and

$$\theta_{3v} = \sin^{-1}\left(\frac{3\times 4\times 10^{-5}}{1.69\times 10^{-4}}\right) \approx \sin^{-1} 0.710 \approx 45.23°$$

where we have assumed the wavelengths of the violet and red colors to be 4×10^{-5} and 7×10^{-5} cm, respectively. Since $\theta_{2r} > \theta_{3v}$, the second- and third-order spectra will overlap. Further since $\sin\theta_{3r} > 1$, the third-order spectrum for the red color will not be observed.

(b) Since $d \sin\theta = m\lambda$, we have for small $\Delta\lambda$:

$$(d \cos\theta)\,\Delta\theta = m(\Delta\lambda)$$

or

$$\Delta\theta = \frac{m\Delta\lambda}{d\left\{1-(m\lambda/d)^2\right\}^{1/2}}$$

$$\simeq \frac{2\times 6\times 10^{-8}}{1.69\times 10^{-4}\left[1-\left(2\times 6\times 10^{-5}/1.69\times 10^{-4}\right)^2\right]^{1/2}}$$

$$\simeq 0.0010 \text{ rad} \simeq 3.44 \text{ minutes}$$

Thus, if we are using telescope of angular magnification 10, the two lines will appear to have an angular separation of 34.4 minutes.

18.8.2 Resolving Power of a Grating

In the case of a grating, the resolving power refers to the power of distinguishing two nearby spectral lines and is defined by the

$$R = \frac{\lambda}{\Delta\lambda} \qquad \textbf{(66)}$$

where $\Delta\lambda$ is the separation of two wavelengths which the grating can just resolve; the smaller the value of $\Delta\lambda$, the larger the resolving power.

The Rayleigh criterion (see Sec. 18.4) can again be used to define the limit of resolution. According to this criterion, if the principal maximum corresponding to the wavelength $\lambda + \Delta\lambda$ falls on the first minimum (on the either side of the principal maximum) of the wavelength λ, then the two wavelengths λ and $\lambda + \Delta\lambda$ are said to be just resolved (see Fig. 18.39). If this common diffraction angle is represented by θ and if we are looking at the mth-order spectrum, then the two wavelengths λ and $\lambda + \Delta\lambda$ will be just resolved if the following two equations are simultaneously satisfied:

$$d \sin\theta = m(\lambda + \Delta\lambda) \qquad \textbf{(67)}$$

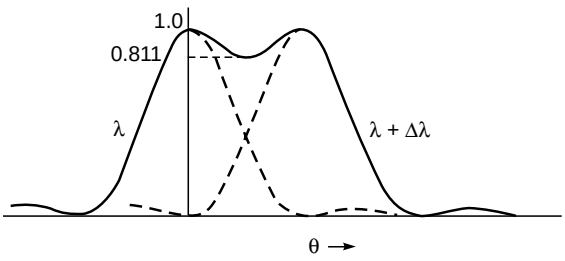


Fig. 18.39 The Rayleigh criterion for the resolution of two spectral lines.

and

$$d \sin\theta = m\lambda + \frac{\lambda}{N} \qquad \textbf{(68)}$$

Thus

$$R = \frac{\lambda}{\Delta\lambda} = mN \qquad \textbf{(69)}$$

which implies that the resolving power depends on the total number of lines in the grating—obviously on only those lines which are exposed to the incident beam (see the derivation in Sec. 18.6). Further, the resolving power is proportional to the order of the spectrum. Thus to resolve the D_1 and D_2 lines of sodium ($\Delta\lambda$ = 6 Å) in the first order, N must be at least $(5.89\times 10^{-5})/(6\times 10^{-8}) \approx 1{,}000$.

From Eq. (69) it appears that the resolving power of the grating will increase indefinitely if N is increased; however, for a given width of the grating D (= Nd), as N is increased d decreases and therefore the maximum value of m also decreases. Thus if d becomes 2.5λ, only first- and second-order spectra will be seen; and if it is further reduced to about 1.5λ, then only the first-order spectrum will be seen.

18.8.3 Resolving Power of a Prism

We conclude this section by calculating the resolving power of a prism. Figure 18.40 gives a schematic description of the experimental arrangement for observing the prism spectrum which is determined through the following formula:

$$n(\lambda) = \frac{\sin\{[A + \delta(\lambda)]/2\}}{\sin(A/2)} \qquad \textbf{(70)}$$

where A represents the angle of the prism and δ the angle of minimum deviation. We assume that the refractive index decreases with λ (which is usually the case) so that δ also decreases with λ. In Fig. 18.40 points P_1 and P_2 represent the images corresponding to λ and $\lambda + \Delta\lambda$, respectively. We are assuming that $\Delta\lambda$ is small so that the same position of the prism corresponds to the minimum deviation position for both wave-

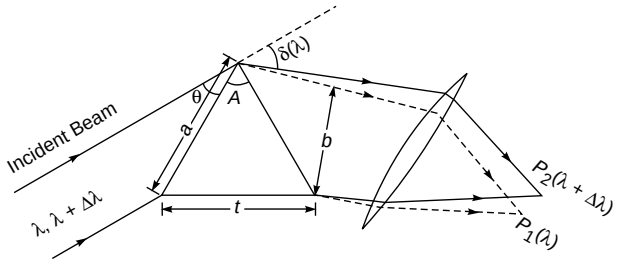


Fig. 18.40 The schematic of the experimental arrangement to observe the prism spectrum. P_1 and P_2 represent the images corresponding to λ and $\lambda + \Delta\lambda$, respectively.

lengths. In an actual experiment one usually has a slit source (perpendicular to the plane of the paper) forming line images at P_1 and P_2. Since the faces of the prism are rectangular, the intensity distribution will be similar to that produced by a slit of width b (see Sec. 18.2).[5] For the lines to be just resolved, the first diffraction minimum [$m = 1$ in Eq. (12)] of λ should fall at the central maximum of $\lambda + \Delta\lambda$; thus we must have

$$\Delta\delta \approx \frac{\lambda}{b} \tag{71}$$

To express $\Delta\delta$ in terms of $\Delta\lambda$, we differentiate Eq. (70):

$$\frac{dn}{d\lambda} = \frac{1}{\sin(A/2)} \cos\left[\frac{A+\delta(\lambda)}{2}\right] \frac{1}{2} \frac{d\delta}{d\lambda}$$

Thus

$$\Delta\delta = \frac{2\sin(A/2)}{\cos\{[A+\delta(\lambda)]/2\}} \frac{dn}{d\lambda} \Delta\lambda$$

Now from Fig. 18.40, we have

$$\theta = \frac{1}{2}[\pi - (A+\delta)]$$

or $$\sin\theta = \frac{b}{a} = \cos\frac{A+\delta}{2}$$

where the length a is shown in the figure. Further

$$\sin\frac{A}{2} = \frac{t/2}{a}$$

where t is the length of the base of the prism. Thus

$$\Delta\delta \approx \frac{t}{b} \frac{dn}{d\lambda} \Delta\lambda \tag{72}$$

Substituting in Eq. (71), we get for the resolving power

$$R = \frac{\lambda}{\Delta\lambda} = t\frac{dn}{d\lambda} \tag{73}$$

Now, for most glasses, the wavelength dependence of the refractive index (in the visible region of the spectrum) can be accurately described by the Cauchy formula

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} + \cdots \tag{74}$$

Thus

$$\frac{dn}{d\lambda} = -\left(\frac{2B}{\lambda^3} + \frac{4C}{\lambda^5} + \cdots\right) \tag{75}$$

the negative sign implying that the refractive index decreases with an increase in wavelength. As an example, we consider telescope crown glass for which[6]

$A = 1.51375$ $B = 4.608 \times 10^{-11}$ cm^2 $C = 6.88 \times 10^{-22}$ cm^4

For $\lambda = 6 \times 10^{-5}$ cm we have

$$\frac{dn}{d\lambda} \simeq -(4.27 \times 10^2 + 3.54)$$
$$\simeq -4.30 \times 10^2 \text{ cm}^{-1}$$

Thus, for $t \simeq 2.5$ cm we have

$$R = \frac{\lambda}{\Delta\lambda} \simeq 1000$$

which is an order of magnitude less than that for typical diffraction gratings with 15,000 lines.

18.9 OBLIQUE INCIDENCE

Until now we have assumed plane waves incident normally on the grating. For an experimental setting it is quite difficult to achieve the condition of normal incidence to a great precision, and it is easily seen that slight deviations from normal incidence will introduce considerable errors. It is, therefore, more practical to consider the more general oblique incidence case (see Fig. 18.41). The wavelength measurement can be carried out by using the method of minimum deviation as we do for prisms.

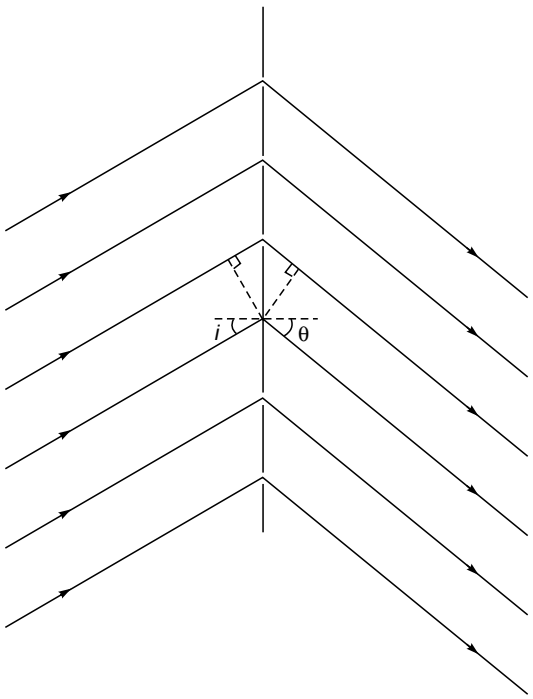


Fig. 18.41 Diffraction of a plane wave incident obliquely on a grating.

[5] Since we have a slit source, we need not consider the diffraction in a direction perpendicular to the plane of the diagram.
[6] Data quoted from Ref. 2.

If the angle of incidence is i, then the path difference of the diffracted rays from two corresponding points in adjacent slits will be $d \sin\theta + d \sin i$ (see Fig. 18.41). Thus, principal maxima will occur when

$$d(\sin\theta + \sin i) = m\lambda \quad \textbf{(76)}$$

or $$d[\sin(\delta - i) + \sin i] = m\lambda \quad \textbf{(77)}$$

when $\delta = i + \theta$ is the angle of deviation. For δ to be minimum, we must have

$$\frac{d}{di}[\sin(\delta - i) + \sin i] = 0 \quad \textbf{(78)}$$

$$-\cos(\delta - i) + \cos i = 0$$

i.e., $$i = \delta - i = \theta \quad \textbf{(79)}$$

or $$i = \frac{\delta}{2} = \theta \quad \textbf{(80)}$$

Hence, at the position of minimum deviation, the grating condition becomes

$$2d \sin\frac{\delta}{2} = m\lambda \quad \textbf{(81)}$$

The minimum deviation position can be obtained in a manner similar to that used in the case of a prism, and since the adjustments are relatively simpler, this provides a more accurate method for the determination of λ.

18.10 X-RAY DIFFRACTION[7]

Visible light is an electromagnetic wave whose wavelength approximately lies between 4000 and 7000 Å. X-rays are also electromagnetic waves whose wavelengths are ~1 Å. Obviously, it is extremely difficult to make slits which are narrow enough for the study of X-ray diffraction patterns. Since the interatomic spacings in a crystal are usually of the order of angstroms, one can use it as a three-dimensional diffraction grating for studying the diffraction of X-rays. Indeed, X-rays have extensively been used to study crystal structures (Ref. 7).

In an ideal crystal, the atoms or molecules arrange themselves in a regular three-dimensional pattern which can be obtained by a three-dimensional repetition of a certain unit pattern. This simplest volume which has all the characteristics of the whole crystal and which completely fills space is called the unit cell. One can think of various identifiable planes in the regular three-dimensional periodic arrangement

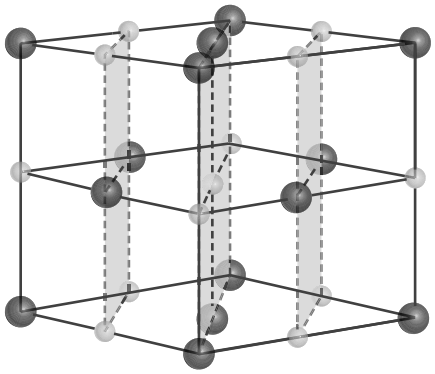


Fig. 18.42 Skewed planes in a NaCl crystal.

(see Fig. 18.42). Miller indices are universally used as a system of notation for planes within a crystal. They specify the orientation of planes relative to the crystal axis without giving the position of the plane in space with respect to the origin. These indices are based on the intercepts of a plane with the three crystal axes, each intercept with an axis being measured in terms of unit cell dimensions (a, b, or c) along that axis. To determine the Miller indices of a plane, the following procedure is used:

1. Find the intercepts (of the plane nearest to the origin) on the three axes, and express them as multiples or fractions of the unit cell dimension.
2. Take the reciprocals of these numbers and multiply by the LCM of the denominators.
3. Enclose in parentheses.

For example, a (111) plane intercepts all three axes at one unit distance [see Fig. 18.43(a)]; a (211) plane intercepts the three axes at $\frac{1}{2}$, 1, and 1 unit distances [see Fig. 18.43(b)]. Similarly, a (110) plane intercepts the z axis at ∞. Miller indices can also be negative; the minus sign is shown above the digit as in $(\bar{1}1\bar{1})$. Figure 18.44 shows the planes characterized by the Miller indices $(\bar{1}11)$ in a simple cubic lattice.

Consider a monochromatic beam of X-rays to be incident on a crystal. In Fig. 18.45 the horizontal dotted lines represent a set of parallel crystal planes with Miller indices (hkl). W_1W_2 and W_3W_4 represent the incident and reflected wave fronts, respectively. Obviously, the secondary wavelets emanating from points A, B, and C are in phase on W_3W_4 (see Sec. 12.4

[7] The author is grateful to Prof. Lalit K. Malhotra for his help in writing this section.

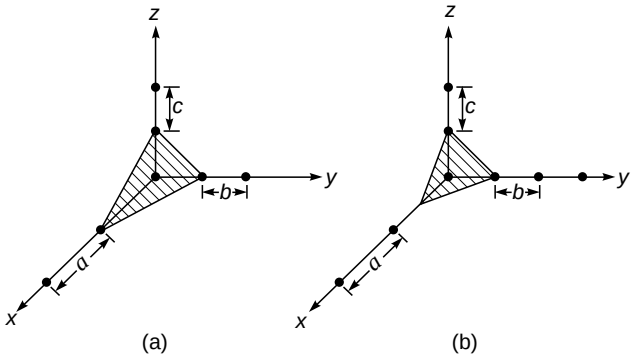


Fig. 18.43 (a) The (111) plane intercepts all three axes at 1 unit distance of each axial dimension. (b) The (211) plane intercepts the three axes at ½, 1, and 1 unit distances.

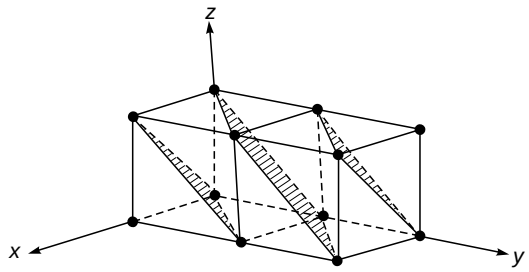


Fig. 18.44 Planes characterized by the Miller indices $(\bar{1}\,1\,1)$ in a simple cubic lattice.

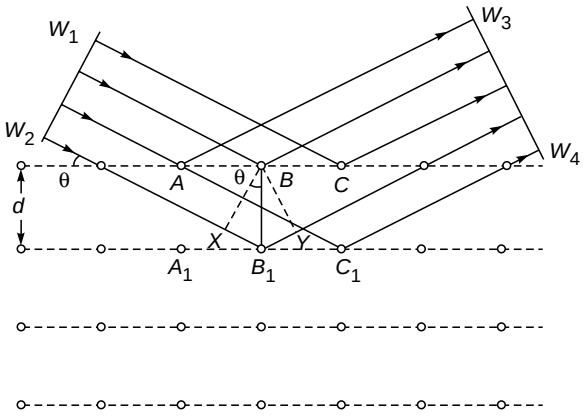


Fig. 18.45 Reflection of a plane wave by a set of parallel crystal planes characterized by the Miller indices (*hkl*). When the Bragg condition $2d \sin\theta = m\lambda$ is satisfied, the waves scattered from different rows will be in phase.

and Fig. 12.7); and the waves emanating from points A_1, B_1, and C_1 will also be in phase on W_3W_4 if

$$XB_1 + B_1Y = m\lambda \qquad m = 1, 2, 3, \ldots \qquad \textbf{(82)}$$

or when

$$2d_{hkl} \sin\theta = m\lambda \qquad \textbf{(83)}$$

where d_{hkl} is the interplanar spacing between crystal planes of indices (*hkl*), $m = 1, 2, 3, \ldots$ is called the order of diffraction, and θ is known as the glancing angle. This equation is known as Bragg's law and gives the angular positions of the reinforced diffracted beams in terms of the wavelength λ of the incoming X-rays and of the interplanar spacings d_{hkl} of the crystal planes. When the condition expressed by Eq. (83) is not satisfied, destructive interference occurs and no reinforced beam will be produced. Constructive interference occurs when the condition given by Eq. (83) is satisfied, leading to peaks in the intensity distribution. For solids which crystallize in cubic structures (which are discussed later), the interplanar spacing d_{hkl} between two closest parallel planes with Miller indices (*hkl*) is given by

$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}} \qquad \textbf{(84)}$$

where *a* represents the lattice constant. Thus knowing the Miller indices, we can find d_{hkl}; and from Bragg's law, we can determine the value of θ at which Bragg's equation can be satisfied.

There are three types of cubic structures: simple cubic, body-centered cubic (BCC) and face-centered cubic (FCC). Figure 18.46 shows a simple cubic structure (abbreviated as SC) in which the atoms are at the corners of a cube which forms what is known as a unit cell. The crystal is built up by the repetition of this unit cell in three dimensions. In addition, if

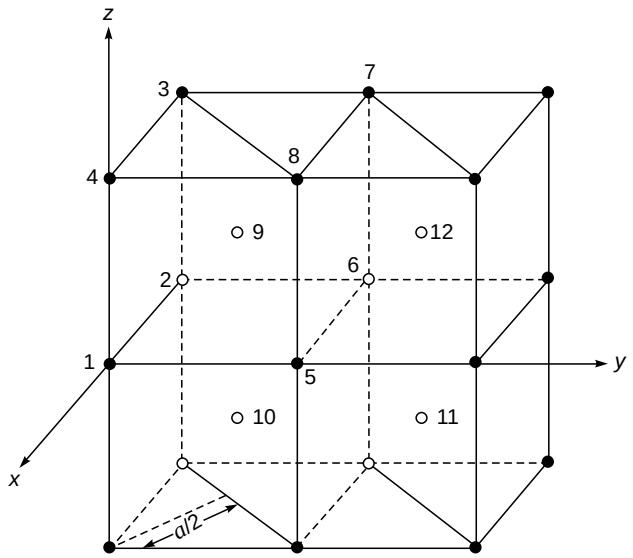


Fig. 18.46 A body-centered cubic (BCC) lattice. The $(\bar{1}10)$ planes are separated by $a/\sqrt{2}$.

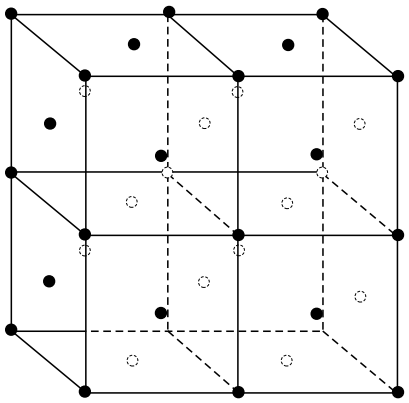


Fig. 18.47 A face-centered cubic (FCC) lattice.

there is an atom at the center of each cube (shown as 9, 10, 11, and 12 in Fig. 18.46), the arrangement is known as a BCC structure. The distance between two adjacent planes characterized by the Miller indices $(\bar{1}10)$ is $a/\sqrt{2}$, which can be verified by simple geometry. On the other hand, if instead of having an atom at the center of the cube there is an atom at the center of each of the six faces of the cube (see Fig. 18.47), we will have the FCC structure. Copper, silver, and gold crystallize in the FCC form with the lattice parameter a = 3.61, 4.09, and 4.08 Å, respectively. Metals such as sodium, barium, and tungsten crystallize in the BCC form with a = 4.29, 5.03, and 3.16 Å, respectively.[8]

Just as there are optical missing orders of a diffraction grating, there are structural extinctions of X-ray reflection from a crystal. For simple cubic structures, reflections from all (*hkl*) planes are possible. However, for the BCC structure, diffraction occurs only on planes whose Miller indices when added total to an even number. Thus for the BCC structure, the principal diffracting planes for a first-order diffraction are (110), (200), (211) (and other similar planes), etc. where $h + k + l$ is an even number. In the case of the FCC crystal structure, the principal diffracting planes are those whose Miller indices are either all even or all odd, e.g., (111), (200), (220), etc.

18.10.1 Experimental Methods of X-ray Diffraction

From Bragg's law $2d_{hkl} \sin\theta = m\lambda$, it is clear that essentially three methods can be used so that Bragg's formula can be satisfied:

	λ	θ
Rotating crystal method	Fixed	Variable (intentional)
Powder method	Fixed	Variable (inherent)
Laue method	Variable	Fixed

When one uses monochomatic X-rays, Bragg's formula cannot be satisfied for an arbitrary value of θ. Hence one rotates the single crystal so that reflection can occur for a discrete set of θ values. This method can be employed only if single crystals of reasonable size are available. If this is not the case, one can still use monochromatic X-rays provided the sample is in powder form so that there are always enough crystallites of the right orientation available to satisfy the Bragg relation. A powder will consist of a large number of randomly oriented microcrystals; each microcrystal is essentially a single crystal. As the X-ray beam passes through such a polycrystalline material, the orientation of any given set of planes, with reference to the X-ray beam, changes from one microcrystal to the other. Thus, corresponding to any given set of planes there will be a large number of crystals for which Bragg's condition will be satisfied, and on the photographic plate one will obtain concentric rings [see Fig. 18.48(a)]; each ring will correspond to a particular value of d_{hkl} and a particular value of m. The appearance of the circular rings can be understood as follows. Consider a set of planes parallel to AB [see Fig. 18.48(b)]. The glancing angle θ is assumed to satisfy the Bragg condition. If the microcrystal is rotated about the direction of the incident X-ray beam, then for all positions of the microcrystal, the glacing angle will be the same for these sets of planes. Further, for each position of the microcrystal, the direction of the diffracted beam will be different, but it will always lie on the surface of the cone whose semivertical angle will be 2θ. Consequently, one will obtain concentric circular rings on the photographic plate; these rings are known as Debye-Scherrer rings.

While using the powder method, the photographic film is put in a cylindrical form surrounding the polycrystalline sample as shown in Fig. 18.49(a). Each Debye-Scherrer ring will produce an arc on the film, and when the film is unrolled, one obtains a pattern as shown in Fig. 18.49(b) and (c). From the position of these arcs one can calculate θ and thus determine the interplanar spacing. From a study of the interplanar spacings one can determine the crystal structure.[9] Although a powder camera with an enclosed film strip has been extensively used in the past, modern X-ray crystal analysis uses

[8] Crystal structures other than cubic are also common; for example, zinc crystallizes into a hexagonal structure, and carbon forms a diamond structure. However, the most important fact is that in all these structures there is a definite periodicity of atoms.

[9] For more details, you may look up Ref. 7.

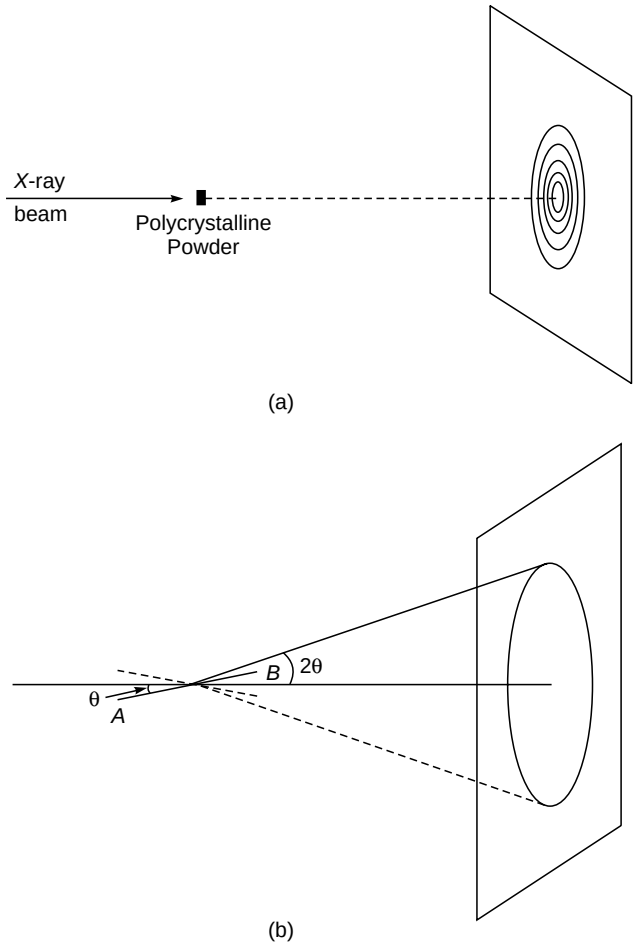


Fig. 18.48 (a) When a monochromatic X-ray beam falls on a polycrystalline sample, one obtains the Debye-Scherrer rings. (b) Diffraction from a polycrystalline sample.

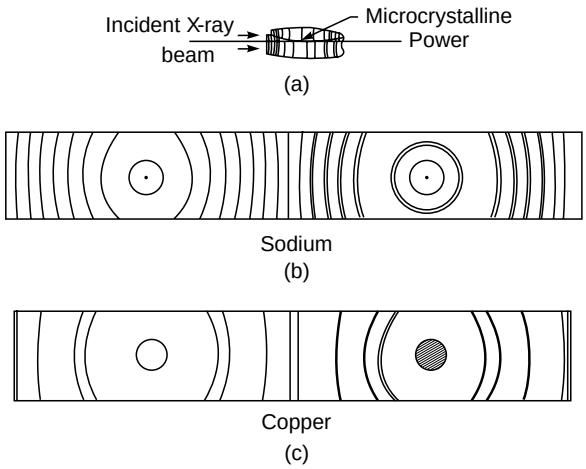


Fig. 18.49 (a) While using the powder method the photographic film is kept in a cylindrical form as shown in the figure. (b) and (c) Schematic diffraction patterns for sodium and copper, respectively.

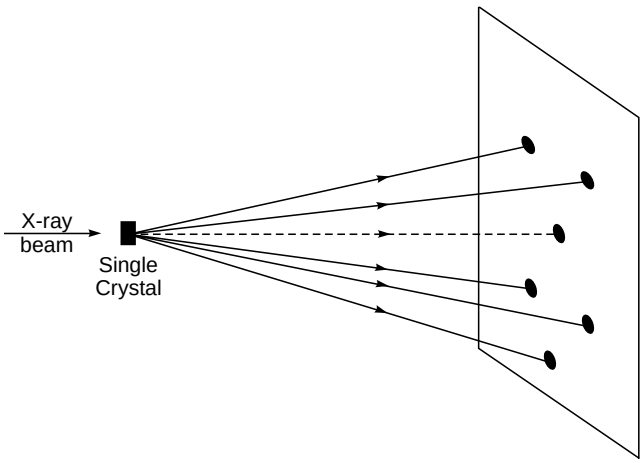


Fig. 18.50 When a polychromatic X-ray beam falls on a single crystal, one obtains Laue spots. Each set of planes chooses its own wavelength to satisfy the Bragg relation given by Eq. (84).

an X-ray diffractometer which has a radiation counter to detect the angle and intensity of the diffracted beam.

Finally there is the Laue method in which the single crystal is held stationary in a beam of white X-rays. Each set of planes then chooses its own wavelength to satisfy the Bragg relation (see Fig. 18.50).

To calculate the angles of diffraction, we substitute Eq. (84) into Bragg's law [Eq. (83)] to obtain

$$\frac{2a}{\sqrt{h^2 + k^2 + l^2}} \sin\theta = m\lambda \qquad \textbf{(85)}$$

We restrict ourselves to only first-order reflections ($m = 1$); higher-order reflections are usually rather weak (see also Prob. 18.22). Thus Eq. (85) can be written in the form

$$\sin\theta = \frac{\lambda}{2a}\sqrt{N} \qquad \textbf{(86)}$$

where

$$N = h^2 + k^2 + l^2$$

Now, for a simple cubic lattice, all values of (hkl) are possible, implying the following possible values of N:

$$N = 1, 2, 3, 4, 5, 6, 7, \ldots \text{(SC)} \qquad \textbf{(87a)}$$

Similarly, for a BCC lattice $h + k + l$ must be even, implying

$$N = h^2 + k^2 + l^2$$
$$= 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, \ldots \text{(BCC)} \qquad \textbf{(87b)}$$

Finally, for an FCC lattice, Miller indices are either all even or all odd, implying

$$N = h^2 + k^2 + l^2$$
$$= 3, 4, 8, 11, 12, 16, 19, 20, 24, 27, \ldots \text{ (FCC)} \quad \textbf{(87b)}$$

For a given structure and for given values of λ and a, one can now easily calculate the different values of θ. For example, if we consider λ = 1.540 and 1.544 Å (corresponding to the $CuK_{\alpha 1}$ and $CuK_{\alpha 2}$ lines), then for sodium (which is a BCC structure with a = 4.2906 Å) the various values of θ are

(14.70°, 14.74°), (21.03°, 21.09°), (26.08°, 26.15°),
(30.50°, 30.59°), (34.58°, 34.68°), (38.44°, 38.56°),
(42.18°, 42.32°), (45.88°, 46.03°), (49.59°, 49.76°),
(53.38°, 53.58°), (57.33°, 57.56°), (61.54°, 61.82°),
(66.22°, 66.56°), (79.41°, 80.23°)

The two values inside the parentheses correspond to the two wavelengths 1.540 and 1.544 Å, respectively. Because of the presence of two wavelengths, one obtains double lines for each family of planes which become resolvable only at higher scattering angles. Similarly one can consider reflections from other structures (see Probs. 18.19, 18.20, and 18.21). Each value of θ will give rise to a Debye-Scherrer ring shown in Figs. 18.48(a) and 18.49(b) and (c).

Finally, the intensity of the diffracted wave depends on the number of atoms per unit area in the plane under consideration. For example, corresponding to the $(\bar{1}10)$ and $(\bar{2}22)$ planes passing through a BCC lattice, there will be one atom and two atoms, respectively, in an area a^2. Thus in the first case the intensity of the diffracted wave will be much more than in the second case.

18.11 THE SELF-FOCUSING PHENOMENON[10]

With the availability of intense laser beams, a large number of interesting nonlinear optical phenomena have been investigated. One such nonlinear phenomenon is the effect on the propagation of a light beam due to the dependence of the refractive index on the intensity of the beam. This leads to the self-focusing (or defocusing) of the beam. To physically understand the self-focusing phenomenon, we assume the nonlinear dependence of the refractive index on the intensity to be of the form

$$n = n_0 + \frac{1}{2} n' E_0^2 \quad \textbf{(88)}$$

where n_0 is the refractive index of the medium in the absence of the electromagnetic field, n' is a constant representing the nonlinear effect,[11] and E_0 represents the amplitude of the electric field. As an example, we consider the incidence of a laser beam (propagating in the z direction) having Gaussian intensity distribution in the transverse direction; i.e., we assume

$$E(x, y, z, t) \approx E_0 \cos(kz - \omega t) \quad \textbf{(89)}$$

with

$$E_0 = E_{00} \exp\left(-\frac{r^2}{a^2}\right) \quad \textbf{(90)}$$

where a represents the width of the Gaussian beam and r ($= \sqrt{x^2 + y^2}$) represents the cylindrical coordinate. In the absence of any nonlinear effects, the beam will undergo diffraction divergence (see Sec. 20.5). However, if the beam is incident on a medium characterized by a positive value of n', the intensity distribution will create a refractive index distribution which will have a maximum value on the axis (i.e., at r = 0) and will gradually decrease with r. Indeed, using Eqs. (88) to (90), we will have

$$n \simeq n_0 + \frac{1}{2} n' E_{00}^2 \exp\left(-\frac{2r^2}{a^2}\right)$$
$$\simeq \left(n_0 + \frac{1}{2} n' E_{00}^2\right) - \frac{1}{2} n_0 \left(\frac{r}{\alpha}\right)^2 \quad \textbf{(91)}$$

where

$$\alpha^2 = \frac{n_0 a^2}{2 n' E_{00}^2} \quad \textbf{(92)}$$

[10] Based on Ref. 8; for a rigorous account see, e.g., Ref. 9.

[11] This dependence may arise from a variety of mechanisms, such as the Kerr effect, electrostriction, and thermal effect. The simplest to understand is the thermal effect, which is due to the fact that when an intense optical beam having a transverse distribution of intensity propagates through an absorbing medium, a temperature gradient is set up. For example, if the beam has a Gaussian transverse intensity variation [i.e., of the form $\exp(-r^2/a^2)$; the direction of propagation being along the z axis], then the temperature will be maximum on the axis (i.e., r = 0) and will decrease with an increase in the value of r. If $dn/dT > 0$, the refractive index will be maximum on the axis, and the beam will undergo focusing; on the other hand, if $dn/dT < 0$, the beam will undergo defocusing (see, e.g., Ref. 9).

The Kerr effect arises due to the anisotropic polarizability of liquid molecules (such as CS_2). An intense light wave will tend to orient the anisotropically polarized molecules such that the direction of maximum polarizability is along the direction of the electric vector; this changes the dielectric constant of the medium. On the other hand, electrostriction (which is important in solids) is the force which a nonuniform electric field exerts on a material medium; this force affects the density of the material, which in turn affects the refractive index. Thus, a beam having nonuniform intensity distribution along its wave front will give rise to a refractive index variation leading to the focusing (or defocusing) of the beam. For a detailed discussion on electrostriction and the Kerr effect, see Refs. 9 to 11.

and in writing Eq. (91) we have expanded the exponential term and have retained only the first two terms. In other words, we are restricting ourselves to small values of r, which is the paraxial approximation. The term $\frac{1}{2}n'E_{00}^2$ is usually very small compared to n_0; so we may write (after squaring)

$$n^2 \simeq n_0^2\left[1-\left(\frac{r}{\alpha}\right)^2\right] \tag{93}$$

We may recall that in Sec. 3.4.1 we considered propagation in a medium whose refractive index decreased parabolically from the axis, and we showed that the beam could undergo periodic focusing (see Fig. 3.25). Indeed we showed that the medium behaved as a converging lens of focal length $\pi\alpha/2$ [see Eq. (48) of Chap. 3]. In the present case also because of nonlinear effects (with $n' > 0$), the medium will act as a converging lens of focal length approximately given by

$$f_{nl} \simeq \frac{\pi}{2}\alpha \simeq \frac{\pi}{2}\left(\frac{n_0}{2n'E_{00}^2}\right)^{1/2} a \tag{94}$$

the subscript (nl) signifying that the effect is due to a nonlinear phenomenon. Thus because of nonlinear effects the beam is said to undergo *self-focusing;* the word *self* signifies the fact that the beam creates its own refractive index gradient, resulting in the focusing of the beam.[12]

Our analysis in Sec. 3.4.1 for the calculation of the focal length was based on ray optics and neglected diffraction effects. Now, in the absence of any nonlinear effects, the beam will spread out due to diffraction, and the angle of divergence will be approximately given by (see Fig. 18.14)

$$\theta_d \simeq \frac{\lambda}{\pi a} = \frac{\lambda_0/n_0}{\pi a} \tag{95}$$

where λ_0 is the free space wavelength. Thus the phenomenon of diffraction can be approximated by a diverging lens of focal length (see Fig. 18.51)

$$f_d \simeq \frac{a}{\theta_d} \simeq \frac{1}{2}ka^2 \tag{96}$$

where $$k = \frac{2\pi}{\lambda} = \frac{2\pi}{\lambda_0}n_0 \tag{97}$$

Clearly if $f_d < f_{nl}$, the diffraction divergence will dominate and the beam will diverge. On the other hand, if $f_{nl} < f_d$, the nonlinear focusing effects will dominate and the beam will undergo self-focusing. For $f_d \approx f_{nl}$, the two effects will cancel each other, and the beam will propagate without any focusing or defocusing.

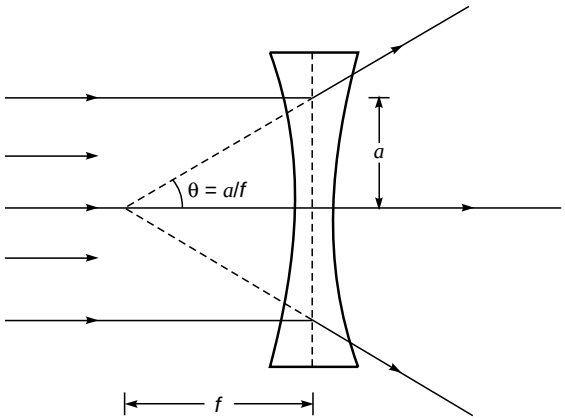


Fig. 18.51 When a plane wave is incident on a diverging lens, the transmitted rays diverge making an angle $\theta \approx a/f$ with the axis.

This is the condition of *uniform waveguide like propagation.* To determine the critical power of the beam, we note that the condition $f_d \approx f_{nl}$ implies

$$\frac{1}{2}ka^2 \approx \frac{\pi}{2}\left(\frac{n_0}{2n'E_{00}^2}\right)^{1/2} a$$

or $$E_{00}^2 \simeq \frac{1}{n_0 n'}\frac{\lambda_0^2}{8a^2} \tag{98}$$

Now the total power of the beam is given by

$$\begin{aligned} P &= \int_0^\infty \text{velocity} \times (\text{energy/unit volume}) \times 2\pi r\,dr \\ &\simeq \int_0^\infty \frac{c}{n_0} \times \frac{1}{2}\varepsilon E_0^2 \times 2\pi r\,dr \\ &\simeq \frac{c}{n_0}\left(\frac{1}{2}n_0^2\varepsilon_0 E_{00}^2\right)\int_0^\infty \exp\left(-\frac{2r^2}{a^2}\right)2\pi r\,dr \\ &\simeq \frac{\pi}{4}n_0 c\varepsilon_0 E_{00}^2 a^2 \end{aligned} \tag{99}$$

where ε ($= n_{00}^2\varepsilon_0$) is the dielectric permittivity of the medium and ε_0 ($= 8.85 \times 10^{-12}$ C^2 N^{-1} m^{-2}) is the dielectric permittivity of free space (see Sec. 19.2). Substituting the expression for E_{00}^2 from Eq. (98) into Eq. (99), we obtain the following expression for the critical power:

$$P_{cr} \simeq \frac{\pi}{32}(c\varepsilon_0)\frac{\lambda_0^2}{n'} \tag{100}$$

[12] If n' were a negative quantity, the refractive index would have increased as we moved away from the axis and the beam would have undergone defocusing. For example, if the refractive index decreases with an increase in temperature, the beam may undergo what is known as thermal defocusing.

Garmire, Chiao, and Townes (Ref. 12) carried out experiments on the self-focusing of a ruby laser beam ($\lambda_0 = 0.6943\ \mu m$) in CS_2 and found that the critical power was 25 ± 5 kW. Equation (100) gives

$$P_{cr} \simeq \frac{3.14}{32} \times 3 \times 10^8 \times 8.85 \times 10^{-12} \times \frac{(0.6943 \times 10^{-6})^2}{2 \times 10^{-20}}$$

$$\simeq 6.3\ \text{kW} \qquad \textbf{(101)}$$

where we have used the following parameters for CS_2: $n_0 \simeq 1.6276$, $n' \simeq 1.8 \times 10^{-11}$ cgs units $\simeq 2 \times 10^{-20}$ mks units. [The mks unit for n' is (meter/volt)^2.] Although the result is wrong by a factor of about 4, one does obtain the correct order; this is indeed the case for all order-of-magnitude calculations. Thus

1. When $P < P_{cr}$, the beam will diverge due to diffraction.
2. When $P = P_{cr}$, the beam will propagate without divergence or convergence. This is the condition for uniform waveguide propagation.
3. When $P > P_{cr}$, we may extrapolate that the beam will undergo focusing, which is indeed borne out by more rigorous analysis. This is known as the self-focusing of the beam.

A detailed study of the self-focusing phenomenon is of considerable importance in laser-induced fusion experiments where there is a nonlinear interaction of the laser beam with the plasma.

18.12 OPTICAL MEDIA TECHNOLOGY—AN ESSAY[13]

Optical media technology has been around for the last 30 years or so. There were various avatars of this technology beginning with laser video disc (an optical medium with analog recording) to the dramatic breakthrough in the form of compact disc ROM, compact disc recordable/rewritable, magneto-optical disc, DVD-ROM, recordable and rewritable to present-day Blu-Ray technology. The optical storage solution using laser typically provides lowest cost per byte; is rugged, transportable, and interchangeable; has fast random access, is durable, and is available in both erasable and nonerasable forms.

Optical data storage are a system in which data are stored and retrieved by light, which happens to be a laser. As for the majority of consumer products, the lasing takes place through a semiconductor laser diode. The optical system

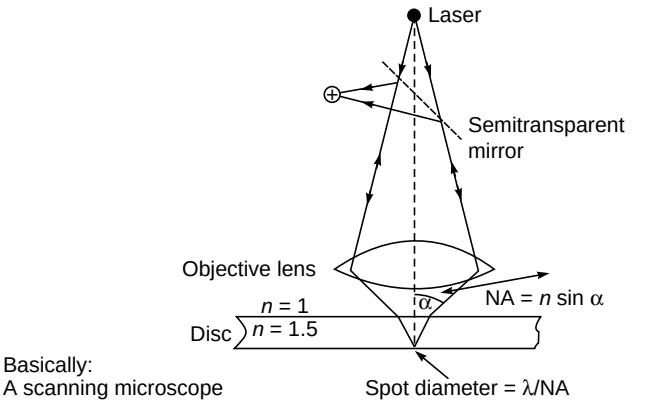


Fig 18.52 Optical pickup unit (OPU) is essentially a scanning optical microscope, which is mounted on a servomotor to scan any desired position of the disc.

consists of two broad parts: a medium or disc and a drive. The discs are all made of optically clear plastic material and contain digitized information in the form of spiral track consisting of finite lengths of "pits" and "lands". An optical drive contains the optical pickup unit (OPU) and servo control system for controlling it, a disc rotation system, and associated electronics mounted on a motherboard.

As shown in Fig. 18.52, the OPU is essentially a scanning optical microscope, which is mounted on a servomotor to scan any desired position of the disc. The OPU is used for both recording and reading of data. The OPU is equipped with a laser diode and a set of optical elements which focus the beam on the disc surface and detect the light reflected back. The spot size of the light is given by the familiar optical concept of numerical aperture (NA) of the OPU lens as shown in Fig. 18.53. The higher the NA; the smaller the spot size. Thus, as discussed later, this aspect is utilized for making smaller spot sizes to cram more data into the same physical structure, resulting in transition to ever-increasing data capacity in the form of DVD and Blu-Ray format.

In a disc, the pits and lands are essentially physical features (protrusions) on the disc surface, which are put there through injection molding. The heights of the pits from the surface are not arbitrary; rather they are fixed. For an ideal case, this height should be equal ~ $\lambda/4$, as then the total path difference after reflection is $\lambda/2$, where λ is the wavelength of the laser used. This ensures that perfect bright and dark fringes are produced. However, in reality this height is actually ~ $\lambda/6$ as some amount of light signal is required from the pits for the servo controlling the OPU on

[13] This essay has been kindly written by Dr. Rajeev Jindal, Mr. Giriraj Nyati, and Mr. Subrata Dutta of Moser Baer India in Greater Noida, India. Moser Baer has done pioneering work in the manufacture of DVDs.

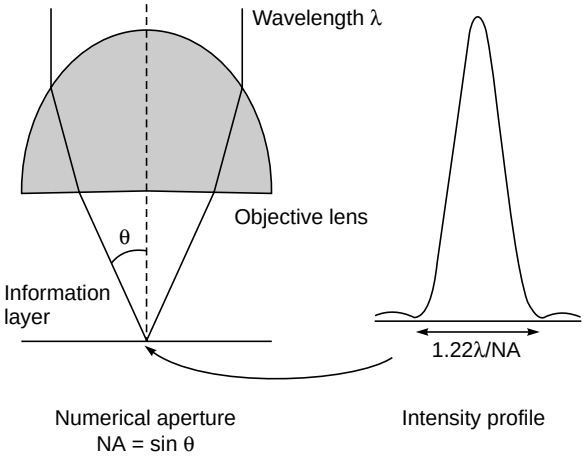


Fig 18.53 The spot size of the light is determined by the wavelength and the NA (numerical aperture) of the OPU lens.

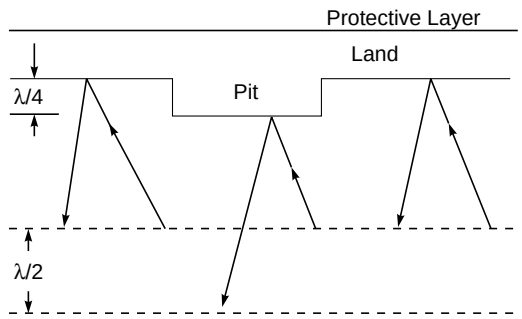


Fig 18.54 The pits and lands are essentially physical features (protrusions) on the disc surface, which are put there through injection molding. The heights of the pits from the surface are not arbitrary; rather they are fixed, being equal to λ/4 where λ is the wavelength of the laser used.

the disc, and hence a perfect dark area is not preferable. As shown in Fig. 18.54, bump (pit) height causes a path difference of ~ λ/2 relative to land. The optical head reads information by capturing reflected light as the laser beam travels across the pits and lands—a transition from either land to pit or pit to land is taken a logic 1 and no transition is taken as logic 0—polarity of pits can be either dark on bright background or reversed. This is shown in Fig. 18.55. The actual working can be seen through a schematic diagram as given in Fig. 18.56. The actual system and a cut-a way section is shown in Figs. 18.57 and 18.58.

A CD-ROM substrate is made of optically clear polycarbonate over which the data marks are made through injection molding. The inner hole has a diameter of 15 mm while the overall diameter of the disc is 120 mm and the thickness is 1.2 mm. The top of the disc is covered with a very thin layer

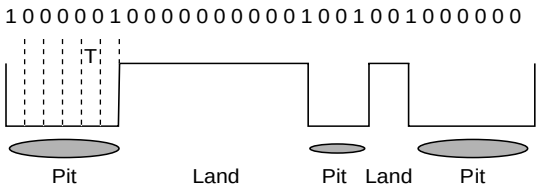


Fig. 18.55 Optical head reads information by capturing reflected light as the laser beam travels across the pits and lands; changes in the light intensity are interpreted as 0s and 1s. Polarity of pits can be either dark on bright background or reversed.

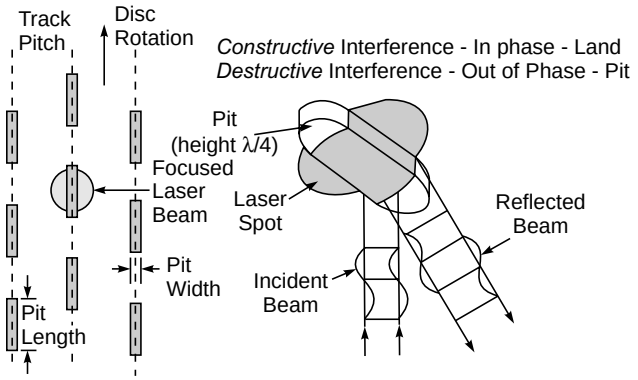


Fig. 18.56 Schematic diagram of reflection from the pit.

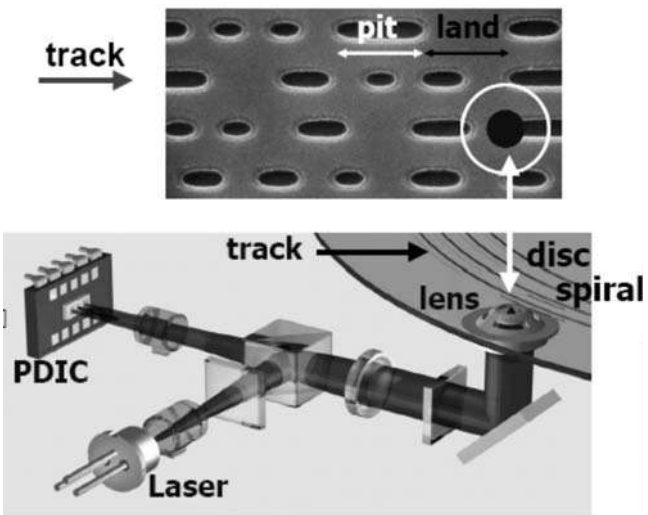


Fig. 18.57 The pits and lands are essentially physical features (protrusions) on the disc surface, which are put there through injection molding. The heights of the pits from the surface are not arbitrary; rather they are fixed, being equal to λ/4 where λ is the wavelength of the laser used. A color figure appears in the insert at the back of the book. Photograph kindly provided by Dr. Rajeev Jindal of Moser Baer, India.

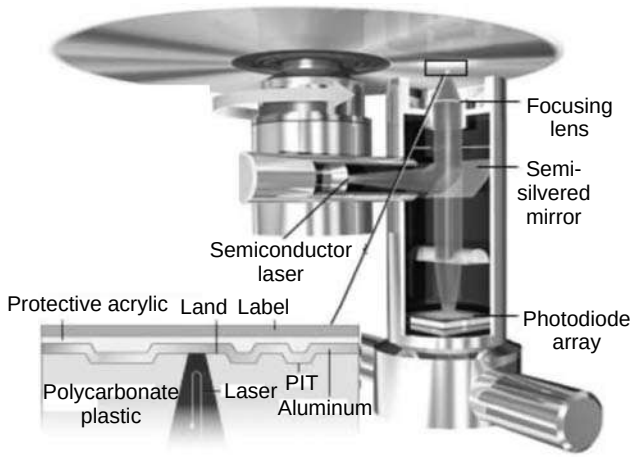


Fig. 18.58 A CD-ROM substrate is made of optically clear polycarbonate over which the data marks are made through injection molding. The inner hole has a diameter of 15 mm while the overall diameter of the disc is 120 mm and the thickness is 1.2 mm. The top of the disc is covered with a very thin layer of silver or gold to form a reflective layer that reflects the laser beam so as to be read back. The reflected light is incident on a quadrant photodetector, which converts the light to suitable electrical pulses, which are subsequently processed to extract relevant data. Details are given in the text. A color figure appears in the insert at the back of the book. Photograph kindly provided by Dr. Rajeev Jindal of Moser Baer, India.

of silver or gold to form a reflective layer which reflects the laser beam so as to be read back. The reflected light is incident on a quadrant photodetector, which converts the light to suitable electrical pulses, which are subsequently processed to extract relevant data. The data capacity of the disc is typically 650 to 700 MB of digital data.

Transition from CD to DVD: Due to the need for higher capacity and better resolution (picture quality), it was decided to go for a medium having a higher capacity called a digital versatile disc (DVD). Due to the need for backward compatibility, it was decided to keep the physical structure of the disc the same, i.e., a plastic substrate of 120 mm diameter with a 15 mm diameter inner hole. However, as shown in Fig. 18.52, the spot size is proportional to the wavelength and inversely proportional to the NA, so in DVD the wavelength has been reduced to 650 nm and NA increased to 0.65. The result has been a reduced spot size. However, the optical path has been reduced (to reduce the aberrations) in the process, resulting in thinner substrate. This entire process is shown in Figs. 18.59 and 18.60. However, ever-increasing hunger for more data has resulted in a new product called Blu-Ray

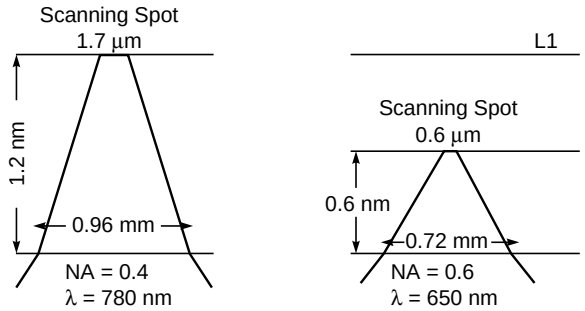


Fig. 18.59 The reduction in the spot size by decreasing the spot size and increasing the numerical aperture.

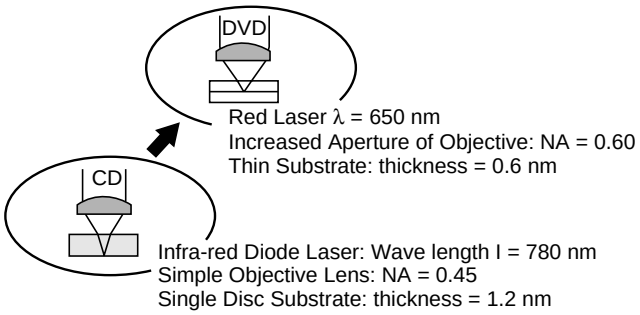


Fig. 18.60 The reduction in the spot size by decreasing the spot size and increasing the numerical aperture.

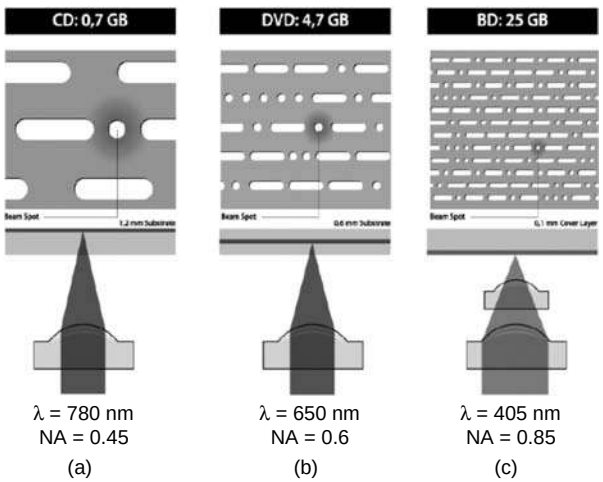


Fig. 18.61 (a) Infrared diode laser (λ = 780 nm) with a simple objective lens with NA = 0.45. (b) Red laser (λ = 650 nm) with increase aperture objective with NA = 0.60. (c) Blue laser (λ = 405 nm) with further increase in NA = 0.85. A color figure appears in the insert at the back of the book. Photograph kindly provided by Dr. Rajeev Jindal of Moser Baer, India.

where using precedent argument the NA has been further increased to 0.85 and the wavelength reduced to 405 nm (blue-violet). The resulting increase in capacity is 25 GB per layer as shown in Figs. 18.61 and 18.62.

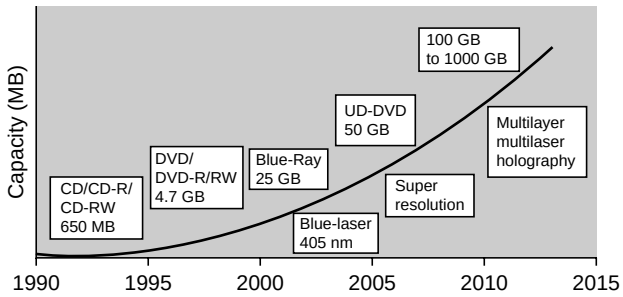


Fig. 18.62 The evolution of optical media technology.

Summary

- Interference corresponds to the situation when we consider the superposition of waves coming out from a number of point sources, and diffraction corresponds to the situation when we consider waves coming out from an area source such as a circular or rectangular aperture or even a large number of rectangular apertures (such as the diffraction grating).
- When a plane wave is incident normally on N parallel slits, the Fraunhofer diffraction pattern is given by

$$I = I_0 \frac{\sin^2 \beta}{\beta^2} \cdot \frac{\sin^2 N\gamma}{\sin^2 \gamma}$$

where

$$\beta = \frac{\pi b \sin\theta}{\lambda} \qquad \gamma = \frac{\pi d \sin\theta}{\lambda}$$

λ is the wavelength of light, θ is the angle of diffraction, b represents the width of each slit, and d is the separation between two slits. When N =1, we have the single-slit diffraction pattern producing a central maximum at $\theta = 0$ and minima when $b \sin\theta = m\lambda$, $m = \pm 1, \pm 2, \ldots$. When $N \geq 2$, the intensity distribution is the product of the single-slit diffraction pattern and the interference pattern produced by N point sources separated by a distance d. For $N = 2$, we obtain Young's double-slit interference pattern. For large values of N, the principal maxima occur when $\gamma = m\pi$, implying

$$d \sin\theta = m\lambda \qquad m = 0, 1, 2, \ldots$$

which is usually referred to as the grating condition.

- The resolving power of the grating is given by

$$R = \frac{\lambda}{\Delta\lambda} = mN$$

where N represents the total number of lines in the grating. For example, in the first-order spectrum ($m = 1$) of a diffraction grating with $N = 10{,}000$, for $\lambda \approx 5000$ Å we get $\Delta\lambda \approx 0.5$ Å.

- Consider a monochromatic beam of X-rays incident on a crystal. The glancing angle θ for which we have reinforced diffracted beams is given by

$$2d_{hkl} \sin\theta = m\lambda$$

where d_{hkl} is the interplanar spacing between crystal planes having Miller indices (hkl); $m = 1, 2, 3, \ldots$ is called the order of diffraction, and θ is known as the glancing angle. The above equation is known as Bragg's law and gives the angular positions of the reinforced diffracted beams.

Problems

18.1 A plane wave ($\lambda = 5000$ Å) falls normally on a long narrow slit of width 0.5 mm. Calculate the angles of diffraction corresponding to the first three minima. Repeat the calculations corresponding to a slit width of 0.1 mm. Interpret physically the change in the diffraction pattern.

[**Ans:** 0.057°, 0.115°, 0.17°; 0.29°, 0.57°, 0.86°]

18.2 A convex lens of focal length 20 cm is placed after a slit of width 0.6 mm. If a plane wave of wavelength 6000 Å falls normally on the slit, calculate the separation between the second minima on either side of the central maximum.

[**Ans:** $\simeq$ 0.08 cm]

18.3 In Prob. 18.2 calculate the ratio of the intensity of the principal maximum to the first maximum on either side of the principal maximum.

[**Ans:** ~ 21]

18.4 Consider a laser beam of circular crosssection of diameter 3 cm and of wavelength 5×10^{-5} cm. Calculate the order of the beam diameter after it has traversed a distance of 3 km.

[**Ans:** ~14 cm. This shows the extremely high directionality of laser beams.]

18.5 A circular aperture of radius 0.01 cm is placed in front of a convex lens of focal length 25 cm and illuminated by a parallel beam of light of wavelength 5×10^{-5} cm. Calculate the radii of the first three dark rings.

[**Ans:** 0.76, 1.4, 2.02 mm]

18.6 Consider a plane wave incident on a convex lens of diameter 5 cm and of focal length 10 cm. If the wavelength of the incident light is 6000 Å, calculate the radius of the first dark ring on the focal plane of the lens. Repeat the calculations for a lens of the same focal length but with diameter of 15 cm. Interpret the results physically.

[**Ans:** 1.46×10^{-4} cm, 4.88×10^{-5} cm]

18.7 Consider a set of two slits each of width $b = 5 \times 10^{-2}$ cm and separated by a distance $d = 0.1$ cm, illuminated by a monochromatic light of wavelength 6.328×10^{-5} cm. If a convex lens of focal length 10 cm is placed beyond the double-slit arrangement, calculate the positions of the minima inside the first diffraction minimum.

[**Ans:** 0.0316 mm, 0.094 mm]

18.8 Show that when $b = d$, the resulting diffraction pattern corresponds to a slit of width $2b$.

18.9 Show that the first-order and second-order spectra will never overlap when the grating is used for studying a light beam containing wavelength components from 4000 to 7000 Å.

18.10 Consider a diffraction grating of width 5 cm with slits of width 0.0001 cm separated by a distance of 0.0002 cm.

What is the corresponding grating element? How many orders would be observable at $\lambda = 5.5 \times 10^{-5}$ cm? Calculate the width of the principal maximum. Would there be any missing orders?

18.11 For the diffraction grating of Prob. 18.10, calculate the dispersion in the different orders. What will be the resolving power in each order?

18.12 A grating (with 15,000 lines per inch) is illuminated by white light. Assuming that white light consists of wavelengths lying between 4000 and 7000 Å, calculate the angular widths of the first- and the second-order spectra. [*Hint:* You should not use Eq. (65). Why?]

18.13 A grating (with 15,000 lines per inch) is illuminated by sodium light. The grating spectrum is observed on the focal plane of a convex lens of focal length 10 cm. Calculate the separation between the D_1 and D_2 lines of sodium. (The wavelengths of the D_1 and D_2 lines are 5890 and 5896 Å, respectively.) [*Hint*: You may use Eq. (65).]

18.14 Calculate the resolving power in the second-order spectrum of a 1 in. grating having 15,000 lines.

18.15 Consider a wire grating of width 1 cm having 1000 wires. Calculate the angular width of the second-order principal maxima, and compare the value with the one corresponding to a grating having 5000 lines in 1 cm. Assume $\lambda = 5 \times 10^{-5}$ cm.

18.16 In the minimum deviation position of a diffraction grating, the first-order spectrum corresponds to an angular deviation of 30°. If $\lambda = 6 \times 10^{-5}$ cm, calculate the grating element.

18.17 Calculate the diameter of a telescope lens if a resolution of 0.1 second of arc is required at $\lambda = 6 \times 10^{-5}$ cm.

18.18 Assuming that the resolving power of the eye is determined by diffraction effects only, calculate the maximum distance at which two objects separated by a distance of 2 m can be resolved by the eye. (Assume pupil diameter to be 2 mm and $\lambda = 6000$ Å.)

18.19 A pinhole camera is essentially a rectangular box with a tiny pinhole in front. An inverted image of the object is formed on the rear of the box. Consider a parallel beam of light incident normally on the pinhole. If we neglect diffraction effects, then the diameter of the image will increase linearly with the diameter of the pinhole. On the other hand, if we assume Fraunhofer diffraction, then the diameter of the first dark ring will go on increasing as we reduce the diameter of the pinhole. Find the pinhole diameter for which the diameter of the geometrical image is approximately equal to the diameter of the first dark ring in the Airy pattern. Assume $\lambda = 6000$ Å and a separation of 15 cm between the pinhole and the rear of the box.

[**Ans:** 0.47 mm]

18.20 Copper is an FCC structure with lattice constant 3.615 Å. An X-ray powder photograph of copper is taken. The X-ray beam consists of wavelengths 1.540 and 1.544 Å. Show that diffraction maxima will be observed at θ = (21.64°, 21.70°), (25.21°, 25.28°), (37.05°, 37.16°), (44.94°, 45.09°), (47.55°, 47.71°), (58.43°, 58.67°), (68.20°, 68.58°), (72.29°, 72.76°).

18.21 Tungsten is a BCC structure with lattice constant 3.1648 Å. Show that in the powder photograph of tungsten (corresponding to an X-ray wavelength of 1.542 Å) one would observe diffraction maxima at θ = 20.15°, 29.17°, 36.64°, 43.56°, 50.39°, 57.55°, 65.74°, and 77.03°.

18.22 (a) In the simple cubic structure if we alternately place Na and Cl atoms, we obtain the NaCl structure. Show that the Na atoms (and the Cl atoms) independently form FCC structures. The lattice constant associated with each FCC structure is 5.6402 Å. Corresponding to the X-ray wavelength 1.542 Å, show that diffraction maxima will be observed at θ = 13.69°, 15.86°, 22.75°, 26.95°, 28.27°, 33.15°, 36.57°, 37.69°, 42.05° 45.26°, 50.66°, 53.98°, 55.10°, 59.84°, 63.69°, 65.06°, 71.27°, 77.45°, and 80.66°.

(b) Show that if we treat NaCl as a simple cubic structure with lattice parameter 2.82 Å, then the maxima at θ = 13.69°, 26.95°, 36.57°, 45.26°, 53.98°, 63.69°, and 77.45° will not be observed. Indeed in the X-ray diffraction pattern of NaCl, the maxima corresponding to these angles will be very weak.

18.23 Show that the *m*th-order reflection from the planes characterized by (*hkl*) can be considered as the same as the first-order reflection from the planes characterized by (*mh mk ml*).

18.24 Calculate the Fraunhofer diffraction pattern produced by a double-slit arrangement with slits of widths *b* and 3*b*, with their centers separated by a distance 6*b*.

18.25 Consider the propagation of a 1 kW laser beam ($\lambda = 6943$ Å, beam diameter $\simeq$ 1 cm) in CS_2. Calculate f_d and f_{nl} and discuss the defocusing (or focusing) of the beam. Repeat the calculations corresponding to a 1000 kW beam, and discuss any qualitative differences that exist between the two cases. The data for n_0 and n_2 are given in Sec. 18.10.

18.26 The values of n_0 and n_2 for benzene are 1.5 and 0.6×10^{-10} cgs units, respectively. Obtain an approximate expression for the critical power.

REFERENCES AND SUGGESTED READINGS

1. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, 1975.
2. F. A. Jenkin and H. E. White, *Fundamentals of Optics*, McGraw-Hill Book Co., New York, 1957.
3. E. Hecht and A. Zajac, *Optics*, Addison-Wesley, Reading, Mass., 1974.
4. A. Nussbaum and R. A. Philips, *Contemporary Optics for Scientists and Engineers*, Prentice-Hall, Englewood Cliffs, N.J., 1976.
5. K. Thyagarajan and A. Ghatak, *Lasers: Theory and Applications*, Plenum Press, New York, 1981. Reprinted by Macmillan India Ltd., New Delhi.
6. A. Ghatak and K. Thyagarajan, *Optical Electronics*, Cambridge University Press, Cambridge, 1989. Reprinted by Foundation Books, New Delhi.
7. A. R. Verma and O. N. Srivastava, *Crystallography for Solid State Physics*, Wiley Eastern, New Delhi, 1982.
8. M. S. Sodha, "Theory of Nonlinear Refraction: Self Focusing of Laser Beams," *Journal of Physics Education (India)*, Vol. 1, No. 2, p. 13, 1973.
9. M. S. Sodha, A. K. Ghatak, and V. K. Tripathi, *Self-Focusing of Laser Beams in Dielectrics, Plasmas and Semiconductors*, Tata McGraw-Hill, New Delhi, 1974.
10. W. K. H. Panofsky and M. Philips, *Classical Electricity and Magnetism*, Addison-Wesley, Reading, Mass., 1962.
11. W. G. Wagner, H. A. Haus, and J. M. Marburger, "Large Scale Self-Trapping of Optical Beams in Paraxial Ray Approximation," *Physical Review Letters*, Vol. 175, p. 256, 1968.
12. E. Garmire, R. V. Chiao, and C. H. Townes, "Dynamics and Characteristics of the Self-Trapping of Intense Light Beams," *Physical Review Letters.*, Vol. 16, p. 347, 1966.
13. J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, New York, 1968.
14. C. J. Ball, *An Introduction to the Theory of Diffraction*, Pergamon Press, Oxford, 1971.
15. J. M. Cowley, *Diffraction Physics*, North-Holland, Amsterdam, 1975.
16. M. Francon, *Diffraction, Coherence in Optics*, Pergamon Press, Oxford, 1966.
17. H. F. Meiners, *Physics Demonstration Experiments*, Vol. 2, The Ronald Press Co., New York, 1970.

Chapter Nineteen

FRAUNHOFER DIFFRACTION II AND FOURIER OPTICS

Fourier analysis is a ubiquitous tool that has found application to diverse areas of physics and engineering.

—Joseph Goodman in the Preface to *Introduction to Fourier Optics*

19.1 INTRODUCTION

In this chapter, we will present a more general analysis of the far-field diffraction of a plane wave by different types of aperture; this is known as Fraunhofer diffraction. We will first derive the formula for what is known as Fresnel diffraction, which will be used in Chap.20. We will then make the far-field approximation, which will give us the Fraunhofer diffraction pattern; this will be shown to be the Fourier transform of the aperture function. We will also derive the Fourier transforming property of a thin lens that forms the basis of Fourier optics and of spatial frequency filtering.

19.2 THE FRESNEL DIFFRACTION INTEGRAL

We consider a plane wave (of amplitude A) incident normally on an aperture as shown in Fig. 19.1. Using the Huygens–Fresnel principle, we will calculate the field produced at point P on a screen SS' which is at a distance z from the aperture. Now, for a spherical wave *diverging* from the origin, the field distribution is given by

$$u \sim \frac{1}{r} e^{ikr}$$

where r is the distance from the source (at the origin) to the observation point. We consider an infinitesimal area $d\xi\, d\eta$ (around point M) on the plane containing the aperture; the field at point P due to waves emanating from this infinitesimal area will be proportional to

$$\frac{Ae^{ikr}}{r} d\xi d\eta \qquad \textbf{(1)}$$

where $r = MP$. To calculate the total field (at point P), we will have to sum over all the infinitesimal areas (in the aperture) to obtain

$$u(P) = C \iint \frac{Ae^{ikr}}{r} d\xi d\eta \qquad \textbf{(2)}$$

where C is a proportionality constant and the integration is over the entire aperture. From a more general theory, one can show that (see, e.g., Refs. 1 to 4; see also Sec.19.3):

$$C = -\frac{ik}{2\pi} = \frac{1}{i\lambda} \qquad \textbf{(3)}$$

We thus obtain

$$u(P) = \frac{A}{i\lambda} \iint \frac{e^{ikr}}{r} d\xi d\eta \qquad \textbf{(4)}$$

If the amplitude and phase distribution on the plane $z = 0$ is given by $A(\xi, \eta)$, then the above integral is modified to

$$u(P) = \frac{1}{i\lambda} \iint A(\xi, \eta) \frac{e^{ikr}}{r} d\xi d\eta \qquad \textbf{(5)}$$

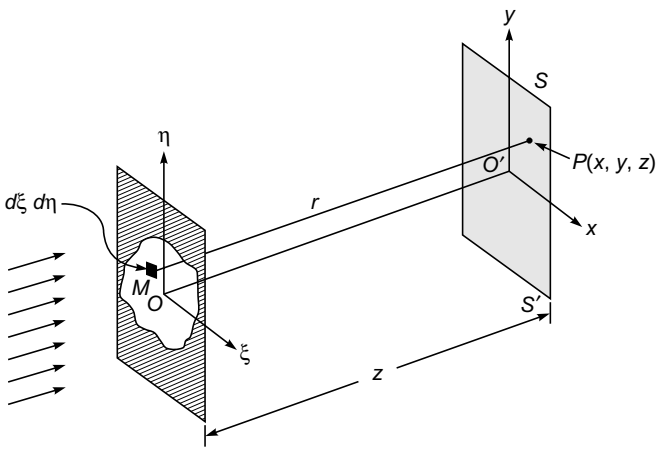


Fig 19.1 A plane wave incident normally on an aperture. The diffraction pattern is observed on screen SS'.

In writing Eqs. (4) and (5), we made two assumptions:

1. The first assumption is that the screen (in the plane of the aperture) does not affect the field at point *P*. This assumption is valid when the dimensions of the aperture are large in comparison to the wavelength. A more accurate analysis would take into consideration the effect of the screen on the field at any point *P*; this, in general, is a very difficult problem.
2. We have used a scalar theory in which we have represented the field by a scalar function *u*; this implies that the electric field is in the same direction everywhere. This assumption will be valid when the line joining point *O* and observation point *P* makes a small angle with the axis.

The quantity *r*, which represents the distance between point *M* [whose coordinates are (ξ, η, 0)] on the plane of the aperture and point *P* (whose coordinates are *x*, *y*, *z*) on the screen (see Fig. 19.1), will be given by

$$r = [(x-\xi)^2 + (y-\eta)^2 + z^2]^{1/2} = z\sqrt{1+\alpha}$$

where

$$\alpha \equiv \frac{(x-\xi)^2}{z^2} + \frac{(y-\eta)^2}{z^2} \tag{6}$$

Now, for $\alpha < 1$, we may write

$$\sqrt{1+\alpha} = 1 + \frac{1}{2}\alpha - \frac{1}{8}\alpha^2 + \cdots \tag{7}$$

If we assume $\alpha << 1$ and neglect quadratic and higher-order terms in the above expansion, we get

$$r \approx z + \frac{(x-\xi)^2}{2z} + \frac{(y-\eta)^2}{2z} \tag{8}$$

Further, in the denominator of Eq. (5) we may safely replace *r* by *z*, so that we may write[1]

$$u(x, y, z) \approx \frac{1}{i\lambda z} e^{ikz} \iint A(\xi, \eta) \times \exp\left\{\frac{ik}{2z}[(x-\xi)^2 + (y-\eta)^2]\right\} d\xi\, d\eta$$

Fresnel diffraction integral (9)

The above equation can be rewritten in the form

$$u(x, y, z) \approx \frac{1}{i\lambda z} e^{ikz} \exp\left[\frac{ik}{2z}(x^2 + y^2)\right] \iint A(\xi, \eta) \times \exp\left[\frac{ik}{2z}(\xi^2 + \eta^2)\right] e^{-i(u\xi + v\eta)} d\xi\, d\eta \tag{10}$$

where

$$u = \frac{2\pi x}{\lambda z} \quad \text{and} \quad v = \frac{2\pi y}{\lambda z} \tag{11}$$

are known as spatial frequencies. Both Eqs. (9) and (10) are usually referred to as the *Fresnel diffraction integral*. In Chap. 20 we will use the above integrals to calculate the Fresnel diffraction pattern. We must mention here that in the Fresnel approximation, we have neglected the terms proportional to α^2; this will be justified if it leads to maximum phase change which is much less than π. Thus the Fresnel approximation will be valid when

$$\frac{1}{8} kz\alpha^2 << \pi \Rightarrow \frac{1}{8}\frac{2\pi}{\lambda}\frac{\left[(x-\xi)^2 + (y-\eta)^2\right]^2}{z^3} << \pi \tag{12}$$

Thus, we must have

$$z >> \left(\frac{1}{4\lambda}\left[(x-\xi)^2 + (y-\eta)^2\right]^2_{\max}\right)^{\frac{1}{3}}$$

Condition for Fresnel approximation to be valid (13)

As an example, we consider a circular aperture of radius *a*; if we observe in a region of dimensions much greater than *a*, then we may neglect the terms involving ξ and η on the right-hand side to obtain

$$z >> \left[\frac{1}{4\lambda}(x^2 + y^2)^2\right]^{\frac{1}{3}} \tag{14}$$

Thus for a circular aperture of radius 0.1 cm, if we observe in a radius of about 1 cm, the maximum value of $x^2 + y^2$ will be about 1 cm^2; if we assume $\lambda \approx 5 \times 10^{-5}$ cm, Eq. (14) will imply $z >> 17$ cm.

[1] For example, for $\lambda = 6 \times 10^{-5}$ cm, the factor cos *kr* becomes $\cos\left(\frac{\pi}{3}10^5 r\right)$. As the value of *r* is changed from, say, 60 to 60.00002 cm, the cosine factor will change from +1 to –0.5. This shows the rapidity with which the exponential factor will vary in the domain of integration, although the change in *r* is extremely small.

19.3 UNIFORM AMPLITUDE AND PHASE DISTRIBUTION

We first consider the absence of any aperture. Thus, at $z = 0$

$$A(\xi,\eta) = A \qquad \text{for all values of } \xi \text{ and } \eta$$

and Eq. (9) can be written as

$$u(x, y, z) = \frac{A}{i\lambda z} e^{ikz} \int_{-\infty}^{+\infty} e^{\frac{ik}{2z}Y^2}\, dX \int_{-\infty}^{+\infty} e^{\frac{ik}{2z}Y^2}\, dY$$

where $X = x - \xi$ and $Y = y - \eta$. If we now use the integral (see App. A)

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2 + \beta x}\, dx = \sqrt{\frac{\pi}{\alpha}} \exp\left(\frac{\beta^2}{4\alpha}\right) \tag{15}$$

we get

$$u(x, y, z) = \frac{A}{i\lambda z} e^{ikz} \sqrt{\frac{\pi 2z}{-ik}} \sqrt{\frac{\pi 2z}{-ik}}$$

or

$$u(x, y, z) = A e^{ikz} \tag{16}$$

as it indeed should for a uniform plane wave. This shows that in spite of all the approximations that we have made, we ended up getting the correct result! The above equation also tells us that the value of C given by Eq. (3) is correct.

19.4 THE FRAUNHOFER APPROXIMATION

In the Fraunhofer approximation, we assume z to be so large that inside the integral in Eq. (10) the function

$$\exp\left[\frac{ik}{2z}(\xi^2 + \eta^2)\right]$$

can be replaced by unity, or the maximum phase change should be much less than π. Thus, *in addition* to the condition given by Eq. (13), we must have

$$z >> \frac{[\xi^2 + \eta^2]_{\max}}{\lambda} \qquad \boxed{\textbf{Condition for Fraunhofer approximation to be valid}} \tag{17}$$

In this approximation, Eq. (10) takes the form

$$u(x, y, z) \approx \frac{1}{i\lambda z} e^{ikz} \exp\left[\frac{ik}{2z}(x^2 + y^2)\right] \times \iint A(\xi, \eta) e^{-i(u\xi + v\eta)} d\xi\, d\eta \qquad \boxed{\textbf{Fraunhofer diffraction integral}} \tag{18}$$

which represents the Fraunhofer diffraction pattern. The integral on the right-hand side is the two-dimensional Fourier transform of the function $A(\xi, \eta)$ (see Sec. 9.6). Thus Eq. (18) gives the very important result that

Fraunhofer diffraction pattern is the Fourier transform of the aperture function.

For a circular aperture of radius a, Eq. (17) would become

$$z >> \frac{a^2}{\lambda} \tag{19}$$

We introduce the Fresnel number

$$N_F = \frac{a^2}{\lambda z} \tag{20}$$

Thus for the Fraunhofer approximation to be valid, we must have

$$N_F << 1 \tag{21}$$

19.5 FRAUNHOFER DIFFRACTION BY A LONG NARROW SLIT

We first consider Fraunhofer diffraction of a plane wave incident normally on a long narrow slit of width b (along the ξ axis) placed on the aperture plane. Figure 19.2 corresponds to a rectangular slit — if the slit is very long along the η axis, then we will have a long narrow slit. For such a

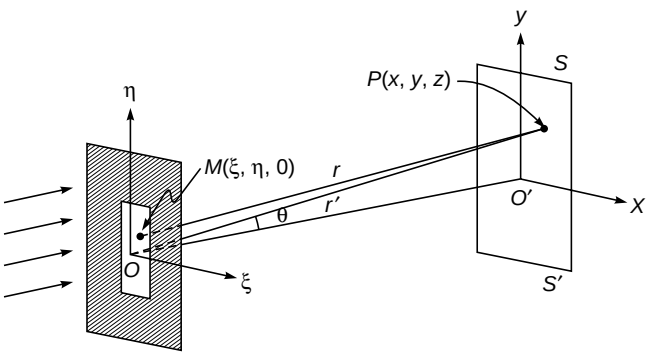


Fig. 19.2 Diffraction of a plane wave incident normally on a rectangular aperture.

case, we will have

$$A(\xi, \eta) = \begin{cases} A & |\xi| < \dfrac{b}{2} \\ 0 & |\xi| > \dfrac{b}{2} \end{cases} \tag{22}$$

for all values of η. Substituting Eq. (22) into Eq. (18), we obtain

$$u(x, y, z) = \frac{A}{i\lambda z} e^{ikz} \exp\left(\frac{ik}{2z}(x^2 + y^2)\right) \times \int_{-b/2}^{+b/2} e^{-iu\xi}\, d\xi \int_{-\infty}^{+\infty} e^{-iv\eta}\, d\eta \tag{23}$$

Now, in Sec. 9.3 we showed

$$\delta(v) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-iv\eta} d\eta \tag{24}$$

and

$$\int_{-b/2}^{+b/2} e^{-iu\xi} d\xi = \frac{1}{-iu} e^{-iu\xi}\Big|_{-b/2}^{+b/2}$$

$$= \frac{2}{u}\frac{e^{iub/2} - e^{-iub/2}}{2i} = b\frac{\sin\beta}{\beta}$$

where

$$\beta = \frac{ub}{2} = \frac{\pi b x}{\lambda z} \approx \frac{\pi b \sin\theta}{\lambda} \tag{25}$$

and $\sin\theta \approx x/z$, with θ representing the angle of diffraction along the x direction. Thus

$$u(x, y, z) = \frac{Ab}{i\lambda z} e^{ikz} \exp\left[\frac{ik}{2z}(x^2 + y^2)\right] \frac{\sin\beta}{\beta} 2\pi\delta(v) \tag{26}$$

Because of the δ function, the intensity is zero except on the x axis; thus the intensity distribution along the x axis will be

$$I = I_0 \frac{\sin^2\beta}{\beta^2} \tag{27}$$

We thus obtain the single-slit diffraction pattern as discussed in Sec. 18.3 [see Figs. 18.3(a) and 18.6].

19.6 FRAUNHOFER DIFFRACTION BY A RECTANGULAR APERTURE

We next consider a rectangular aperture (of dimension $a \times b$) (see Fig. 19.2). The Fraunhofer diffraction of a plane wave incident normally on such a rectangular aperture will be given by

$$u(x, y, z) = \frac{A}{i\lambda z} e^{ikz} \exp\left[\frac{ik}{2z}(x^2 + y^2)\right] \times \int_{-b/2}^{+b/2} e^{-iu\xi}\, d\xi \int_{-a/2}^{+a/2} e^{-iv\eta}\, d\eta \tag{28}$$

where we have chosen the origin to be at the center of the rectangular aperture (see Fig. 19.2). Carrying out the integration as in the previous section, we obtain

$$u(x, y, z) = \frac{Aba}{i\lambda z} e^{ikz} \exp\left[\frac{ik}{2z}(x^2 + y^2)\right]\left(\frac{\sin\beta}{\beta}\right)\left(\frac{\sin\gamma}{\gamma}\right) \tag{29}$$

where β is given by Eq. (25),

$$\gamma = \frac{va}{2} = \frac{\pi a y}{\lambda z} \approx \frac{\pi a \sin\phi}{\lambda} \tag{30}$$

and $\sin\phi \approx y/z$, with ϕ representing the angle of diffraction along the y direction. Thus we may write for the intensity distribution

$$I(P) = I_0 \frac{\sin^2\gamma}{\gamma^2}\frac{\sin^2\beta}{\beta^2} \tag{31}$$

The above equation represents the Fraunhofer diffraction pattern by a rectangular aperture. We must remember that Eqs. (29) and (31) are valid when both Eqs. (13) and (17) are satisfied. The intensity distribution due to a square aperture ($a = b$) is shown in Fig. 19.3; the figure corresponds to $a = b =$ 0.01 cm and $z = 100$ cm, and we have assumed $\lambda = 5 \times 10^{-5}$ cm.

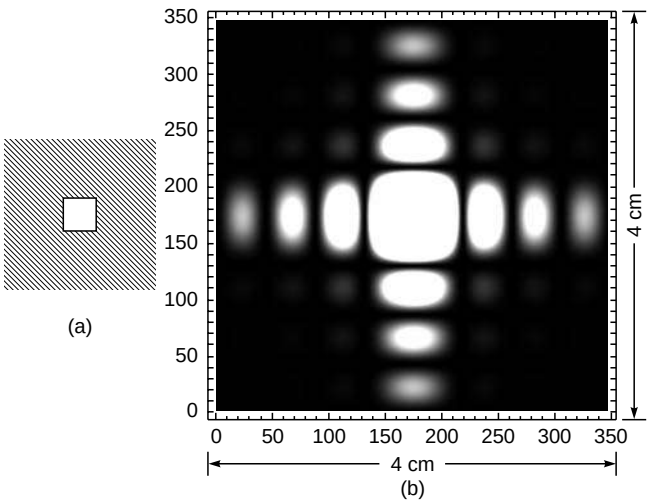


Fig. 19.3 (a) A square aperture of side 0.01 cm. (b) The corresponding (computer-generated) Fraunhofer diffraction pattern on a screen at a distance of 100 cm from the aperture; $\lambda = 5 \times 10^{-5}$ cm.

Now, if we observe in a region of radius 0.5 cm [i.e., $(x^2 + y^2) < 0.25\ \text{cm}^2$], then Eq. (13) gives

$$z >> \left[\frac{1}{4\times 5\times 10^{-5}}(0.25)^2\right]^{1/3} \approx 7\ \text{cm}$$

Further Eq. (17) gives

$$z >> \left[\frac{1}{5\times 10^{-5}}2\times(0.01)^2\right] \approx 4\ \text{cm}$$

We have chosen z = 100 cm, and we get the diffraction pattern as shown in Fig. 19.3. Although in the above we have assumed that we are observing a region of radius 0.5 cm, we have plotted the diffraction pattern for $-2\ \text{cm} < x, y < +2\ \text{cm}$. Note that along the x axis, the intensity will be zero when

$$\beta = \left(\frac{\pi b x}{\lambda z}\right) = m\pi; \qquad m = 0, 1, 2, 3, \ldots \quad \textbf{(32)}$$

or

$$x = \frac{m\lambda}{b} z$$
$$= 0.5\ \text{cm}, 1.0\ \text{cm}, 1.5\ \text{cm}, 2.0\ \text{cm}, \ldots$$

corresponding to m = 1, 2, 3, 4, . . ., respectively; this is consistent with the positions of the minima in Fig. 19.3.

For the case of a long narrow slit (i.e. for $a \to \infty$), the function

$$\frac{a \sin\gamma}{\gamma} = \frac{\sin(a\pi \sin\phi/\lambda)}{(\pi \sin\phi/\lambda)}$$

becomes very sharply peaked around $\phi = 0$. Since $\phi = 0$ implies $y = 0$, there is no diffraction along the y axis (see Sec. 19.4).

19.7 FRAUNHOFER DIFFRACTION BY A CIRCULAR APERTURE

We consider a plane wave incident normally on a circular aperture as shown in Fig. 19.4. On the plane of the circular aperture we choose cylindrical coordinates (see Fig. 19.5)

$$\xi = \rho\cos\phi \qquad \text{and} \qquad \eta = \rho\sin\phi \qquad \textbf{(33)}$$

Further, because of the circular symmetry of the system, the diffraction pattern will be of the form of concentric circular rings with their centers at point O'. Consequently, we may calculate the intensity distribution only along the x axis (i.e., at points for which $y = 0$) and in the final result replace x by $\sqrt{x^2 + y^2}$. Now, when $y = 0$,

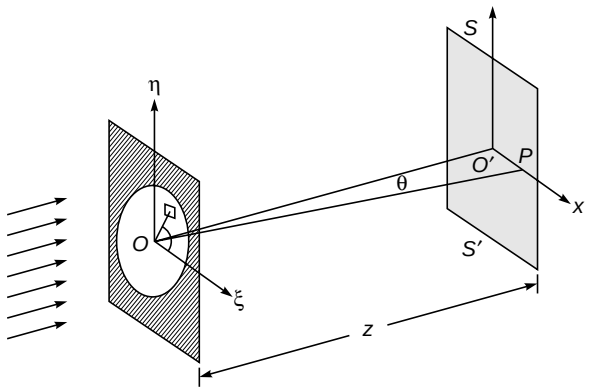


Fig. 19.4 Diffraction of a plane wave incident on a circular aperture of radius *a*.

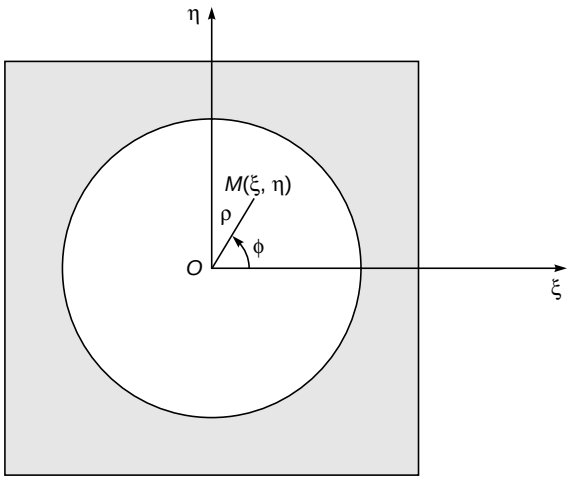


Fig. 19.5 Cylindrical coordinates (ρ, ϕ) on the plane of the circular aperture.

$$v = 0 \qquad \text{and} \qquad \sin\theta \approx \frac{x}{z} \qquad \textbf{(34)}$$

where θ is the angle that OP makes with the z axis. Thus

$$u = \frac{2\pi x}{\lambda z} = k\sin\theta$$

and Eq. (18) becomes

$$u(P) = \frac{A}{i\lambda z} e^{ikz} \exp\left(\frac{ikr^2}{2z}\right) \int_0^a \int_0^{2\pi} e^{-ik\rho\sin\theta\cos\phi}\, \rho\, d\rho\, d\phi \qquad \textbf{(35)}$$

Thus

$$u(P) = \frac{A}{i\lambda z} e^{ikz} \exp\left(\frac{ikr^2}{2z}\right) \frac{1}{(k\sin\theta)^2} \int_0^{ka\sin\theta} \zeta\, d\zeta \int e^{-i\zeta\cos\phi}\, d\phi$$

$$= \frac{A}{i\lambda z} e^{ikz} \exp\left(\frac{ikr^2}{2z}\right) \frac{2\pi}{(k\sin\theta)^2} \int_0^{ka\sin\theta} \zeta\, J_0(\zeta)\, d\zeta \qquad \textbf{(36)}$$

where $\zeta = k\rho \sin\theta$ and use has made of the following well-known relation[2]

$$J_0(\zeta) = \frac{1}{2\pi}\int_0^{2\pi} e^{\pm i\zeta\cos\phi}\, d\phi \qquad (37)$$

If we further use the relation

$$\frac{d}{d\zeta}[\zeta J_1(\zeta)] = \zeta J_0(\zeta) \qquad (38)$$

then Eq. (36) becomes

$$u(P) = \frac{A}{i\lambda z} e^{ikz} \exp\left(\frac{ikr^2}{2z}\right) \frac{2\pi}{(k\sin\theta)^2}\left[\zeta J_1(\zeta)\right]_0^{ka\sin\theta}$$

$$= \frac{A}{i\lambda z} e^{ikz} \exp\left(\frac{ikr^2}{2z}\right) \pi a^2 \left[\frac{2J_1(v)}{v}\right]$$

where $v = ka \sin\theta$. Thus the intensity distribution is given by

$$I(P) = I_0\left[\frac{2J_1(v)}{v}\right] \qquad (39)$$

where I_0 is the intensity at point O' (see Fig. 19.4). This is the famous Airy pattern which has been discussed in Sec. 18.3. We already mentioned that the diffraction pattern (in plane SS') will consist of concentric rings with their centers at point O'. If $F(r)$ represents the fractional energy contained in a circle of radius r, then

$$F(r) = \frac{\int_0^r I(\sigma)\, 2\pi\sigma\, d\sigma}{\int_0^\infty I(\sigma)\, 2\pi\sigma\, d\sigma} \qquad (40)$$

where $I(\sigma)2\pi\sigma\, d\sigma$ is proportional to the energy contained in the annular region whose radii lie between σ and $\sigma + d\sigma$. Clearly

$$\sin\theta \approx \frac{\sigma}{z} \qquad (41)$$

Since $v = ka \sin\theta$, we obtain

$$\sigma = \frac{z}{ka} v \qquad (42)$$

and Eq. (40) becomes

$$F(r) = \frac{\int_0^v \left[2J_1(v)/v\right]^2 v\, dv}{\int_0^\infty \left[2J_1(v)/v\right]^2 v\, dv} \qquad (43)$$

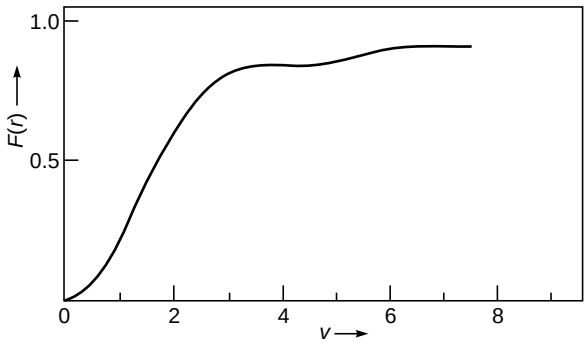


Fig. 19.6 The fractional energy contained in a circle of radius r.

where we have used Eq. (39) for the intensity distribution. Now

$$\frac{J_1^2(v)}{v} = J_1(v)\left[J_0(v) - \frac{dJ_1(v)}{dv}\right]$$

$$= -\left[J_0(v)\frac{dJ_0(v)}{dv} + J_1(v)\frac{dJ_1(v)}{dv}\right]$$

$$= -\frac{1}{2}\frac{d}{dv}\left[J_0^2(v) + J_1^2(v)\right] \qquad (44)$$

Thus

$$F(r) = \frac{J_0^2(v) + J_1^2(v)\Big|_0^v}{J_0^2(v) + J_1^2(v)\Big|_0^\infty} = 1 - J_0^2(v) - J_1^2(v) \qquad (45)$$

The above function is plotted in Fig. 19.6; one can deduce from the curve that about 84% of light is contained within the circle bounded by the first dark ring, and about 91% of the light is contained in the circle bounded by the first two dark rings, etc. The Fraunhofer pattern by an annular aperture is discussed in Prob. 19.5.

19.8 ARRAY OF IDENTICAL APERTURES

We next consider an array of N identical apertures as shown in Fig. 19.7. The Fraunhofer diffraction pattern will be the sum of the fields produced by the individual apertures and will be given by [see Eq. (18)]

$$u = C\left(\iint_{S_1} + \iint_{S_2} + \cdots\right) \exp\left[-i(u\xi + v\eta)\right] d\xi\, d\eta \qquad (46)$$

[2] The identities associated with Bessel functions can be found in most books on mathematical physics; see, e.g., Refs. 5 to 7.

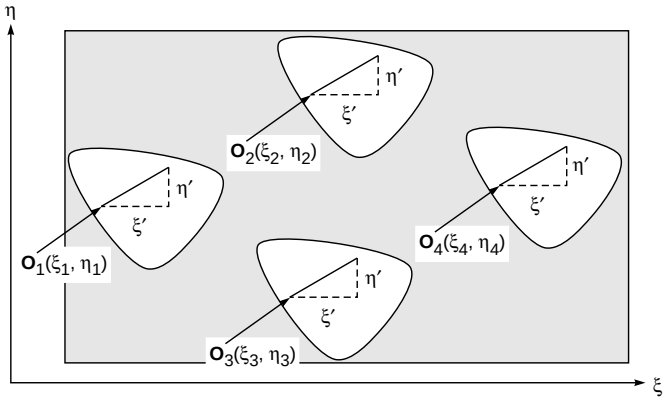


Fig. 19.7 Diffraction of a plane wave incident normally on an array of N identical apertures.

where each integral represents the contribution from a particular aperture. Let $O_1, O_2, O_3, \ldots$ represent points that are identically situated inside the apertures. For example, if the apertures are rectangular in nature, then $O_1, O_2, O_3, \ldots$ could represent the centers of the (rectangular) aperture. Let (ξ_1, η_1), (ξ_2, η_2), $(\xi_3, \eta_3), \ldots$ represent the coordinates of points $O_1, O_2, O_3, \ldots$, respectively; then

$$u = C \sum_{n=1}^{N} \iint e^{-i[u(\xi_n + \xi') + v(\eta_n + \eta')]} \, d\xi' \, d\eta' \qquad \textbf{(47)}$$

where (ξ', η') represents the coordinates of an arbitrary point in a given aperture with respect to the point (ξ_n, η_n) as shown in Fig. 19.7. Thus

$$u = u_s \sum_{n=1}^{N} e^{-i[u\xi_n + v\eta_n]} \qquad \textbf{(48)}$$

where

$$u_s = C \iint e^{-i(u\xi' + v\eta')} \, d\xi' \, d\eta' \qquad \textbf{(49)}$$

is the field produced by a single aperture. Thus, the resultant intensity distribution is given by

$$I = I_s I_1 \qquad \textbf{(50)}$$

where I_s represents the intensity produced by a single aperture and

$$I_1 = \left| \sum_{n=1}^{N} e^{-i[u\xi_n + v\eta_n]} \right|^2 \qquad \textbf{(51)}$$

represents the intensity distribution produced by N point sources.

As an example, we consider N equally spaced identical apertures as shown in Fig. 19.8. Without loss of generality, we may assume

$$\xi_n = (n-1)d \qquad \text{and} \qquad \eta_n = 0 \qquad n = 1, 2, 3, \ldots, N$$

Thus

$$\sum_{n=1,2,3}^{N} e^{-iu(n-1)d} = 1 + e^{-iud} + \cdots + e^{-i(N-1)ud} = \frac{1 - e^{-iNud}}{1 - e^{-iud}}$$

$$= \exp\left[-\frac{1}{2} i(N-1)ud\right] \frac{\sin N\gamma}{\sin \gamma} \qquad \textbf{(52)}$$

where

$$\gamma = \frac{ud}{2} = \frac{\pi d \sin \theta}{\gamma} \qquad \textbf{(53)}$$

and

$$\sin \theta = \frac{x}{z} \qquad \textbf{(54)}$$

We therefore obtain

$$I_1 = \left| \sum_{n=1,2,3\ldots}^{N} e^{-iu(n-1)d} \right|^2 = \frac{\sin^2 N\gamma}{\sin^2 \gamma} \qquad \textbf{(55)}$$

which is the interference pattern produced by N identically placed point sources; this is the same result as derived in Sec. 18.7. When $N = 2$, we obtain the interference pattern produced by two point sources.

If each aperture is a long narrow slit, we obtain the diffraction pattern produced by a grating [see Eq. (50) of Chap. 18]. On the other hand, if each aperture is circular, we obtain the product of the Airy pattern and the interference pattern produced by two point sources (see Figs 17.4 and 17.5).

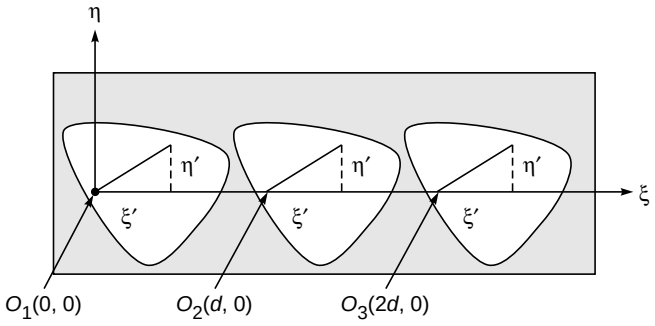


Fig. 19.8 Diffraction of a plane wave incident normally on an array of N identical equally spaced apertures.

19.9 SPATIAL FREQUENCY FILTERING

In the next section, we will show that if $g(x, y)$ represents the filed distribution on the front focal plane of a corrected lens (i.e., on plane P_1 in Fig. 19.9), then on the back focal plane P_2 of the lens, one obtains the Fourier transform of $g(x, y)$, and the z axis represents the optical axis of the lens. Thus if $G(x, y)$ represents the field distribution on the back focal plane P_2 then it is related to $g(x, y)$ through the following relation:

$$G(u,v) = \frac{1}{\lambda f}\iint g(x', y')\exp[-i(ux' + vy')]\,dx'\,dy' \qquad \textbf{(56)}$$

where,

$$u \equiv \frac{2\pi x}{\lambda f} \qquad \text{and} \qquad v \equiv \frac{2\pi y}{\lambda f} \qquad \textbf{(57)}$$

represent the spatial frequencies. Further, λ represents the wavelength of light and f is the focal length of the lens. If we compare Eq. (56) with Eq. (29) of Chap. 9, we find that

> *The field distribution on the back focal plane of a corrected lens is the Fourier transform of the field distribution on the front plane.*

This important property of a corrected lens forms the basis of the subject of spatial frequency filtering which finds applications in many diverse areas (see, e.g., Refs. 2, 4, and 8 to 10). We note here that in writing Eq. (56), we have neglected an (unimportant) phase factor on the right-hand side (see Sec. 19.11).

We first consider a plane wave incident normally on the lens. This implies that $g(x, y)$ is a constant (= g_0, say) and

$$G(u, v) = \frac{g_0}{\lambda f}\iint \exp[-i(ux' + vy')]\,dx'\,dy' \qquad \textbf{(58)}$$

Now, if we use

$$\int_{-\infty}^{+\infty} e^{-iux}dx = 2\pi\delta(u) \qquad \textbf{(59)}$$

[see Eq. (32) of Chap. 9], we obtain

$$G(u, v) = \frac{g_0}{\lambda f}\,4\pi^2\delta(u)\delta(v) \qquad \textbf{(60)}$$

where $\delta(u)$ and $\delta(v)$ represent the Dirac delta functions. Since $\delta(u) = 0$ for $u \neq 0$, one can infer from Eq. (60) that the intensity is zero at all points except at the point $x = 0$, $y = 0$. This is to be expected because a plane wave gets focused to a point by a corrected lens.[3]

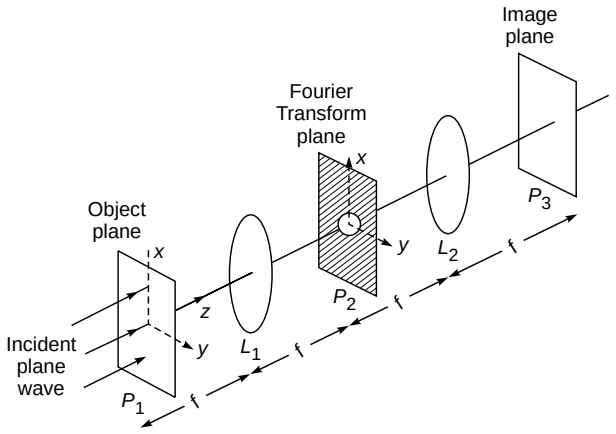


Fig. 19.9 Plane P_2 is the Fourier transform plane where the spatial frequency components of the object (placed in plane P_1) are displayed. In the above figure, a small hole is placed on the axis (in plane P_2) which filters out the high-frequency components.

Another interesting example is a one-dimensional cosinusoidal field distribution in the object plane, i.e.,

$$g(x, y) = g_0\cos(2\pi\alpha x) \qquad \textbf{(61)}$$

where α is a constant.[4] We have assumed no y dependence of the field. If we use the identity

$$\cos\theta = \tfrac{1}{2}(e^{i\theta} + e^{-i\theta})$$

we would obtain

$$G(u, v) = \frac{g_0}{\lambda f}\int\frac{1}{2}\left(e^{i2\pi\alpha x'} + e^{-i2\pi\alpha x'}\right)e^{-iux'}dx' \times \int e^{ivy'}dy' \qquad \textbf{(62)}$$

If we now use Eq. (59), we get

$$G(u, v) = \frac{g_0}{f}\,2\pi^2\,[\delta(u - 2\pi\alpha) + \delta(u + 2\pi\alpha)]\delta(v) \qquad \textbf{(63)}$$

Thus, we will obtain two spots in plane P_2. These two spots will be lying on the x axis (where $v = 0$) at $u = \pm 2\pi\alpha$ (i.e.; at $x = \pm\lambda f\alpha$). Physically this can be understood from the following consideration: When a plane wave is incident normally on plane P_1 (see Fig. 19.9), the time dependence is of the form $\cos\omega t$. If in plane P_1, we have an object whose

[3] We are assuming a very large dimension of the aperture of the lens; as such, the limits of integration in Eq. (56) are assumed to be from $-\infty$ to $+\infty$. This is a good approximation in most cases.

[4] On plane P_1 (see Fig. 19.7), if we place the negative of the photograph shown in Fig. 14.11(b) with the y axis along the length of a fringe and assume a plane wave to be incident normally on the film, then the field distribution is proportional to $\cos^2(2\pi\alpha x)$ which is equal to $\frac{1}{2}[1 + \cos(2\pi\alpha x)]$.

transmittance is proportional to cos ($2\pi\alpha x$), then the field to the right of plane P_1 is proportional to

$$\cos \omega t \cos (2\pi\alpha x)$$

$$= \frac{1}{2}[\cos (\omega t + 2\pi\alpha x) + \cos (\omega t - 2\pi\alpha x)] \qquad \textbf{(64)}$$

We know that for a plane wave with $k_y = 0$, the field variation is of the form (see Example 11.6)

$$\cos (\omega t - k_x x - k_z z) \qquad \textbf{(65)}$$

where $k_x = k \sin\theta$, $k_z = k \cos\theta$, $k = 2\pi/\lambda$, and θ is the angle that the propagation vector **k** makes with the z axis. At $z = 0$, the field becomes

$$\cos(\omega t - k_x x) \qquad \textbf{(66)}$$

Comparing the above equation with Eq. (64), we find that the two terms on the RHS of Eq. (64) represent two plane waves propagating along directions making angles $-\theta$ and $+\theta$ with the z axis, where

$$\sin\theta = \frac{k_x}{k} = \frac{2\pi\alpha}{2\pi/\lambda} = \alpha\lambda \qquad \textbf{(67)}$$

These plane waves will obviously focus to two points at $x = -\lambda f\alpha$ and $x = +\lambda f\alpha$ on the x axis in plane P_2. Since α represents the spatial frequency associated with the object, one essentially obtains, on the back focal plane, the spatial frequency spectrum of the object.

We are familiar with the fact that a general time varying signal can be expressed as a superposition of pure sinusoidal signals [see Eq. (33) of Chap. 9]. In a similar manner, the field variation across an arbitrary object (placed on plane P_1), can be expressed as a superposition of sinusoidal variations, and one would get the corresponding (spatial) frequency components on plane P_2. For this reason, plane P_2 is often termed as the Fourier transform plane.

As another example, if the amplitude variation of the object is of the form

$$g(x, y) = A \cos 2\pi\alpha x + B \cos 2\pi\beta x \qquad \textbf{(68)}$$

then one would obtain four spots on plane P_2 (all lying on the x axis); these spots will appear at $x = \pm\lambda f\alpha$, $\pm\lambda f\beta$. Since the Fourier transform of the Fourier transform is the original function itself[5] (see Chap. 9), if we place plane P_2 on the front focal plane of lens L_2, then on its back focal plane (i.e., in plane P_3 in Fig. 19.9) we will obtain the amplitude distribution associated with the object. If we now put stops at the points ($x = +\lambda f\alpha$, $y = 0$) and ($x = -\lambda f\alpha$, $y = 0$) on plane P_2, then the field distribution on plane P_3 is proportional to cos $2\pi\beta x$. Thus, we have been able to filter out the spatial frequency α. This is the basic principle behind spatial frequency filtering.

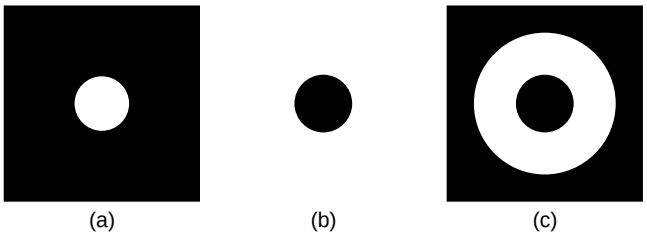


Fig. 19.10 (a) Low-pass filter, (b) high-pass filter, and (c) band-pass filter; filters are to be put on plane P_2.

For an arbitrary object, if we put a small hole on plane P_2, then it will filter out the high-frequency components [see Fig. 19.10(a)]; if we put a small stop on the axis, we filter out the low-frequency components [see Fig. 19.10(b)]. On the other hand, an annular aperture on plane P_2 will act as a band-pass filter as shown in Fig. 19.10(c).

As a simple application, we consider a halftone photograph (like that in a newspaper), which consists of a large number of spots of varying shades that produce the image pattern. Since the spots are closely spaced, it represents a high-frequency noise, and the overall image has much smaller frequencies associated with it. Thus, if we put a transparency similar to that shown in Fig. 19.11(a) and allow only the low-frequency components to pass through (as shown in Fig. 19.9), we will obtain, on plane P_3, an image which does not contain the unwanted high-frequency noise [see Fig 19.11(c)].

The subject of spatial frequency filtering finds applications in many other areas such as a contrast improvement, character recognition, etc. (see, e.g., Refs. 2, 4, and 9).

19.9.1 The 4f Correlator

The 4f correlator is based on the convolution theorem discussed in Sec. 9.7. A plane wave is assumed to be incident on a transparency containing one two-dimensional function $g(x, y)$ which is placed on the front focal plane of the first lens as shown in Fig. 19.12. The Fourier transform of $g(x, y)$ [= $G(u, v)$] is formed on the back focal plane of the lens. A transmission mask containing the Fourier transform of the second function $h(x, y)$ [= $H(u, v)$] is placed on this plane. Thus the product $G(u, v)H(u, v)$ lies on the front focal plane of the second lens, and therefore on its back focal plane, we will obtain the Fourier transform of $G(u, v)H(u, v)$ which is nothing but the convolution of $g(x, y)$ and $h(x, y)$. This concept is of considerable use in many applications (see, e.g., Refs. 2, 4, and 8 to10).

[5] There will, however, be an inversion; i.e., $f(x, y)$ will become $f(-x, -y)$ on plane P_3. This can also be seen by simple ray tracing.

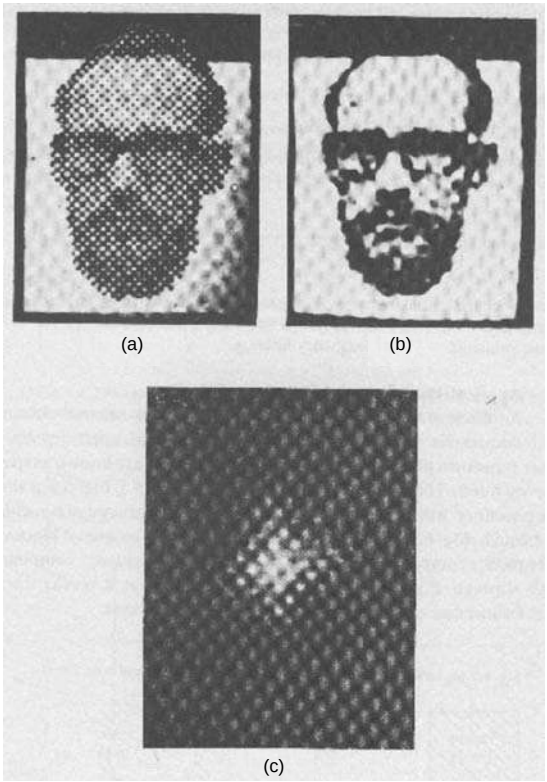


Fig. 19.11 (a) A photograph consisting of regularly spaced black-and-white squares of varying sizes. When a pinhole is placed in the Fourier transform plane to block the high-frequency components, an image of the form shown in (b) is obtained; the frequency spectrum is shown in (c). Notice that in (b) shades of gray appear as well details such as the missing part of the eyeglass frame [Photographs reprinted with permission from R. A. Phillips. "Spatial Filtering Experiments for Undergraduate Laboratories," *American Journal of Physics,* Vol. 37, 536, 1969; Copyright © 1969, American Association of Physics Teachers].

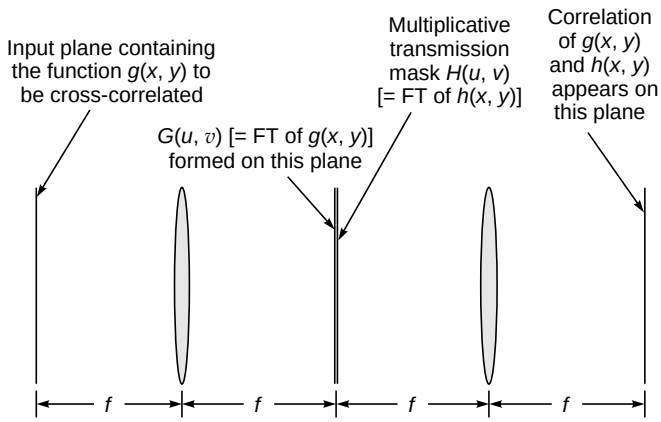


Fig. 19.12 The 4f correlator.

19.10 THE FOURIER TRANSFORMING PROPERTY OF A THIN LENS

In this section we will derive the Fourier transforming property of a thin lens [see Eq. (56)]. We will first show that the effect of a thin lens of focal length f is to multiply the incident field distribution by a factor p_L given by

$$p_L = \exp\left[-\frac{ik}{2f}(x^2 + y^2)\right] \tag{69}$$

Consider an object point O at a distance d_1 from an aberrationless thin lens of focal length f (see Fig. 19.13). If the image point I is at a distance d_2 from the lens, then d_2 is given by (see Sec. 4.4)

$$\frac{1}{d_1} + \frac{1}{d_2} = \frac{1}{f} \tag{70}$$

where d_1 and d_2 represent the magnitude of the distances of the object and image points from the lens. The phase factor corresponding to the disturbance emanating from point O is simply $\exp(+ikr)$, where r is the distance measured from point O. Now

$$r = (x^2 + y^2 + d_1^2)^{1/2} = d_1\left(1 + \frac{x^2 + y^2}{d_1^2}\right)^{1/2}$$

$$\approx d_1 + \frac{x^2 + y^2}{2d_1}$$

where in writing the last expression, we have assumed $x, y << d_1$; i.e., we have confined ourselves to a region close to the axis of the lens—this is known as the paraxial approximation. Thus, the phase distribution on the transverse plane P_2 at a distance d_1 from point O (i.e., immediately in front of the

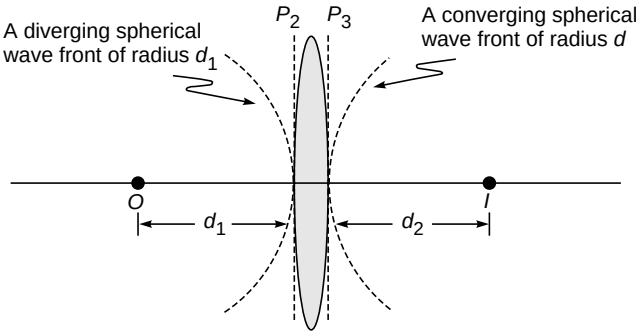


Fig. 19.13 Spherical waves emanating from an object point O, after refraction through a convex lens, emerge as spherical waves converging to the image point I.

lens—see Fig. 19.13) is given by

$$\exp(+ikr) \approx \exp\left[ik\left(d_1 + \frac{x^2+y^2}{2d_1}\right)\right]$$

Since the image is formed at I, the incident spherical wave emerges as another spherical wave of radius d_2, which under the paraxial approximation is

$$\exp\left[-ik\left(d_2 + \frac{x^2+y^2}{2d_2}\right)\right]$$

The negative sign inside the square brackets refers to the fact that we now have a converging spherical wave. Thus, if p_L represents the factor that when multiplied to the incident phase distribution gives the phase distribution of the emergent wave, then

$$\exp\left[-ik\left(d_2 + \frac{x^2+y^2}{2d_2}\right)\right] = \exp\left[+ik\left(d_1 + \frac{x^2+y^2}{2d_1}\right)\right] p_L$$

or

$$p_L = \exp[-ik(d_1+d_2)] \exp\left\{-\frac{ik}{2}\left[\left(\frac{1}{d_1}+\frac{1}{d_2}\right)(x^2+y^2)\right]\right\} \quad \textbf{(71)}$$

where the subscript L on p corresponds to the fact that we are referring to a lens. If we use Eq. (70) and neglect the first factor in the above equation, because it is independent of x and y, we obtain Eq. (69). Thus the effect of a thin lens on an incident field is to multiply the incident phase distribution by a factor that is given by Eq. (69). For a plane wave incident along the axis, the emerging disturbance will be simply p_L, which can be seen to be the paraxial approximation of a converging spherical wave front of radius f.

Now, let $g(x, y)$ represent the field distribution on plane P_1 (see Fig. 19.14). We first want to determine the field distribution on plane P_2, i.e., at a distance f from plane P_1 (see Fig. 19.14). Obviously the field will undergo Fresnel diffraction, and on plane P_2 it will be given by [using Eq. (9)]

$$u(x, y)|_{P_2} = \frac{1}{i\lambda f} \exp(ikf) \times \iint g(\xi, \eta) \times \exp\left\{\frac{ik}{2f}\right\} [(x-\xi)^2 + (x-\eta)^2]\, d\xi\, d\eta \quad \textbf{(72)}$$

Now, as shown earlier in this section, the effect of a thin lens of focal length f is to multiply the incident field distribution by the factor p_L given by Eq. (69). Thus on plane P_3, the field distribution will be given by

$$u(x, y)|_{P_3} = \frac{1}{i\lambda f} e^{ikf} \exp[-i\alpha(x^2+y^2)] \times \iint g(\xi, \eta) \times \exp\{i\alpha[(x-\xi)^2 + (y-\eta)^2]\}\, d\xi\, d\eta \quad \textbf{(73)}$$

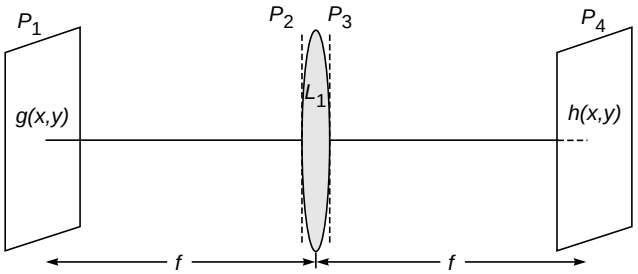


Fig. 19.14 A field distribution $g(x, y)$ placed at the front focal plane of a lens produces a field distribution $h(x, y)$ in plane P_4 at the back focal plane of the lens. The field $g(x, y)$ first undergoes Fresnel diffraction from plane P_1 to P_2, then it gets multiplied by a phase factor due to the presence of the lens, and the resultant field again undergoes Fresnel diffraction from plane P_3 to P_4 to produce the field distribution $h(x, y)$.

where

$$\alpha = \frac{k}{2f} = \frac{\pi}{\lambda f} \quad \textbf{(74)}$$

From plane P_3 the field will again undergo Fresnel diffraction, and therefore on plane P_4, it will be given by [using Eq. (9)]

$$u(x, y)|_{P_4} = \frac{1}{i\lambda f} e^{ikf} \iint u(\zeta, \tau)|_{P_3} \times \exp\{i\alpha[(x-\zeta)^2 + (y-\tau)^2]\}\, d\zeta\, d\tau \quad \textbf{(75)}$$

Substituting for $u|_{P_3}$ from Eq. (73), we get

$$u(x, y)|_{P_4} = \left(\frac{1}{i\lambda f} e^{ikf}\right)^2 I(x, y) \quad \textbf{(76)}$$

where

$$I(x, y) = \int\int_{-\infty}^{+\infty} g(\xi, \eta) H(x, y, \xi, \eta)\, d\xi\, d\eta \quad \textbf{(77)}$$

$$H(x, y, \xi, \eta) = \int\int_{-\infty}^{+\infty} \exp\{-i\alpha(\zeta^2+\tau^2)\} \times \exp\{i\alpha[(\zeta-\xi)^2 + (\tau-\eta)^2]\} \times \exp\{i\alpha[(x-\zeta)^2 + (y-\tau)^2]\}\, d\zeta\, d\tau = H_\xi(x) H_\eta(y) \quad \textbf{(78)}$$

$$H_\xi(x) = \int_{-\infty}^{+\infty} \exp\,[i\alpha(\xi^2 - 2\xi\zeta + x^2 - 2x\zeta + \zeta^2)] d\zeta \quad \textbf{(79)}$$

and a similar expression for H_η. Now,

$$\xi^2 - 2\xi\zeta + x^2 - 2x\zeta + \zeta^2 = \zeta^2 - 2\zeta(x+\xi) + (x+\xi)^2 - (x+\xi)^2 + \xi^2 + x^2 = (\zeta - g)^2 - 2x\xi$$

where $g = x + \xi$. Thus

$$H_\xi = \exp(-2i\alpha x\xi)\int_{-\infty}^{+\infty} \exp[i\alpha(\zeta - g)^2]\,d\zeta$$

or

$$H_\xi(x) = e^{-2i\alpha x\xi}\sqrt{\frac{\pi}{-i\alpha}} \qquad \textbf{(80)}$$

and a similar expression for $H_\eta(y)$. Thus

$$I(x, y) = \int\int_{-\infty}^{+\infty} g(\xi, \eta)\, H_\xi(x)\, H_\eta(y)\, d\xi\, d\eta$$

$$= \frac{\pi}{-i\alpha}\int\int_{-\infty}^{+\infty} g(\xi, \eta) e^{-2i\alpha(x\xi + y\eta)} d\xi\, d\eta$$

$$= i\lambda f \int\int_{-\infty}^{+\infty} g(\xi, \eta) e^{-i(u\xi + v\eta)} d\xi\, d\eta$$

where we have used Eq. (74) and

$$u = 2\alpha x = \frac{2\pi x}{\lambda f} \qquad \text{and} \qquad v = 2\alpha y = \frac{2\pi y}{\lambda f} \qquad \textbf{(81)}$$

represent the spatial frequencies in the x and y directions, respectively. If we substitute the above expression for $I(x, y)$ in Eq. (76), we obtain

$$u(x, y)|_{P_4} = \frac{1}{\lambda f}\int\int_{-\infty}^{+\infty} g(\xi, \eta)\, e^{-i(u\xi + v\eta)} d\xi\, d\eta$$

where we have neglected the unimportant constant phase factors. Equation (81) is the same as Eq. (56) and gives the important result that

The field distribution on the back focal plane of a corrected lens is the Fourier transform of the field distribution on the front plane.

Note that in writing the limits in the integral from $-\infty$ to $+\infty$, we have assumed the lens to be of infinite extent; the error involved is usually very small because in almost all practical cases

$$a/\lambda >>> 1$$

where a represents the aperture of the lens.

Summary

- If the amplitude and phase distribution on the plane $z = 0$ are given by $A(\xi, \eta)$, then the Fresnel diffraction pattern is given by

$$u(x, y, z) \approx \frac{1}{i\lambda z} e^{ikz}\iint A(\xi, \eta) \times \exp\left\{\frac{ik}{2z}[(x - \xi)^2 + (y - \eta)^2\right\} d\xi\, d\eta$$

where $k = \dfrac{2\pi}{\lambda}$

- The Fraunhofer diffraction pattern is the Fourier transform of the aperture function and is given by

$$u(x, y, z) \approx \frac{1}{i\lambda z} e^{ikz} \exp\left[\frac{ik}{2z}(x^2 + y^2)\right] \times \iint A(\xi, \eta) e^{-i(u\xi + v\eta)}\, d\xi\, d\eta$$

For a plane wave incident normally on a circular aperture of radius a, the Fraunhofer diffraction pattern is given by

$$I(P) = I_0\left[\frac{2J_1(v)}{v}\right]$$

where $v = ka \sin\theta$.

- If $g(x, y)$ and $G(x, y)$ represent the filed distributions on the front focal plane and on the back focal plane of a corrected lens, then

$$G(u, v) = \frac{1}{\lambda f}\iint g(x', y') \exp[-i(ux' + vy')]\, dx'\, dy'$$

where $u \equiv 2\pi x/\lambda f$ and $\nu \equiv 2\pi y/\lambda f$ represent the spatial frequencies. Thus on the back focal plane of the lens one obtains the Fourier transform of $g(x, y)$; the z axis represents the optical axis of the lens. This important property of a corrected lens forms the basis of the subject of spatial frequency filtering.

Problems

19.1 Consider a rectangular aperture of dimensions 0.2 mm × 0.3 mm with a screen placed at a distance of 100 cm from the aperture. Assume a plane wave with $\lambda = 5 \times 10^{-5}$ cm incident normally on the aperture. Calculate the positions of maxima and minima in a region 0.2 cm × 0.2 cm of the screen. Show that both Fresnel and Fraunhofer approximations are satisfied.

19.2 In Prob. 19.1, assume a convex lens (of focal length 20 cm) placed immediately after the aperture. Calculate the positions of the first three maxima and minima on the x axis (implying $\phi = 0$) and also on the y axis (implying $\theta = 0$).

19.3 The Fraunhofer diffraction pattern of a circular aperture (of radius 0.5 mm) is observed on the focal plane of a convex lens of focal length 20 cm. Calculate the radii of the first and second dark rings. Assume $\lambda = 5.5 \times 10^{-5}$ cm.

[**Ans:** 0.13 mm, 0.18 mm]

19.4 In Prob. 19.3, calculate the area of the patch (on focal plane) which will contain 95% of the total energy.

19.5 Obtain the diffraction pattern of an annular aperture bounded by circles of radii a_1 and a_2 ($> a_1$). [*Hint:* The integration limits of ρ in Eq. (103) must be a_1 and a_2.]

REFERENCES AND SUGGESTED READINGS

1. M. Born and E. Wolf, *Principles of Optics*, 7th ed., Cambridge University Press, Cambridge, United Kingdom, 1999.
2. J. W. Goodman, *Introduction to Fourier Optics*, 3d ed., Roberts & Co., Englewood, Colo., 2005.
3. M.V. Klein and T. E. Furtak, *Optics*, Wiley, New York, 1986.
4. A. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978. Reprinted by Macmillan India, New Delhi, 1981.
5. J. Irving and N. Mullineux, *Mathematics in Physics and Engineering*, Academic Press, New York, 1959.
6. G. Arfken, *Mathematical Methods for Physicists*, 2d ed., Academic press, New York, 1970.
7. A. K. Ghatak, I.C. Goyal, and S. J. Chua, *Mathematical Physics*, Macmillan India, New Delhi, 1985.
8. E. G. Steward, *Fourier Optics: An Introduction*, 2d ed., Dover Publications, New York, 2004.
9. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, New York, 1991.
10. E. Hecht, *Optics*, Pearson Education, Singapore, 2002. http://en.wikipedia.org/wiki/Fourier_optics.

Chapter Twenty

FRESNEL DIFFRACTION

One of your commissioners, M. Poisson, had deduced from the integrals reported by the author [Fresnel] the singular result that the centre of the shadow of an opaque circular screen must, when the rays penetrate there at incidences which are only a little oblique, be just as illuminated as if the screen did not exist. The consequences have been submitted to the test of a direct experiment, and observation has perfectly confirmed the calculation.

—Dominique Arago to the French Academy of Sciences[1]

Important Milestones

1816	*Augustin Fresnel developed the theory of diffraction using the wave theory of light.*
1817	*Using Fresnel's theory, Poisson predicted a bright spot at the center of the shadow of an opaque disc—this is usually referred to as the Poisson spot.*
1818	*Fresnel and Arago carried out the experiment to demonstrate the existence of the Poisson spot, validating the wave theory.*
1874	*Marie Cornu developed a graphical approach to study Fresnel diffraction—this came to be known as Cornu's spiral.*

20.1 INTRODUCTION

In Chap. 18 we had mentioned that the phenomenon of diffraction can be broadly classified under two categories: Under the first category comes the Fresnel class of diffraction in which either the source or the screen (or both) is at a finite distance from the diffracting aperture. In the second category comes the Fraunhofer class of diffraction (discussed in the last two chapters) in which the wave incident on the aperture is a plane wave and the diffraction pattern is observed on the focal plane of a convex lens, so that the screen is effectively at an infinite distance from the aperture. In this chapter. we will discuss the Fresnel class of diffraction and also study the transition to the Fraunhofer region. The underlying principle in the entire analysis is the Huygens–Fresnel principle according to which

> *Each point on a wave front is a source of secondary disturbance, and the secondary wavelets emanating from different points mutually interfere.*

To appreciate the implications of this principle, we consider the incidence of a plane wave on a circular hole of radius a as shown in Fig. 20.1. In Sec. 18.3 we showed that the beam will undergo diffraction divergence and the angular spreading will be given by

$$\Delta\theta \sim \frac{\lambda}{2a}$$

Thus, when $a >>> \lambda$, the intensity at a point R (which is deep inside the geometrical shadow) will be negligible; on the other hand, if $a \sim \lambda$, there will be almost uniform spreading out of the beam, resulting in an (almost) uniform illumination of the screen. This phenomenon is a manifestation of the fact that when $a >>> \lambda$, the secondary wavelets emanating from different points on the circular aperture so beautifully interfere to produce (almost) zero intensity in the geometrical shadow and a large intensity inside the circular region (see Fig. 20.1). However, if $a \sim \lambda$, then the aperture almost acts as a point source, resulting in a uniform illumination of the screen (see Fig. 12.3).

[1] The author found this quotation in Ref. 1.

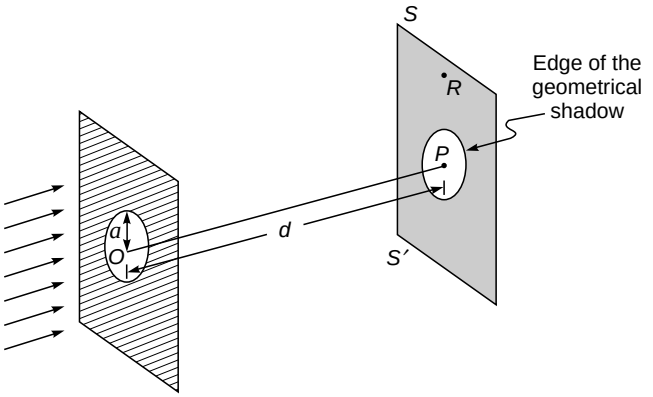


Fig. 20.1 Diffraction of a plane wave incident normally on a circular aperture of radius a.

We will first introduce the concept of Fresnel half-period zones to have a qualitative understanding of the Fresnel diffraction pattern; this will be followed by a more rigorous analysis of the Fresnel class of diffraction and its transition to the Fraunhofer region.

20.2 FRESNEL HALF-PERIOD ZONES

Let us consider a plane wave front WW' propagating in the z direction as shown in Fig. 20.2. To determine the field at an arbitrary point P due to the disturbances reaching from different portions of the wave front, we make the following construction: From point P we drop a perpendicular PO on the wave front. If $PO = d$, then with point P as center we draw spheres of radii $d + \lambda/2$, $d + 2\lambda/2$, $d + 3\lambda/2$, . . ., these spheres will intersect WW' in circles as shown in Fig. 20.2. The radius

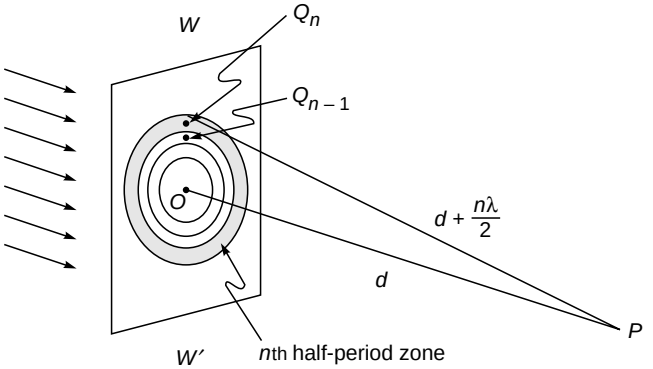


Fig. 20.2 Construction of Fresnel half-period zones.

of the nth circle will obviously be given by

$$r_n = \left[\left(d + n\frac{\lambda}{2}\right)^2 - d^2\right]^{1/2}$$

$$= \sqrt{n\lambda d}\left(1 + \frac{n\lambda}{4d}\right)^{1/2}$$

or

$$r_n \approx \sqrt{n\lambda d} \qquad \textbf{(1)}$$

where we have assumed $d >>> \lambda$; this is indeed justified for practical systems using visible light. Of course, we are assuming that n is not a very large number. The annular region between the nth circle and $(n - 1)$st circle is known as the nth half-period zone; the area of the nth half-period zone is given by

$$A_n = \pi r_n^2 - \pi r_{n-1}^2$$

$$\approx \pi [n\lambda d - (n-1)\lambda d] = \pi\lambda d \qquad \textbf{(2)}$$

Thus the areas of all the half-period zones are approximately equal. Now the resultant disturbance produced by the nth zone will be π out of phase with the disturbance produced by the $(n - 1)$st [or the $(n + 1)$st] zone. This can be easily seen from the following consideration: For infinitesimal area surrounding a point Q_n in the nth half-period zone, there is a corresponding infinitesimal area surrounding point Q_{n-1} in the $(n - 1)$st half-period zone such that

$$Q_nP - Q_{n-1}P = \frac{\lambda}{2}$$

which corresponds to a phase difference of π. Since the areas of the zones are approximately equal, one can have a one-to-one correspondence between points in various zones. Thus, the resultant amplitude at point P can be written as

$$u(P) = u_1 - u_2 + u_3 - u_4 + \cdots + (-1)^{m+1}u_m + \cdots \qquad \textbf{(3)}$$

where u_n represents the net amplitude produced by the secondary wavelets emanating from the nth zone; the alternate negative and positive signs represent the fact that the resultant disturbances produced by two consecutive zones are π out of phase with respect to each other. The amplitude produced by a particular zone is proportional to the area of the zone and inversely proportional to the distance of the zone from point P; further, it also depends on an obliquity factor which is proportional to $\frac{1}{2}(1 + \cos\chi)$, where χ is the angle that the normal to the zone makes with line QP; this obliquity factor comes out automatically from rigorous diffraction theory.[2] Thus we may write

$$u_n = \text{constant}\ \frac{A_n}{Q_nP}\frac{1 + \cos\chi}{2} \qquad \textbf{(4)}$$

[2] See, e.g., Ref. 2.

where A_n represents the area of the nth zone. It can be shown that if we use the exact expression for r_n, the area of the zones increases with n; however, this slight increase in area is exactly compensated by the increased distance of the zone from point P. In spite of this, the amplitudes u_1, u_2, $u_3 \ldots$ decrease monotonically because of increased obliquity. Thus we may write

$$u_1 > u_2 > u_3 > \cdots \tag{5}$$

The series expressed by Eq. (3) can be approximately summed due to a method by Schuster. We rewrite Eq. (3) as

$$u(P) = \frac{u_1}{2} + \left[\frac{u_1}{2} - u_2 + \frac{u_3}{2}\right] + \left[\frac{u_3}{2} - u_4 + \frac{u_5}{2}\right] + \cdots \tag{6}$$

where the last term is either $\frac{1}{2}u_m$ or $\frac{1}{2}u_{m-1} - u_m$ according to whether m is odd or even. If the obliquity factor is such that

$$u_n > \tfrac{1}{2}(u_{n-1} + u_{n+1}) \tag{7}$$

then the quantities inside the brackets in Eq. (6) will be negative; consequently,

$$\begin{aligned} u(P) &< \frac{1}{2}u_1 + \frac{1}{2}u_m \quad (m \text{ odd}) \\ u(P) &< \frac{1}{2}u_1 + \frac{1}{2}u_{m-1} - u_m \approx \frac{u_1}{2} - \frac{u_m}{2} \quad (m \text{ even}) \end{aligned} \tag{8}$$

where we have assumed that the amplitudes of the fields produced by consecutive zones differ only slightly. To obtain the upper limits, we rewrite Eq. (3) in the form

$$u(P) = u_1 - \frac{u_2}{2} - \left[\frac{u_2}{2} - u_3 + \frac{u_4}{2}\right] - \left[\frac{u_4}{2} - u_5 + \frac{u_6}{2}\right] \tag{9}$$

where the last term is now $-\frac{1}{2}u_{m-1} + u_m$, when m is odd, and $-\frac{1}{2}u_m$; when m is even. Since the quantities inside the brackets are negative, we obtain

$$\begin{aligned} u(P) &> u_1 - \frac{u_2}{2} - \frac{u_{m-1}}{2} + u_m \approx \frac{u_1}{2} + \frac{u_m}{2} \quad m \text{ odd} \\ u(P) &> u_1 - \frac{u_2}{2} - \frac{u_m}{2} \approx \frac{u_1}{2} - \frac{u_m}{2} \quad m \text{ even} \end{aligned} \tag{10}$$

Using Eqs. (8) and (10), we may approximately write

$$\begin{aligned} u(P) &\approx \frac{u_1}{2} + \frac{u_m}{2} \quad m \text{ odd} \\ u(P) &\approx \frac{u_1}{2} - \frac{u_m}{2} \quad m \text{ even} \end{aligned} \tag{11}$$

If we can neglect u_m in comparison to u_1, then the Eq. (11)[3] gives the remarkable result that

$$u(P) \approx \frac{u_1}{2} \tag{12}$$

implying that *the resultant amplitude produced by the entire wave front is only one-half of the amplitude produced by the first half-period zone.*

20.2.1 Diffraction by a Circular Aperture

We may use the above analysis to study the diffraction of a plane wave by a circular aperture. Let point P be at a distance d from the circular aperture (see Fig. 20.1). We assume that the radius of the circular aperture a can be increased from zero onward. As a increases, the intensity at point P will also increase until the circular aperture contains the first half-period zone; this happens when $a = \sqrt{\lambda d}$. The resultant amplitude at point P is u_1 which is twice the value of the amplitude for the unobstructed wave front [see Eq. (12)]. The intensity is therefore $4I_0$, where I_0 represents the intensity at point P due to the unobstructed wave front. If we further increase a, then $u(P)$ will start decreasing and when the circular aperture contains the first two half-period zones (which happens when $a = \sqrt{2\lambda d}$), the resultant amplitude ($= u_1 - u_2$) is almost zero. *Thus, by increasing the hole diameter, the intensity at point P decreases almost to zero*. This interesting result is once again due to the validity of the Huygens–Fresnel principle and hence would be valid for sound waves also. We may generalize the above result by noting that if

$$a = \sqrt{(2n+1)\lambda d} \qquad n = 0, 1, 2, \ldots \text{ (maxima)}$$

the aperture will contain an odd number of half-period zones and the intensity will be maximum; on the other hand, if

$$a = \sqrt{2n\lambda d} \qquad n = 1, 2, \ldots \text{ (minima)}$$

[3] If one assumes a form of the obliquity factor as given by Eq. (4), then it decreases from 1 to $\frac{1}{2}$ as m increases from 1 to ∞; this implies that $|u_m|$ can never be smaller than $u_1/2$. However, when m is large, a slight shift of point P on the axis will change the amplitude from $u_1/2 + u_m/2$ to $u_1/2 - u_m/2$; the changes will occur with such great rapidity that one can only observe the average value which will be $u_1/2$.

the aperture will contain an even number of half-period zones and the intensity will be minimum. To have a numerical appreciation, we note that for $d = 50$ cm and $\lambda = 5 \times 10^{-5}$ cm, the radii of the first, second, and third zones are 0.500, 0.707, and 0.866 mm, respectively. As a corollary of the above analysis, we can consider a circular aperture of a fixed radius a and study the intensity variation along the axis. Whenever the distance

$$d = \frac{a^2}{(2n+1)\lambda} \qquad n = 0, 1, 2, \ldots \text{ (maxima)}$$

point P (see Fig. 20.1) will correspond to a maximum. Similarly, when

$$d = \frac{a^2}{2n\lambda} \qquad n = 1, 2, \ldots \text{ (minima)}$$

point P will correspond to a minimum. The intensity distribution on screen SS' at off-axis points can be approximately calculated by using the half-period zones, but such a calculation is fairly cumbersome. However, from the symmetry of the problem, one can deduce that the diffraction pattern has to be in the form of concentric circular rings with their centers at point P.

20.2.2 Diffraction by an Opaque Disc—The Poisson Spot

If instead of the circular aperture we have a circular disc [see Fig. 20.3(a)] and if the disc obstructs the first p half-period zones, then the field at point P is

$$u(P) = u_{p+1} - u_{p+2} + \cdots \approx \frac{u_{p+1}}{2} \qquad \textbf{(13)}$$

Thus, we should always obtain a bright spot on the axis behind a circular disc (the more rigorous theory also predicts the same result—see Sec. 20.4.2). This is called the *Poisson spot*. In 1816 the French physicist Augustin Fresnel developed the mathematical theory of diffraction using the wave theory of light. Simeon Poisson, the famous mathematician, used Fresnel's theory to predict a bright spot at the center of the shadow of an opaque disc. Poisson was a great supporter of the corpuscular theory of light, and he said that since the bright spot is against common sense, the wave theory must be wrong. Shortly afterward, Fresnel and Arago carried out the experiment to demonstrate the existence of the Poisson spot [see Fig. 20.3(b)], validating the wave theory.

20.3 THE ZONE PLATE

A beautiful application of the concept of Fresnel half-period zones lies in the construction of the zone plate which consists of a large number of concentric circles whose radii are proportional to the square root of natural numbers and the alternate annular regions of which are blackened (see Fig. 20.4). Let the radii of the circles be $\sqrt{1}\,K$, $\sqrt{2}\,K$, $\sqrt{3}\,K$, $\sqrt{4}\,K, \ldots$ where K is a constant and has the dimension of length. We consider a point P_1 which is at a distance K^2/λ from the zone plate; for this point the blackened

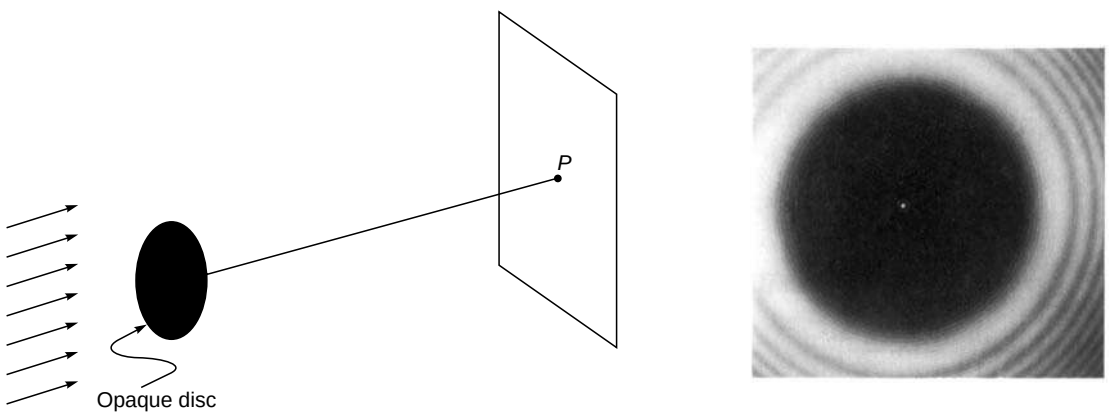


Fig. 20.3 (a) When a plane wave is incident normally on an opaque disc, a bright spot is always formed on an axial point. This spot is known as the Poisson spot. (b) The Poisson spot at the center of the shadow of a penny; the screen is 20 m from the coin, and the source of light is also 20 m from the coin [Photograph reprinted with permission from P. M. Rinard, "Large Scale Diffraction Patterns from Circular Objects," *American Journal of Physics,* Vol. 44, p. 70, 1976; Copyright 1976, American Association of Physics Teachers].

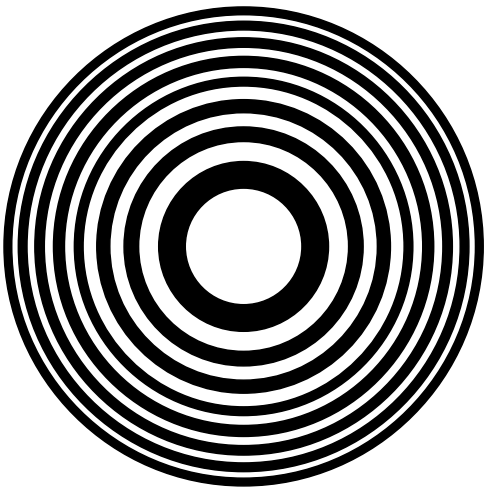


Fig. 20.4 The zone plate.

rings correspond to the 2nd, 4th, 6th, . . . half-period zones. Thus, the even zones are obstructed and the resultant amplitude at P_1 [see Fig. 20.5(a)] will be

$$u_1 + u_3 + u_5 + \cdots \tag{14}$$

producing an intense maximum. For point P_3 (which is at a distance $K^2/3\lambda$) the first blackened ring contains the 4th, 5th, 6th zones, the second blackened ring contains the 10th, 11th and 12th zones, etc.; thus the resultant amplitude is

$$(u_1 - u_2 + u_3) + (u_7 - u_8 + u_9) + \cdots \tag{15}$$

which would again correspond to a maximum, but it would not be as intense as point P_1. Between points P_1 and P_3 there will be a point P_2 (at a distance $K^2/2\lambda$) where the resultant amplitude is

$$(u_1 - u_2) + (u_5 - u_6) + \cdots \tag{16}$$

implying that corresponding to P_2 the first blackened ring contains the 3rd and 4th half-period zones, etc. Obviously, point P_2 will correspond to a minimum. Thus, if a plane wave is incident normally on a zone plate, then the corresponding focal points are at distances

$$\frac{K^2}{\lambda}, \frac{K^2}{3\lambda}, \frac{K^2}{5\lambda}, \ldots \tag{17}$$

from the zone plate. Elementary calculations will show that the zone plate suffers from considerable chromatic aberrations (see Prob. 20.5).

Example 20.1 Assume a plane wave ($\lambda = 5 \times 10^{-5}$ cm) to be incident on a circular aperture of radius 0.5 mm. We will calculate the positions of the brightest and darkest points on the axis. For the brightest point, the aperture should contain only the first zone, and thus we must have (see Fig. 20.1)

$$(0.05)^2 = OP(5 \times 10^{-5})$$

Thus $OP = 50$ cm. Similarly the darkest point would be at a distance

$$\frac{(0.05)^2}{2 \times 5 \times 10^{-5}} = 25 \text{ cm}$$

Example 20.2 Consider a zone plate with radii

$$r_n = 0.1\sqrt{n} \quad \text{cm}$$

For $\lambda = 5 \times 10^{-5}$ cm, we will calculate the positions of various foci. The most intense focal point will be at a distance

$$\frac{r_1^2}{\lambda} = \frac{0.01}{5 \times 10^{-5}} = 200 \text{ cm}$$

The other focal points will be at distances of 200/3, 200/5,and 200/7 cm, etc. Between any two consecutive foci there will be dark points on the axis corresponding to which the first circle will contain an even number of half-period zones.

The zone plate can also be used for imaging points on the axis; e.g., if we have a point source at S, then a bright image will be formed at P, where point P should be such that [see Fig. 20.5(b)]

$$SL + LP - SP = \frac{\lambda}{2} \tag{18}$$

the point L being on the periphery of the first circle of the zone plate [see Fig. 20.5(b)]. If the radius of the first circle is r_1, then

$$\begin{aligned} SL + LP - SP &= \sqrt{a^2 + r_1^2} + \sqrt{b^2 + r_1^2} - (a + b) \\ &\approx a\left(1 + \frac{r_1^2}{2a^2}\right) + b\left(1 + \frac{r_1^2}{2b^2}\right) - (a + b) \\ &\approx \frac{r_1^2}{2}\left(\frac{1}{a} + \frac{1}{b}\right) \end{aligned} \tag{19}$$

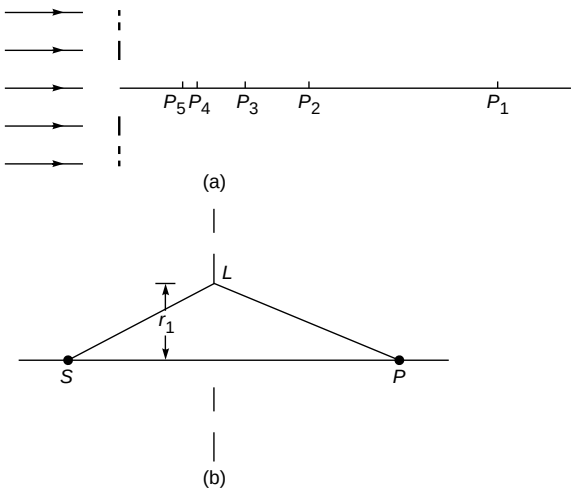


Fig. 20.5 (a) For a plane wave incident on a zone plate, the maximum intensity occurs at points P_1, P_3, etc. The minima occur at P_2, P_4, ... (b) Imaging of a point object by a zone plate.

Thus Eq. (18) becomes

$$\frac{1}{a}+\frac{1}{b}=\frac{1}{f} \tag{20}$$

where $f = r_1^2/\lambda$ represents the focal length. Equation (20) resembles the lens law. A very interesting demonstration experiment of the zone plate can be carried out by using microwave sources ($\lambda \sim 1$ cm) and, instead of the dark rings, having aluminum rings on a perspex sheet of dimension ~40 cm × 40 cm.

20.4 FRESNEL DIFFRACTION—A MORE RIGOROUS APPROACH

In Sec. 19.2 we gave a more rigorous analysis of the diffraction of a plane wave by different types of aperture. We considered a plane wave (of amplitude A) incident normally on an aperture as shown in Fig. 20.6. Using the Huygens–Fresnel principle, we showed that the field produced at point P on screen SS' (which is at a distance d from the aperture) is given by

$$u(P) = \frac{A}{i\lambda}\iint \frac{e^{ikr}}{r}\, d\xi\, d\eta \tag{21}$$

where the integration is over the area of the aperture. Now, if the amplitude and phase distribution on the plane $z = 0$ is given by $A(\xi, \eta)$, then the above integral is modified to

$$u(P) = \frac{1}{i\lambda}\iint A(\xi, \eta)\frac{e^{ikr}}{r}\, d\xi\, d\eta \tag{22}$$

Further, in the Fresnel approximation [see Eq. (9) of Chap.19] the above integral takes the form

$$u(x, y, z) \approx \frac{1}{i\lambda z}\, e^{ikz}\iint A(\xi, \eta) \times \exp\left\{\frac{ik}{2z}[x-\xi)^2 + (y-\eta)^2]\right\} d\xi\, d\eta \tag{23}$$

20.4.1 Diffraction of a Plane Wave Incident Normally on a Circular Aperture

We assume a plane wave incident normally on a circular aperture of radius a as shown in Fig. 20.7. The z axis is normal to the plane of the aperture, and screen SS' is assumed to be normal to the z axis. It is obvious from the symmetry of the problem that we will obtain circular fringes on screen SS'; however, it is very difficult to calculate the actual intensity variation on the screen. Therefore, for the sake of mathematical simplicity, we will calculate the variation of intensity only along the z axis. Obviously, it will be more convenient to use the circular system of coordinates. In this system, the coordinates of an arbitrary point M on the aperture will be (ρ, ϕ), where ρ is the distance the point M from the center O and ϕ is the angle that OM makes with the ξ axis (see Fig. 20.7), and a small element area dS surrounding point M will be $\rho\, d\rho\, d\phi$. Thus, using Eq. (21) we get

$$u(P) \approx -\frac{iA}{\lambda}\int_0^{2\pi}\int_0^{a}\frac{e^{ikr}}{r}\rho\, d\rho\, d\phi \tag{24}$$

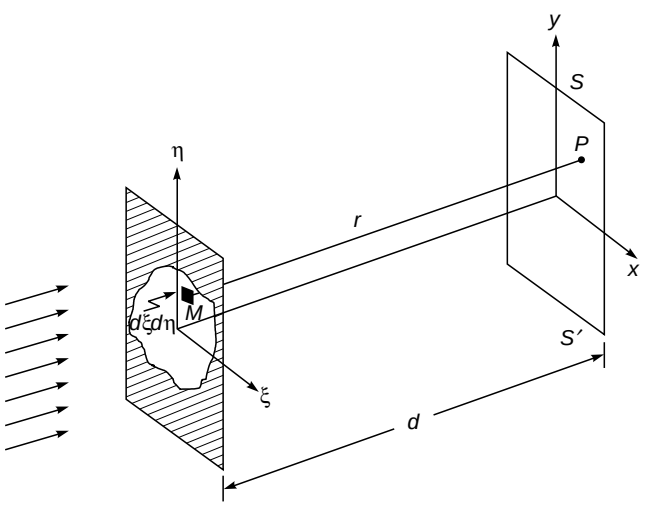


Fig. 20.6 A plane wave incident normally on an aperture.

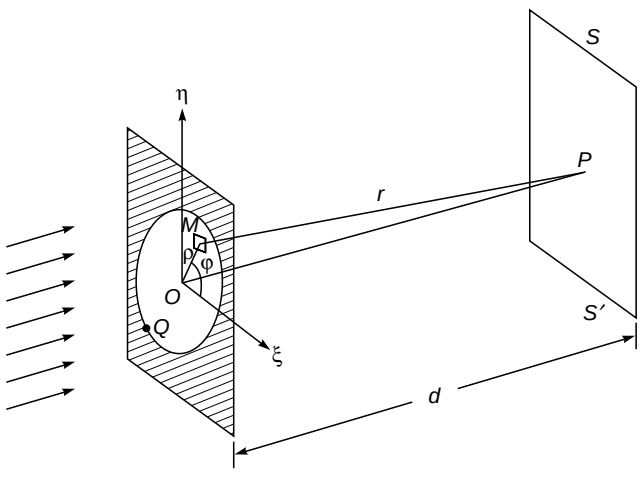


Fig. 20.7 Diffraction of a plane wave incident normally on a circular aperture of radius *a;* point Q is an arbitrary point on the periphery of the aperture.

Now

$$\rho^2 + d^2 = r^2$$

Thus

$$\rho\, d\rho = r\, dr$$

and Eq. (24) becomes

$$u(P) \approx -\frac{iA}{\lambda} \int_0^{2\pi} \int_d^{\sqrt{a^2+d^2}} e^{ikr}\, dr\, d\phi \qquad \textbf{(25)}$$

The integration is very simple, and since $k = 2\pi/\lambda$, we readily obtain

$$u(P) \approx Ae^{ikd}\,(1 - e^{ip\pi}) \qquad \textbf{(26)}$$

where we have defined p by the equation

$$k\left(\sqrt{a^2 + d^2} - d\right) = p\pi$$

The above equation implies

$$QP - OP = \frac{p\lambda}{2}$$

where Q is a point on the periphery of the circular aperture (see Fig. 20.7). From Eq. (26) we readily get

$$I(P) = 4I_0 \sin^2 \frac{p\pi}{2} \qquad \textbf{(27)}$$

where I_0 is the intensity associated with the incident plane wave. Equation (27) tells us that the intensity is zero or maximum when p is an even or odd integer, i.e., when $QP - OP$ is an even or odd multiple of $\lambda/2$. This can be understood physically by using the concept of Fresnel half-period zones discussed in Sec. 20.2. Thus, if the aperture contains an even number of half-period zones, the intensity at point P will be negligibly small; and conversely, if the circular aperture contains an odd number of zones, the intensity at P will be maximum. Now, when $d << a$ (as is usually the case)

$$p \approx \frac{k}{\pi}\left[d\left(1 + \frac{a^2}{2d^2}\right) - d\right]$$

or

$$p \approx \frac{a^2}{\lambda d} \qquad \textbf{(28)}$$

which is known as the Fresnel number of the aperture. In Fig. 20.8 we have plotted the corresponding intensity variation as a function of the dimensionless parameter

$$\frac{\lambda d}{a^2}$$

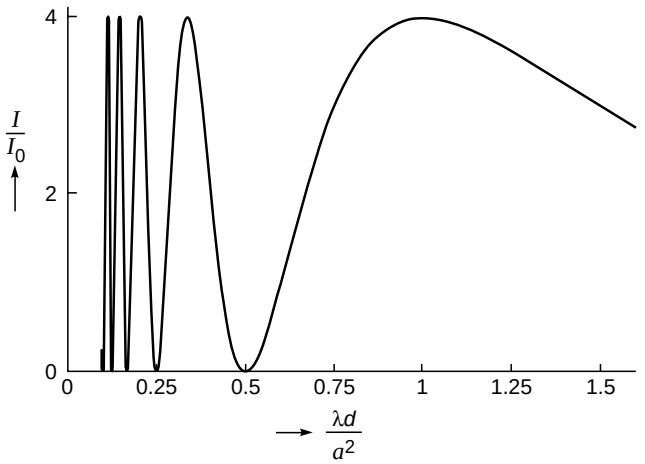


Fig. 20.8 The intensity variation on an axial point corresponding to a plane wave incident on a circular aperture of radius a.

The figure shows that when the (circular) aperture contains an even number of half-period zones, the intensity at point P will be zero; and when the aperture contains an odd number of zones, the intensity at point P will be maximum.

20.4.2 Diffraction by a Circular Disc

We next consider the diffraction pattern produced by an opaque disc of radius a (see Fig. 20.3). Once again we assume that the observation point lies on the axis of the disc. Equation (21) tells us that to calculate the field we have to carry out an integration over the open region of the aperture. Obviously, if $u_1(P)$ and $u_2(P)$, respectively, represent the fields at point P due to a circular aperture and an opaque disc (of the same radius), then

$$u_1(P) + u_2(P) = u_0(P) \qquad \textbf{(29)}$$

where $u_0(P)$ represents the field in the absence of any aperture; Eq. (29) is known as Babinet's principle. Thus,

$$u_2(P) = u_0(P) - u_1(P)$$
$$= u_0(P) - u_0(P)(1 - e^{ip\pi})$$

or

$$u_2(P) = u_0(P)\, e^{ip\pi} \qquad \textbf{(30)}$$

where for $u_1(P)$ we have used Eq. (26). Thus the intensity at point P on the axis of a circular disc is

$$I_2(P) = |u_2(P)|^2 = I_0(P) \qquad \textbf{(31)}$$

which gives us the remarkable result that the intensity at a point on the axis of an opaque disc is equal to the intensity at the point in the absence of the disc! This is the Poisson spot discussed in Sec. 20.2.2.

20.5 GAUSSIAN BEAM PROPAGATION

When a laser oscillates in its fundamental transverse mode, the transverse amplitude distribution is Gaussian (see Sec.26.5). Also, the output of a single mode fiber is very nearly Gaussian. Therefore, the study of the diffraction of a Gaussian beam is of great importance. We assume a Gaussian beam propagating along the z direction whose amplitude distribution on the plane $z = 0$ is given by

$$A(\xi,\eta) = a \exp\left(-\frac{\xi^2+\eta^2}{w_0^2}\right) \tag{32}$$

implying that the phase front is plane at $z = 0$. From Eq. (32) it follows that at a distance w_0 from the z axis, the amplitude falls by a factor $1/e$ (i.e., the intensity reduces by a factor $1/e^2$). This quantity w_0 is called the *spot size* of the beam. Substituting Eq. (32) into Eq. (23) and carrying out the integration, we obtain (see App. D)

$$u(x,y,z) \approx \frac{a}{1+i\gamma}\exp\left[-\frac{x^2+y^2}{w^2(z)}\right]e^{i\Phi} \tag{33}$$

where

$$\gamma = \frac{\lambda z}{\pi w_0^2} \tag{34}$$

$$w(z) = w_0(1+\gamma^2)^{1/2} = w_0\left(1+\frac{\lambda^2 z^2}{\pi^2 w_0^4}\right)^{1/2} \tag{35}$$

$$\Phi = kz + \frac{k}{2R(z)}(x^2+y^2) \tag{36}$$

$$R(z) \equiv z\left(1+\frac{1}{\gamma^2}\right) = z\left(1+\frac{\pi^2 w_0^4}{\lambda^2 z^2}\right) \tag{37}$$

Thus the intensity distribution is given by

$$I(x,y,z) = \frac{I_0}{1+\gamma^2}\exp\left[-\frac{2(x^2+y^2)}{w^2(z)}\right] \tag{38}$$

which shows that the transverse intensity distribution remains Gaussian with the beam width increasing with z which essentially implies diffraction divergence. As can be seen from Eq. (35), for small values of z, the width increases quadratically with z, but for large values of $z >> w_0^2/\lambda$, we obtain

$$w(z) \approx w_0\frac{\lambda z}{\pi w_0^2} = \frac{\lambda z}{\pi w_0} \tag{39}$$

which shows that the width increases linearly with z. We define the diffraction angle as

$$\tan\theta = \frac{w(z)}{z} \approx \frac{\lambda}{\pi w_0} \tag{40}$$

showing that the rate of increase in the width is proportional to the wavelength and inversely proportional to the initial width of the beam.To get some numerical values, we assume $\lambda = 0.5\ \mu$m. Then for $w_0 = 1$ mm

$$2\theta \approx 0.018^\circ \quad \text{and} \quad w \approx 1.88\ \text{mm} \quad \text{at } z = 10\ \text{m} \tag{41}$$

Similarly, for $w_0 = 0.25$ mm,

$$2\theta \approx 0.073^\circ \quad \text{and} \quad w \approx 6.35\ \text{mm} \quad \text{at } z = 10\ \text{m} \tag{42}$$

(see Fig. 20.9). Notice that θ increases with a decrease in w_0 (the smaller the size of the aperture, the greater the diffraction). Further, for a given value of w_0, the diffraction effects decrease with λ. In Fig. 20.10 we have shown the decrease in diffraction divergence for $w_0 = 0.25$ mm as the wavelength is decreased from 5000 to 500 Å; indeed as $\lambda \to 0$, $\theta \to 0$ and there is no diffraction which is the geometric optics limit. From Eq. (38) one can readily show that

$$\int_{-\infty}^{+\infty}\int I(x,y,z)\,dx\,dy = \frac{\pi w_0^2}{2} I_0 \tag{43}$$

which is independent of z. This is to be expected, as the total energy crossing the entire xy plane will not change with z.

Now, for a spherical wave *diverging* from the origin, the field distribution is given by

$$u \sim \frac{1}{r}e^{ikr} \tag{44}$$

Now, on the plane $z = R$ (see Fig. 20.11)

$$r = (x^2+y^2+R^2)^{1/2} \tag{45}$$

Thus

$$r = R\left(1+\frac{x^2+y^2}{R^2}\right)^{1/2}$$

$$\approx R + \frac{x^2+y^2}{2R} \tag{46}$$

where we have assumed $|x|, |y| << R$. Thus on the plane $z = R$, the phase distribution (corresponding to a spherical wave of radius R) is given by

$$e^{ikr} \approx e^{ikR}e^{\frac{ik}{2R}(x^2+y^2)} \tag{47}$$

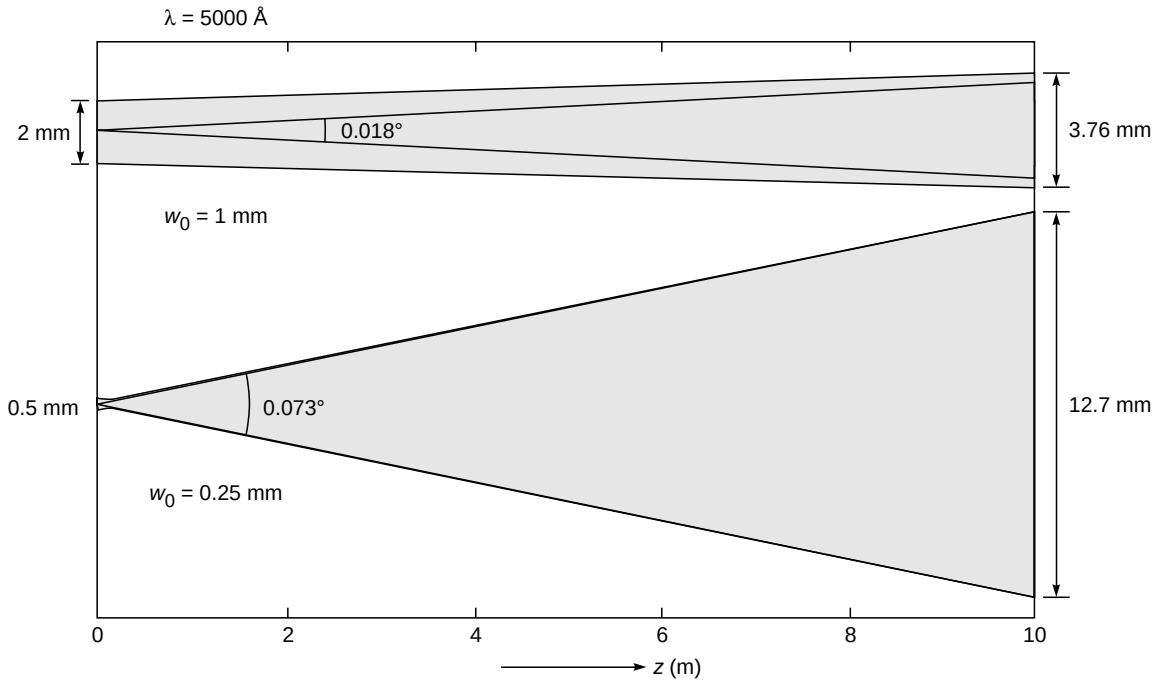


Fig. 20.9 Diffraction divergence of a Gaussian beam whose phase front is plane at $z = 0$. The figure shows the increase in the diffraction divergence as the initial spot size is decreased from 1 to 0.25 mm; the wavelength is assumed to be 5000 Å.

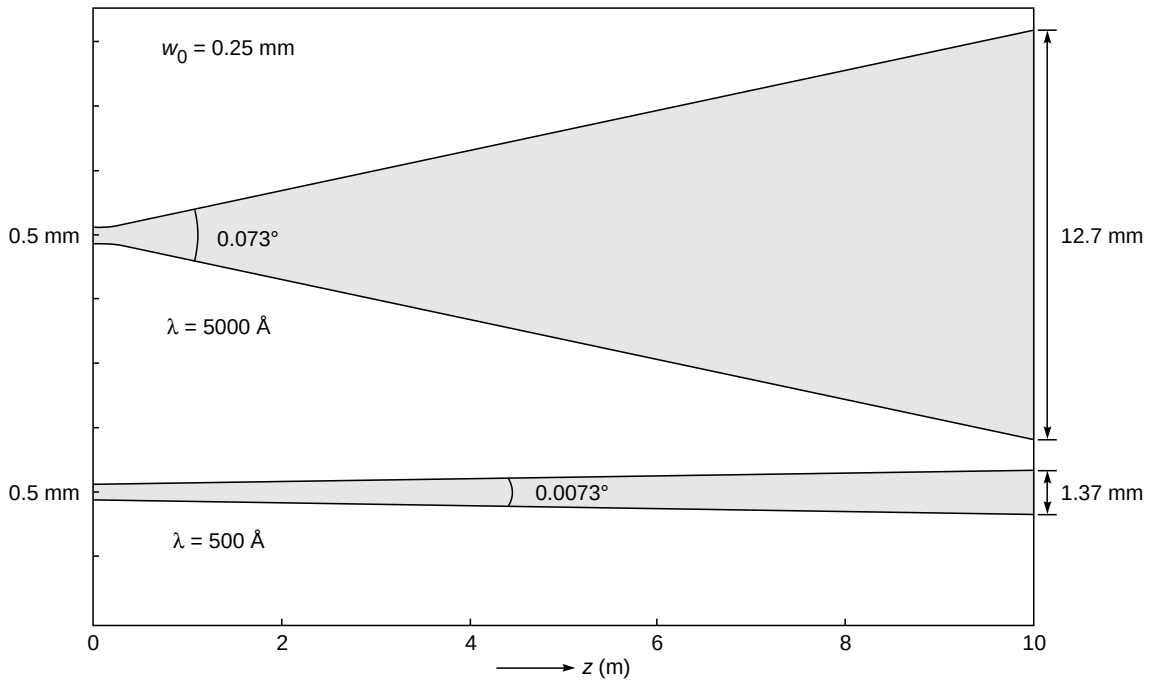


Fig. 20.10 Diffraction divergence of a Gaussian beam whose phase front is plane at $z = 0$. The figure shows the decrease in divergence as the wavelength is decreased from 5000 to 500 Å; the initial spot size w_0 is assumed to be 0.25 mm.

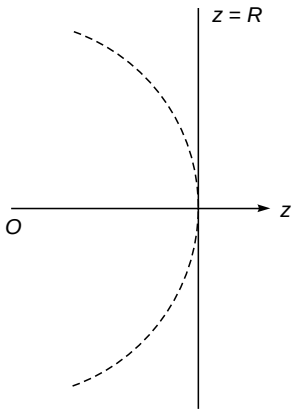


Fig. 20.11 A spherical wave diverging from point *O*. The dashed curve represents a section of the spherical wave front at a distance *R* from the source.

From the above equation it follows that a phase variation of the type

$$\exp\left[i\frac{k}{2R}(x^2 + y^2) \right] \qquad \textbf{(48)}$$

(on the *xy* plane) represents a *diverging* spherical wave of radius *R*. If we compare the above expression with Eqs. (42) and (43), we obtain the following approximate expression for the radius of curvature of the phase front:

$$R(z) \approx z\left(1 + \frac{\pi w_0^4}{\lambda^2 z^2}\right) \qquad \textbf{(49)}$$

which is shown in Fig. 20.11. Thus as the beam propagates, the phase front which was plane at $z = 0$ becomes curved. In Fig. 20.12 we have shown a Gaussian beam resonating between two identical spherical mirrors of radius *R*; the plane $z = 0$, where the phase front is plane and the beam has the minimum spot size, is referred to as the *waist* of the Gaussian beam. For the beam to resonate, the phase front must have a radius of curvature equal to *R* on the mirrors. For this to happen, we must have

$$R \approx \frac{d}{2}\left(1 + \frac{4\pi w_0^4}{\lambda^2 d^2}\right) \qquad \textbf{(50)}$$

where *d* is the distance between the two mirrors. We will use the above analysis in Sec. 26.5 and discuss the modes of a laser.

Note that although in the derivation of Eq. (33) we have assumed *z* to be large, Eq. (33) does give the correct field distribution even at $z = 0$.

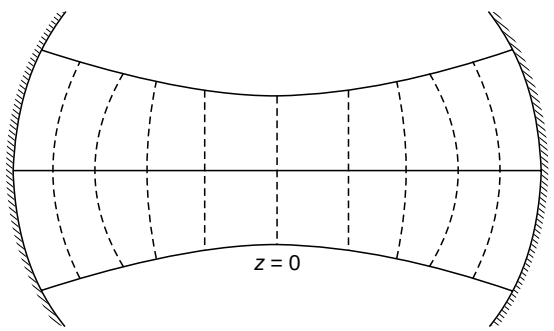
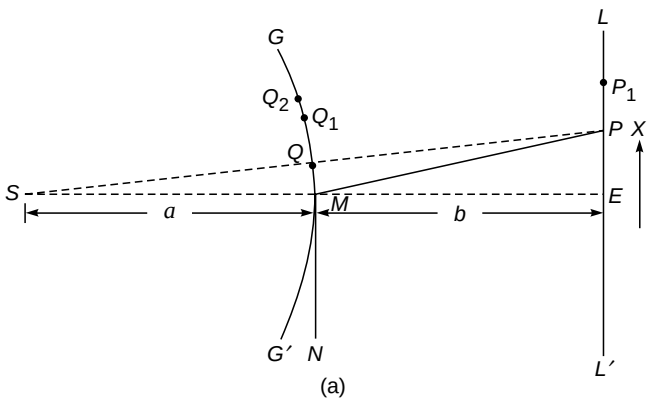


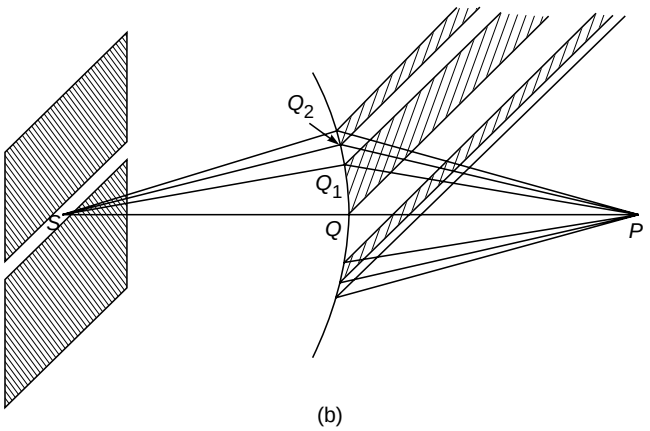
Fig. 20.12 Diffraction divergence of a Gaussian beam whose phase front is plane at $z = 0$. The dashed curves represent the phase fronts; see also Fig. 40 in the insert at the back of the book.

20.6 DIFFRACTION BY A STRAIGHT EDGE

Let us consider a straightedge *MN* placed perpendicular to the plane of the paper and parallel to a long, narrow slit *S* (see Fig. 20.13). We wish to calculate the intensity variation on screen *LL′*. From the geometry of the arrangement, it is



(a)



(b)

Fig. 20.13 (a) Diffraction at a straightedge. (b) Half-period strips of a cylindrical wave front.

obvious that on the screen there will be no intensity variation along the direction parallel to the length of the edge. Thus, the fringes (wherever they occur) will be straight lines parallel to the edge. We will first give a very approximate theory based on Fresnel half-period zones; this will be followed by a more rigorous analysis.

20.6.1 Analysis Using Half-Period Zones

In this section we will give a very approximate theory based on Fresnel half-period zones. The wave front emanating from the slit is cylindrical, and to find the amplitude at an arbitrary point P (on the screen), we draw half-period strips in the following manner: Let $GQMG'$ represent a section of the wave front, and point Q lies on the line joining S and P. Points Q_1 and Q_2 on the wave front are such that

$$\begin{aligned} SQ_1 + Q_1P - SQP &= \frac{\lambda}{2} \\ SQ_2 + Q_2P - SQP &= \frac{2\lambda}{2} \\ &\vdots \end{aligned} \qquad \textbf{(51)}$$

The half-period strips will be on the surface of the cylindrical wave front as shown in Fig. 20.13(b). However, unlike the Fresnel half-period zones, the areas of the half-period strips will not be equal, and thus the analysis becomes quite difficult. Even then one can draw the following conclusions:

1. Corresponding to the edge of the geometrical shadow [which is shown as E in Fig. 20.13(a)], one-half of the wave front is obstructed by the edge. Hence the amplitude will be given by

   $$u(E) = \tfrac{1}{2}u_0 \qquad \textbf{(52)}$$

   where u_0 represents the amplitude that would be produced by the unobstructed wave front (i.e., in the absence of the edge). Thus the intensity will be given by

   $$I(E) = \tfrac{1}{4}I_0 \qquad \textbf{(53)}$$

2. Let us next assume that point P satisfies the following relation:

   $$SM + MP - SQP = \frac{\lambda}{2} \qquad \textbf{(54)}$$

   Thus only the first half-period strip of the lower part of the wave front contributes, and the resultant amplitude is approximately

   $$\frac{u_1}{2} + \frac{u_1}{4} = \frac{3u_1}{4} = \frac{3}{2}\left(\frac{u_1}{2}\right) = \frac{3}{2}u_0 \qquad \textbf{(55)}$$

   where $u_1/2$ is the amplitude produced by the first half-period strip in the lower portion and $u_1/4$ is the resultant amplitude produced by the upper half of the wave front [see Eq. (12)]. The intensity is $\frac{9}{4}I_0$. For a point P_1 such that

   $$SM + MP_1 - SP_1 = \lambda \qquad \textbf{(56)}$$

   we will have a minimum, and the resultant amplitude will be

   $$\left(\frac{u_1}{2} - \frac{u_2}{2}\right) + \frac{u_1}{4} \qquad \textbf{(57)}$$

   In general, an arbitrary point P will correspond to maximum intensity if

   $$SM + MP - SP = (2n+1)\frac{\lambda}{2} \qquad n = 0, 1, 2, \ldots \qquad \textbf{(58)}$$

   and minimum intensity if

   $$SM + MP - SP = 2n\frac{\lambda}{2} \qquad n = 1, 2, \ldots \qquad \textbf{(59)}$$

   Now,

   $$MP = (b^2 + x^2)^{1/2} \approx b\left(1 + \frac{1}{2}\frac{x^2}{b^2}\right) = b + \frac{x^2}{2b}$$

   $$SP = [(a+b)^2 + x^2]^{1/2} \approx a + b + \frac{x^2}{2(a+b)}$$

   Hence,

   $$\begin{aligned} SM + MP - SP &\approx a + b + \frac{x^2}{2b} - (a+b) - \frac{x^2}{2(a+b)} \\ &\approx \frac{a}{2(a+b)b}x^2 \end{aligned}$$

   Thus, when

   $$x \cong \left[(2n+1)\frac{b(a+b)}{a}\lambda\right]^{1/2} \qquad n = 1, 2, \ldots \qquad \textbf{(60)}$$

   we will have a maximum. For example, for $a = b = 25$ cm and $\lambda = 5 \times 10^{-5}$ cm,

   $$\left[\frac{b(a+b)}{a}\lambda\right]^{1/2} = 5 \times 10^{-2} \text{ cm}$$

Thus the first maximum will occur at a distance of 0.05 cm from the edge of the shadow, and the second and third maxima will occur at distances of 0.0866 and 0.112 cm, respectively. The distance between two consecutive maxima will decrease as we go away from the edge of the geometrical shadow. Similarly, the positions of the minima are given by

$$x \simeq \left[2n\frac{b(a+b)}{a}\lambda\right]^{1/2} \qquad n = 1, 2, 3, \ldots \qquad \textbf{(61)}$$

and for the above parameters they will occur at distances of 0.07 cm, 0.10 cm, etc. By determining the positions of these maxima and minima, one can calculate the wavelength. The precise variation of the intensity is difficult to calculate from this analysis; a more rigorous theory will be given now.

20.6.2 More Rigorous Analysis of the Straightedge Diffraction Pattern

Before we discuss the straightedge diffraction pattern, we introduce the Fresnel integrals.

Fresnel Integrals: The Fresnel integrals are defined by the following equations:

$$C(\tau) = \int_0^{\tau} \cos\left(\frac{1}{2}\pi u^2\right) du \qquad \textbf{(62)}$$

and

$$S(\tau) = \int_0^{\tau} \sin\left(\frac{1}{2}\pi u^2\right) du \qquad \textbf{(63)}$$

Since the integrands are even functions of τ, the Fresnel integrals $C(\tau)$ and $S(\tau)$ are odd functions of τ:

$$C(-\tau) = -C(\tau) \qquad \text{and} \qquad S(-\tau) = -S(\tau) \qquad \textbf{(64)}$$

Further, since

$$\int_{-\infty}^{\infty} e^{-\alpha x^2} dx = \sqrt{\frac{\pi}{\alpha}} \qquad \textbf{(65)}$$

we have

$$\int_{-\infty}^{\infty} e^{i\pi u^2/2} du = \sqrt{\frac{\pi}{-i\pi/2}} = \sqrt{2}e^{i\pi/4} = 1 + i \qquad \textbf{(66)}$$

Now,

$$\int_{-\infty}^{+\infty} \exp\left(i\frac{\pi u^2}{2}\right) du$$

$$= 2\left[\int_0^{\infty} \cos\left(\frac{1}{2}\pi u^2\right) du + i\int_0^{\infty} \sin\left(\frac{1}{2}\pi u^2\right) du\right]$$

$$= 2[C(\infty) + i\,S(\infty)]$$

Thus, using Eq.(66), we get $C(\infty) = \frac{1}{2} = S(\infty)$.

To summarize, the Fresnel integrals have the following important properties:

$$C(\infty) = S(\infty) = \tfrac{1}{2} \qquad C(0) = S(0) = 0 \qquad \textbf{(67)}$$

$$C(-\tau) = -C(\tau) \qquad \text{and} \qquad S(-\tau) = -S(\tau) \qquad \textbf{(68)}$$

The values of the Fresnel integrals for typical values of τ are tabulated in Table 20.1.

Figure 20.14 gives a parametric representation of the Fresnel integrals and is known as Cornu's spiral. The horizontal and the vertical axes represent $C(\tau)$ and $S(\tau)$ respectively, and the numbers written on the spiral are the values of τ. For example, as can be seen from the figure, for $\tau = 1.0$, $C(\tau) \approx 0.77989$ and $S(\tau) \approx 0.43826$.

We now return to the calculation of the straightedge diffraction pattern which we had qualitatively discussed in Sec. 20.6.1. In this section we will make a more rigorous

Table 20.1 Table of Fresnel Integrals*

$$C(\tau) = \int_0^{\tau} \cos\left(\frac{\pi}{2}v^2\right) dv \qquad S(\tau) = \int_0^{\tau} \sin\left(\frac{\pi}{2}v^2\right) dv$$

τ	$C(\tau)$	$S(\tau)$	τ	$C(\tau)$	$S(\tau)$
0.0	0.00000	0.00000	2.6	0.38894	0.54999
0.2	0.19992	0.00419	2.8	0.46749	0.39153
0.4	0.39748	0.03336	3.0	0.60572	0.49631
0.6	0.58110	0.11054	3.2	0.46632	0.59335
0.8	0.72284	0.24934	3.4	0.43849	0.42965
1.0	0.77989	0.43826	3.6	0.58795	0.49231
1.2	0.71544	0.62340	3.8	0.44809	0.56562
1.4	0.54310	0.71353	4.0	0.49843	0.42052
1.6	0.36546	0.63889	4.2	0.54172	0.56320
1.8	0.33363	0.45094	4.4	0.43833	0.46227
2.0	0.48825	0.34342	4.6	0.56724	0.51619
2.2	0.63629	0.45570	4.8	0.43380	0.49675
2.4	0.55496	0.61969	5.0	0.56363	0.49919
			∞	0.5	0.5

*Adapted from Ref. 5; a more detailed table (with greater accuracy) has been given there.

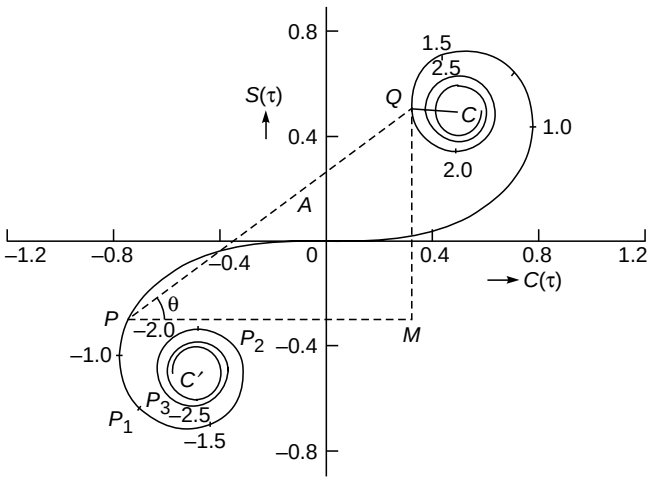


Fig. 20.14 Cornu's spiral which is a parametric plot of $C(\tau)$ and $S(\tau)$.

analysis of the diffraction of a plane wave incident normally on a straight edge (see Fig. 20.15). Once again, there will be no variation of intensity along the x axis and therefore, without any loss of generality, we may assume the coordinates of an arbitrary point P (on the screen) to be $(0, y)$, where the origin has been assumed to be on the edge of the geometrical shadow. If the x and y coordinates of an arbitrary point M on the plane of the straightedge are denoted by ξ and η, then

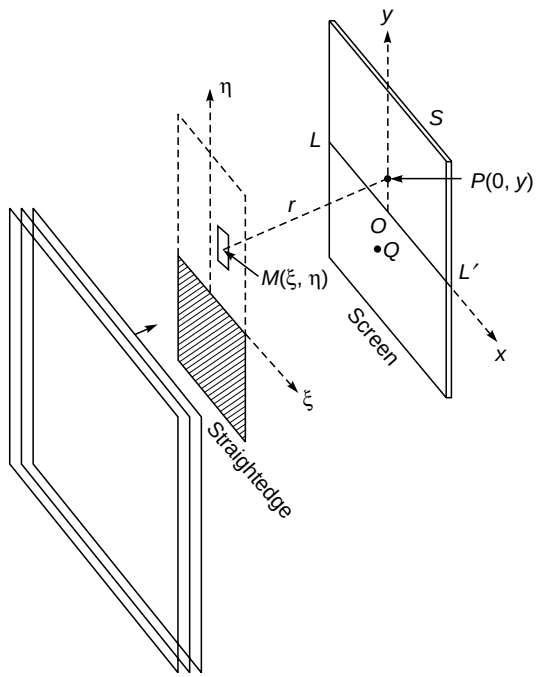


Fig. 20.15 Diffraction of a plane wave incident normally on a straightedge.

$$r = MP = [\xi^2 + (\eta - y)^2 + d^2]^{1/2}$$
$$= d\left[1 + \frac{\xi^2 + (\eta - y)^2}{d^2}\right]^{1/2}$$
$$\approx d + \frac{\xi^2 + (\eta - y)^2}{2d} \qquad \textbf{(69)}$$

where d is the distance between the straightedge and the screen. On substituting the expression for r from Eq. (69) into Eq. (21), we obtain

$$u(P) \approx -\frac{i}{\lambda}\frac{A}{d}\int_{-\infty}^{\infty} d\xi \int_{0}^{\infty} d\eta \exp\left\{ik\left[d + \frac{\xi^2 + (\eta - y)^2}{2d}\right]\right\} \qquad \textbf{(70)}$$

where, in the denominator of the integrand, we have replaced r by its minimum value[4] d. To express the above expression in terms of the Fresnel integrals, we introduce two dimensionless variables u and v such that

$$\frac{1}{2}\pi u^2 = \frac{k}{2d}\xi^2 = \frac{\pi}{\lambda d}\xi^2$$
$$\frac{1}{2}\pi v^2 = \frac{k}{2d}(\eta - y)^2 = \frac{\pi}{\lambda d}(\eta - y)^2$$

Thus we may assume u and v to be *defined* by the following equations:

$$u = \sqrt{\frac{2}{\lambda d}}\,\xi$$
$$v = \sqrt{\frac{2}{\lambda d}}\,(\eta - y) \qquad \textbf{(71)}$$

With these substitutions, Eq. (70) becomes

$$u(P) = -\frac{i}{2}u_0 \int_{-\infty}^{+\infty} \exp\left(\frac{i\pi u^2}{2}\right) du \int_{v_0}^{\infty} \exp\left(\frac{i\pi v^2}{2}\right) dv \qquad \textbf{(72)}$$

where

$$v_0 = -\sqrt{\frac{2}{\lambda d}}\,y \qquad \textbf{(73)}$$

and

$$u_0 = A\,e^{ikd}$$

represents the field at point P in the absence of the straightedge. To calculate the intensity distribution, we use the

[4] This is justified because in carrying out the integration, only a small region around the point $r = d$ contributes; the contribution due to far-off points is small because of the rapid oscillations of the exponential term in the integrand (see also footnote on page 290).

Fresnel integrals; thus

$$\int_{-\infty}^{+\infty} \exp\left(i\frac{\pi u^2}{2}\right) du = 2\,[C(\infty) + S(\infty)]$$
$$= 1 + i \qquad \textbf{(74)}$$

Further,

$$\int_{v_0}^{\infty} \exp\left(\frac{i\pi v^2}{2}\right) dv = \left[\int_0^{\infty} \cos\left(\frac{\pi}{2}v^2\right) dv - \int_0^{v_0} \cos\left(\frac{\pi}{2}v^2\right) dv\right]$$
$$+ i\left[\int_0^{\infty} \sin\left(\frac{\pi}{2}v^2\right) dv - \int_0^{v_0} \sin\left(\frac{\pi}{2}v^2\right) dv\right]$$
$$= \left[\frac{1}{2} - C(v_0)\right] + i\left[\frac{1}{2} - S(v_0)\right] \qquad \textbf{(75)}$$

Substituting in Eq. (72), we obtain

$$u(P) = -\frac{i}{2}u_0\,(1+i)\left\{\left[\frac{1}{2} - C(v_0)\right] + i\left[\frac{1}{2} - S(v_0)\right]\right\}$$
$$= \frac{1-i}{2}u_0\left\{\left[\frac{1}{2} - C(v_0)\right] + i\left[\frac{1}{2} - S(v_0)\right]\right\} \qquad \textbf{(76)}$$

Note that a large value of y corresponds to a point which is very far above the edge of the geometrical shadow. For such a point v_0 tends to $-\infty$ [see Eq. (73)] and we obtain

$$u(P) = \frac{1-i}{2}u_0\left[\left(\frac{1}{2}+\frac{1}{2}\right) + i\left(\frac{1}{2}+\frac{1}{2}\right)\right]$$
$$= u_0 \qquad \textbf{(77)}$$

Thus, as expected, the amplitude at such a point is the same as that in the absence of the edge. This also justifies the value of the constant given by Eq. (21). On the other hand, when point P is deep inside the geometrical shadow (i.e., when $y \to -\infty$ and hence $v_0 \to \infty$), we obtain

$$C(v_0) = S(v_0) \to \tfrac{1}{2}$$

giving $$u(P) \to 0$$

as it should indeed be. The intensity distribution corresponding to Eq. (76) is given by

$$I(P) = \frac{1}{2}I_0\left\{\left[\frac{1}{2} - C(v_0)\right]^2 + \left[\frac{1}{2} - S(v_0)\right]^2\right\} \qquad \textbf{(78)}$$

If point P is such that it lies on the edge of the geometrical shadow [i.e., on the line LL' (see Fig. 20.15], then $y = 0$ and hence $v_0 = 0$; thus

$$I(P) = \frac{1}{2}I_0\left(\frac{1}{4}+\frac{1}{4}\right) = \frac{1}{4}I_0 \qquad \textbf{(79)}$$

where we have used the fact that $C(0) = S(0) = 0$. Thus the intensity on the edge of the geometrical shadow is one-fourth of the intensity that it would have been in the absence of the edge [see also Eq. (53)]. To determine the field at an arbitrary point P, we may use Table 20.1 to calculate the RHS of Eq. (78). The intensity variation is plotted in Fig. 20.16 from which one can make the following observations:

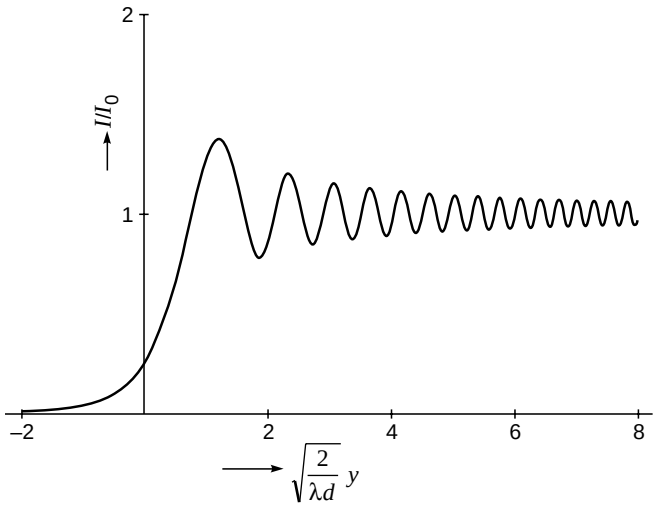


Fig. 20.16 The intensity variation corresponding to the straightedge diffraction pattern.

1. Figure 20.16 represents a universal curve; i.e., for given values of λ and d, one simply has to calculate v_0 as the observation point moves along the y axis. For example, the first three maxima occur at

$$v_0 \approx -1.22 \quad \text{with} \quad I \approx 1.37 I_0$$
$$v_0 \approx -2.34 \quad \text{with} \quad I \approx 1.20 I_0 \qquad \text{maxima}$$
$$v_0 \approx -3.08 \quad \text{with} \quad I \approx 1.15 I_0$$

   Similarly, the first three minima occur at

$$v_0 \approx -1.87 \quad \text{with} \quad I \approx 0.778 I_0$$
$$v_0 \approx -2.74 \quad \text{with} \quad I \approx 0.843 I_0 \qquad \text{minima}$$
$$v_0 \approx -3.39 \quad \text{with} \quad I \approx 0.872 I_0$$

   Thus, as y increases, the intensity modulation decreases (see Fig. 20.16).

2. For a given experimental setup, the determination of the positions of maxima and minima is quite straightforward. For example, for $\lambda = 6 \times 10^{-5}$ cm and $d = 120$ cm,

$$y = -\sqrt{\frac{\lambda d}{2}}\,v_0 = -0.06 v_0 \qquad \text{cm}$$

   Thus the first three maxima will occur at

$$y \approx 0.732,\ 1.404, \text{ and } 1.848 \text{ mm}$$

respectively. Similarly, the first three minima will occur at

$$y \approx 1.122,\ 1.644,\ \text{and}\ 2.034\ \text{mm}$$

respectively. These results may be compared with those obtained in Sec. 20.6.1.

3. As we go inside the geometrical shadow, the intensity monotonically decreases to zero.
4. One could have also studied the intensity variation directly from Cornu's spiral (see Fig. 20.14). This is due to the fact that associated with Cornu's spiral, we have the following interesting property. Let us write

$$[C(\tau_2) - C(\tau_1)] + i[S(\tau_2) - S(\tau_1)] \equiv Ae^{i\theta} \qquad \textbf{(80)}$$

Thus

$$C(\tau_2) - C(\tau_1) = A\cos\theta$$

and

$$S(\tau_2) - S(\tau_1) = A\sin\theta$$

Let points P and Q on Cornu's spiral (see Fig. 20.14) correspond to $\tau = \tau_1$ and $\tau = \tau_2$, respectively. It is obvious that

$$PM = C(\tau_2) - C(\tau_1) = A\cos\theta$$

and

$$QM = S(\tau_2) - S(\tau_1) = A\sin\theta$$

Thus the length of the line joining points P and Q is A, and the angle that the line makes with the abscissa is θ. To use Cornu's spiral, we rewrite Eq. (76):

$$u = \frac{1-i}{2}u_0\left\{\left[\frac{1}{2} - C(v_0)\right] + i\left[\frac{1}{2} - S(v_0)\right]\right\}$$

Let us first consider a point of observation Q in the geometrical shadow region. Consequently v_0 will be positive. Let point Q on the spiral (see Fig. 20.14) correspond to $\tau = v_0$. Since point C in the curve corresponds to $\tau = \infty$, we have

$$\frac{1}{2} - C(v_0) + i\left[\frac{1}{2} - S(v_0)\right] = QCe^{i\psi}$$

where ψ is the angle that QC makes with the abscissa [see Eq. (80)]. Thus,

$$u(Q) = \frac{1-i}{2}QCe^{i\psi}u_0$$

or

$$I(Q) = \frac{1}{2}(QC)^2 I_0 \qquad \textbf{(81)}$$

We can easily see that as the point of observation moves into the shadow region, the value of v_0 increases. Thus point Q keeps on moving on the spiral toward point C, and the length QC decreases uniformly. Hence in the shadow region the intensity uniformly decreases to zero (see Figs. 20.16 and 20.17).

As we move away from the edge of the geometrical shadow to the illuminated region, the value of v_0 becomes negative and the corresponding point P (on Cornu's spiral) lies in the third quadrant as shown in Fig. 20.14. The intensity is again given by

$$I(P) = \frac{1}{2}(PC)^2 I_0$$

As the value of v_0 becomes more and more negative, the length PC keeps on increasing until point P reaches P_1 which corresponds to $v_0 \approx -1.22$. The intensity at this point is maximum, and the length $P_1C \approx 1.67$. Thus, the corresponding intensity is

$$I(P_1) \approx \frac{1}{2}(1.67)^2 I_0 \approx 1.37 I_0 \qquad \textbf{(82)}$$

As the value of v_0 becomes further negative, the length PC starts decreasing until it reaches point P_2. Thus, the intensity keeps on oscillating with decreasing amplitude about I_0 as we move more and more into the illuminated region (see Figs. 20.16 and 20.17).

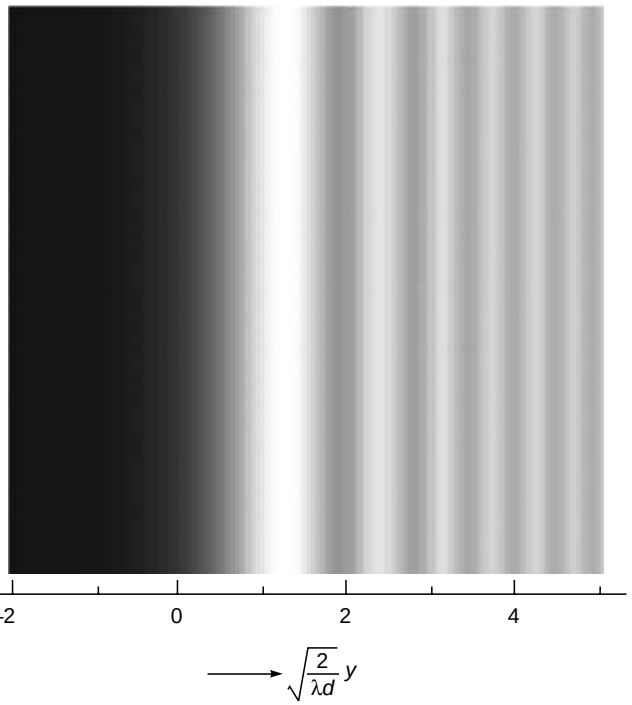


Fig. 20.17 Computer-generated intensity distribution corresponding to the straightedge diffraction pattern.

20.7 DIFFRACTION OF A PLANE WAVE BY A LONG NARROW SLIT AND TRANSITION TO THE FRAUNHOFER REGION

We next consider a plane wave incident normally on a long narrow slit (of width b) as shown in Fig. 20.18. We wish to calculate the intensity distribution at an arbitrary point P on screen SS'. Lines LL' and MM' represent the edges of the geometrical shadow. Once again, there will be no variation of the intensity along the x axis, and we may (without any loss of generality) assume the coordinates of point P to be $(0,y)$. The field at point P will again be given by Eq. (70) except that the limits of the η integral will be $-b/2$ and $+b/2$ (we are assuming the origin to be at the center of the slit).

$$u(P) = -\frac{iA}{\lambda d}\int_{-\infty}^{\infty} d\xi \int_{-b/2}^{+b/2} d\eta \exp\left\{ik\left[d + \frac{\xi^2 + (\eta - y)^2}{2d}\right]\right\}$$

Carrying out manipulations similar to those in the previous section, we obtain

$$u(P) = -\frac{i}{2}u_0 \int_{-\infty}^{\infty} \exp\left(\frac{i\pi u^2}{2}\right)du \int_{-(v_2+v_1)}^{-(v_2-v_1)} \exp\left(\frac{i\pi v^2}{2}\right)dv$$

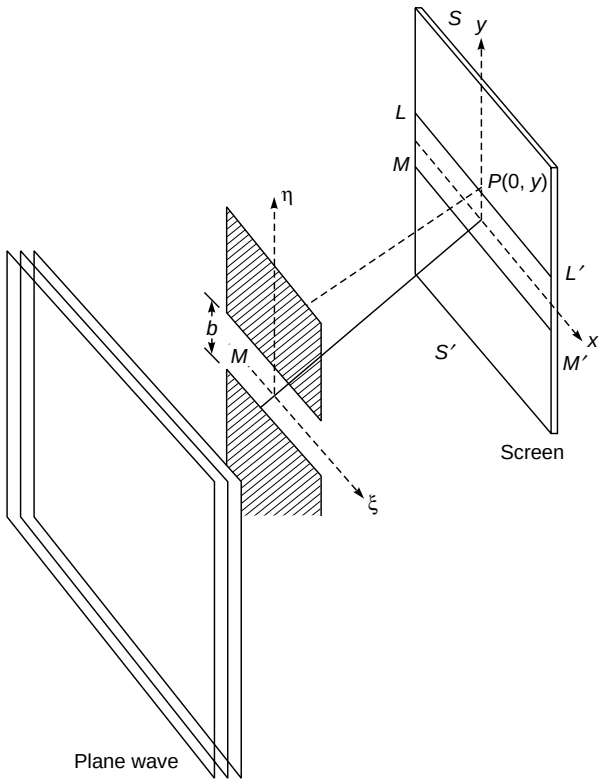


Fig. 20.18 Diffraction of a plane wave incident normally on a long narrow slit.

where

$$u = \sqrt{\frac{2}{\lambda d}}\xi, \qquad v = \sqrt{\frac{2}{\lambda d}}(\eta - y)$$

and

$$v_1 = \sqrt{\frac{2}{\lambda d}}\frac{b}{2} \qquad v_2 = \sqrt{\frac{2}{\lambda d}}\, y$$

Using Eq. (66), we obtain

$$u(P) = -\frac{i}{2}u_0(1+i) \times \left\{\int_0^{-(v_2-v_1)} \left[\cos\left(\frac{\pi}{2}v^2\right) + i\sin\left(\frac{\pi}{2}v^2\right)\right]dv - \int_0^{-(v_2+v_1)} \left[\cos\left(\frac{\pi}{2}v^2\right) + i\sin\left(\frac{\pi}{2}v^2\right)\right]dv\right\}$$

or

$$u(P) = \frac{1-i}{2}u_0\{[C(v_2+v_1) - C(v_2-v_1)] + i[S(v_2+v_1) - S(v_2-v_1)]\} \qquad \textbf{(83)}$$

where we have used Eq. (64). Thus the intensity distribution is

$$I(P) = \tfrac{1}{2}I_0\{[C(v_2+v_1) - C(v_2-v_1)]^2 + [S(v_2+v_1) - S(v_2-v_1)]^2\} \qquad \textbf{(84)}$$

For a given system λ, d, and b are known and determine v_1; e.g., for $\lambda = 5 \times 10^{-5}$ cm, $d = 100$ cm, and $b = 0.1$ cm, one obtains $v_1 = 2.0$; further, as y varies on the screen, the quantity v_2 also changes. In Figs. 20.19, 20.20, 20.21, and 20.22 we have plotted the intensity variation as a function of v_2 for

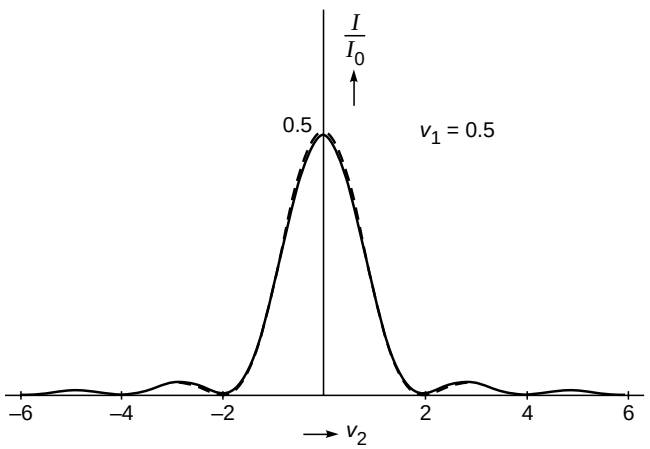


Fig. 20.19 The intensity distribution produced by diffraction of a plane wave by a long narrow slit corresponding to $v_1 = 0.5$. The dashed curves correspond to Eq. (86).

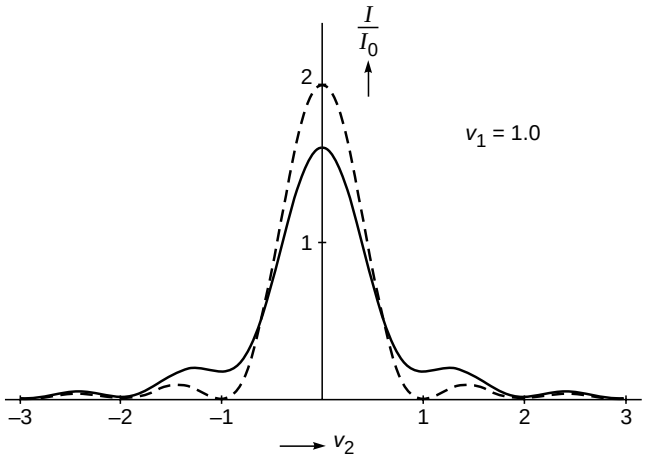


Fig. 20.20 The intensity distribution produced by diffraction of a plane wave by a long narrow slit corresponding to v_1 = 1.0. The dashed curves correspond to Eq. (86).

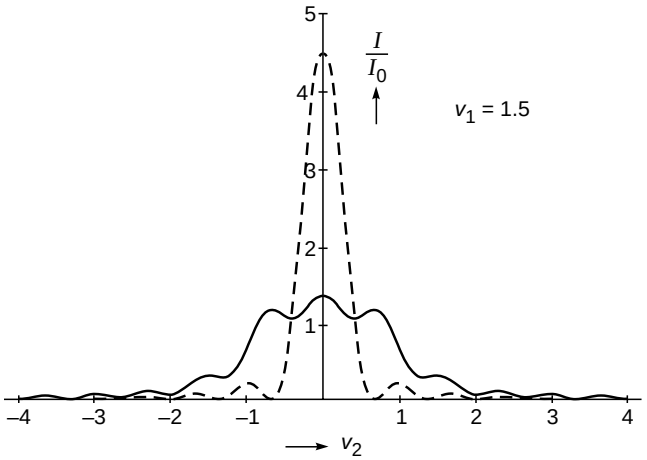


Fig. 20.21 The intensity distribution produced by diffraction of a plane wave by a long narrow slit corresponding to v_1 = 1.5. The dashed curves correspond to Eq. (86).

v_1 = 0.5, 1.0, 1.5, and 5.0, respectively. One can see that for a large value of v_1 (i.e., when the slit width is very large) the diffraction pattern is similar to that produced by two straightedges. This is indeed what we should have also expected. On the other hand, for small values of v_1 (i.e., when the observation screen is far away from the aperture), the diffraction pattern is essentially of the Fraunhofer type. To show this explicitly, we notice that

$$v_2 = \sqrt{\frac{2}{\lambda d}}\, y = \sqrt{\frac{2d}{\lambda}}\,\frac{y}{d} \approx \sqrt{\frac{2d}{\lambda}}\,\theta \qquad \textbf{(85)}$$

where θ represents the angle of diffraction (see Fig. 20.23). Clearly, in the Fraunhofer region since d is very large, the

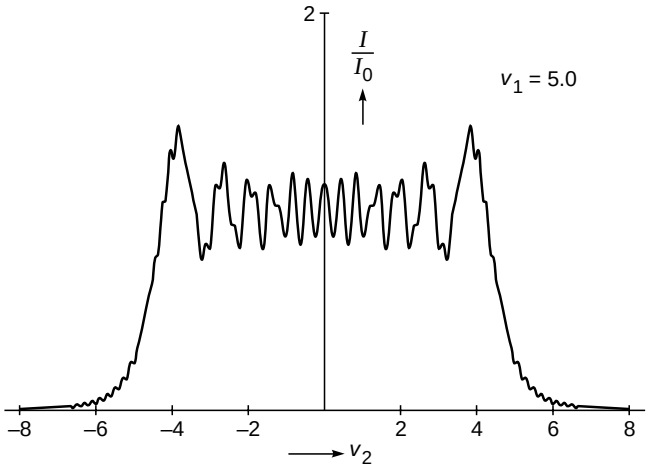


Fig. 20.22 The intensity distribution produced by diffraction of a plane wave by a long narrow slit corresponding to v_1 = 5.0.

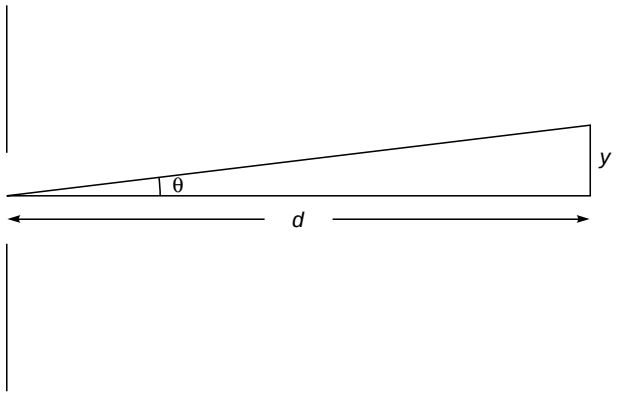


Fig. 20.23 In the Fraunhofer region, d is very large.

value of v_2 will also be very large, and thus we must look for expressions of the Fresnel integrals in the limit of $v \to \infty$. Now, we may write

$$C(v) = \int_0^v \cos\left(\frac{\pi}{2}v^2\right)dv$$

$$= \int_0^\infty \cos\left(\frac{\pi}{2}v^2\right)dv - \int_v^\infty \cos\left(\frac{\pi}{2}v^2\right)dv$$

$$= \frac{1}{2} - \int_v^\infty \frac{1}{\pi v}\cos\left(\frac{\pi}{2}v^2\right)\pi v\, dv$$

$$= \frac{1}{2} - \frac{1}{\pi v}\sin\left(\frac{\pi}{2}v^2\right)\Bigg|_v^\infty + \int_v^\infty \frac{1}{\pi v^2}\sin\left(\frac{\pi}{2}v^2\right)dv$$

$$\cong \frac{1}{2} + \frac{1}{\pi v}\sin\left(\frac{\pi}{2}v^2\right)$$

where we have neglected terms which would be of order $1/v^3$. Similarly,

$$S(v) = \frac{1}{2} - \frac{1}{\pi v}\cos\left(\frac{\pi}{2}v^2\right)$$

Since v_2 is large and v_1 is small, we have

$$C(v_2 + v_1) - C(v_2 - v_1) \approx \left\{\frac{1}{2} + \frac{1}{\pi v_2}\sin\left[\frac{\pi}{2}(v_2 + v_1)^2\right]\right\} - \left\{\frac{1}{2} + \frac{1}{\pi v_2}\sin\left[\frac{\pi}{2}(v_2 - v_1)^2\right]\right\}$$

$$\equiv \frac{2}{\pi v_2}\cos\left[\frac{\pi}{2}(v_2^2 + v_1^2)\right]\sin \pi v_1 v_2$$

Similarly,

$$S(v_2 + v_1) - S(v_2 - v_1) \approx \frac{2}{\pi v_2}\sin\left[\frac{\pi}{2}\left(v_2^2 + v_1^2\right)\right]\sin \pi v_1 v_2$$

Thus, in the Fraunhofer limit, Eq. (84) becomes

$$I(P) = \frac{1}{2} I_0 \left(\frac{4}{\pi^2 v_2^2}\sin^2 \pi v_1 v_2\right)$$

$$= I_{00}\frac{\sin^2\beta}{\beta^2} \qquad \textbf{(86)}$$

where

$$I_{00} = 2I_0 v_1^2$$

and

$$\beta = \pi v_1 v_2 = \frac{\pi}{\lambda d} by \approx \frac{\pi b}{\lambda}\theta \qquad \textbf{(87)}$$

and

$$\theta \approx \frac{y}{d} \qquad \textbf{(88)}$$

represents the diffraction angle. Equation (86) shows that the intensity distribution is indeed of the Fraunhofer type (see Sec. 18.2). In Figs. 20.19 to 20.22 the dashed curves correspond to Eq. (86), and one can see that the intensity distribution is almost of the Fraunhofer type for $v_1 \le 0.5$.

Summary

- The underlying principle in the theory of diffraction is the Huygens–Fresnel principle according to which *each point on a wave front is a source of secondary disturbance, and the secondary wavelets emanating from different points mutually interfere.*
- For a plane wave incident normally on a circular aperture of radius *a*, the intensity variation on an axial point *P* is given by

$$I = I_0 \sin^2\frac{p\pi}{2}$$

where

$$p \approx \frac{a^2}{\lambda d}$$

λ is the wavelength and *d* is the distance of point *P* from the center of the circular aperture. The quantity *p* is known as the Fresnel number of the aperture. When $p = 1, 3, 5, 7, \ldots$, we have maximum intensity and the circular aperture will contain (with respect to point *P*) an odd number of Fresnel half-period zones; and when $p = 2, 4, 6, 8, \ldots$, we have minimum intensity and the circular aperture will contain an even number of half-period zones.

- If instead of the circular aperture we have an opaque disc, then we always obtain a bright spot on the axis behind the disc; this is called the *Poisson spot*.
- For a Gaussian beam (whose phase front is plane at $z = 0$), the variation of the spot size is given by

$$w(z) \approx w_0\left(1 + \frac{\lambda^2 z^2}{\pi^2 w_0^4}\right)^{1/2}$$

where w_0 is the spot size at $z = 0$. For large values of *z*,

$$w(z) \approx \frac{\lambda z}{\pi w_0}$$

which shows that the width increases linearly with *z*. We define the diffraction angle as

$$\tan\theta = \frac{w(z)}{z} \approx \frac{\lambda}{\pi w_0}$$

showing that the rate of increase in width is proportional to the wavelength and inversely proportional to the initial width of the beam; this is characteristic of diffraction. The corresponding radius of curvature of the wave front is given by

$$R(z) \approx z\left(1 + \frac{\pi^2 w_0^4}{\lambda^2 z^2}\right)$$

- For a plane wave incident normally on a straightedge, the intensity variation on a screen (at a distance *d* from the straightedge) is given by

$$I = \frac{1}{2} I_0 \left\{\left[\frac{1}{2} - C(v_0)\right]^2 + \left[\frac{1}{2} - S(v_0)\right]^2\right\}$$

where I_0 is the intensity in the absence of the straightedge,

$$v_0 = -\sqrt{\frac{2}{\lambda d}}\, y$$

with *y* being the distance from the edge of the geometrical shadow, and

$$C(x) = \int_0^x \cos\left(\tfrac{1}{2}\pi u^2\right) du$$

and

$$S(x) = \int_0^x \sin\left(\tfrac{1}{2}\pi u^2\right) du$$

are known as Fresnel integrals. The intensity monotonically goes to zero as we go deep inside the geometrical shadow. As we move away from the edge of the geometrical shadow to the illuminated region, one obtains maxima at $v_0 \approx -1.22$ ($I \approx 1.37I_0$), -2.34 ($I \approx 1.20I_0$), -3.08 ($I \approx 1.15I_0$) ... and minima at $v_0 \approx -1.87$ ($I \approx 0.78I_0$), -2.74 ($I \approx 0.84I_0$), -3.39 ($I \approx 0.87I_0$), ...

- For a plane wave incident normally on a long narrow slit of width b, the intensity variation on a screen (at a distance d from the slit) is given by

$$I = \frac{1}{2} I_0 \{[C(v_2 + v_1) - C(v_2 - v_1)]^2 + [S(v_2 + v_1) - S(v_2 - v_1)]^2\}$$

where

$$v_1 = \sqrt{\frac{2}{\lambda d}}\frac{b}{2} \qquad v_2 = \sqrt{\frac{2}{\lambda d}}\, y$$

and y is the distance from the midpoint of the edges of the geometrical shadow. As v_1 becomes large, we obtain the intensity distribution corresponding to two straightedges, and for $v_1 \to 0$ we get the Fraunhofer diffraction pattern.

Problems

20.1 Consider a plane wave of wavelength 6×10^{-5} cm incident normally on a circular aperture of radius 0.01 cm. Calculate the positions of the brightest and the darkest points on the axis.

[**Ans:** $d \approx 1.67$ cm, 0.56 cm, 0.33 cm, ... (maxima); $d \approx 0.83$ cm, 0.42 cm, ... (minima)]

20.2 What would happen if the circular aperture in Prob. 20.1 were replaced by a circular disc of the same radius?

20.3 (a) A plane wave ($\lambda = 6 \times 10^{-5}$ cm) is incident normally on a circular aperture of radius a.

(i) Assume $a = 1$ mm. Calculate the values of z (on the axis) for which maximum intensity will occur. Plot the intensity as a function of z and interpret physically.

(ii) Assume $z = 50$ cm. Calculate the values of a for which minimum intensity will occur on the axial point. Plot the intensity variation as a function of a and interpret physically.

(b) Repeat the calculations for $\lambda = 5 \times 10^{-5}$ cm and discuss chromatic aberration of a zone plate.

[**Ans:** (a) (i) $z \approx 166.7$ cm, 55.6 cm, 33.3 cm, ... (maxima);
(ii) minimum intensity will occur when $a \approx 0.0775$ cm, 0.110 cm, 0.134 cm, ...]

20.4 Consider a circular aperture of diameter 2 mm illuminated by a plane wave. The most intense point on the axis is at a distance of 200 cm from the aperture. Calculate the wavelength.

[**Ans:** 5×10^{-5} cm]

20.5 If a zone plate has to have a principal focal length of 50 cm corresponding to $\lambda = 6 \times 10^{-5}$ cm, obtain an expression for the radii of different zones. What is its principal focal length for $\lambda = 5 \times 10^{-5}$ cm?

[**Ans:** $\sqrt{0.3n}$ mm, 60 cm]

20.6 In a zone-plate, the second, fourth, sixth, ... zones are blackened; what would happen if instead the first, third, fifth, etc., zones were blackened?

20.7 (a) A plane wave is incident normally on a straightedge (see Fig. 20.24). Show that the field at an arbitrary point P is given by

$$u(P) = \frac{1-i}{2} u_0 \left\{\left[\frac{1}{2} - C(v_0)\right] + i\left[\frac{1}{2} - S(v_0)\right]\right\}$$

where $$v_0 = -\sqrt{\frac{2}{\lambda d}}\, y$$

(b) Assume $\lambda_0 = 5000$ Å and $d = 100$ cm. Using Table 20.1, write approximately the values of I/I_0 at points O, P, ($y = 0.5$ mm), Q ($y = 1$ mm), and R ($y = -1$ mm), where O is at the edge of the geometrical shadow.

[**Ans:** (b) $I/I_0 \approx 1.26$, 0.24, 0.01]

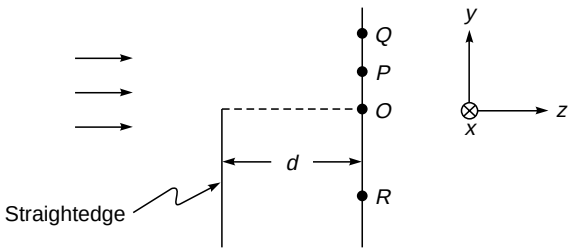


Fig. 20.24

20.8 Consider a straightedge being illuminated by a parallel beam of light with $\lambda = 6 \times 10^{-5}$ cm. Calculate the positions of the first two maxima and minima on a screen at a distance of 50 cm from the edge.

[**Ans:** the first two maxima occur at $y \approx 0.0473$ and 0.0906 cm. The first two minima occur at $y \approx 0.0724$ and 0.1061 cm.]

20.9 In a straightedge diffraction pattern, one observes that the most intense maximum occurs at a distance of 1 mm from the edge of the geometrical shadow. Calculate the wavelength of light, if the distance between the screen and the straightedge is 300 cm.

[**Ans:** ≈ 4480 Å]

20.10 In a straightedge diffraction pattern, if the wavelength of the light used is 6000 Å and if the distance between the screen and the straightedge is 100 cm, calculate the distance between the most intense maximum and the next maximum. Find approximately the distance in centimeters inside the geometrical shadow where $I/I_0 = 0.1$.

[**Ans:** $y \approx 0.027$ cm]

20.11 Consider a plane wave falling normally on a narrow slit of width 0.5 mm. If the wavelength of light is 6×10^{-5} cm, calculate the distance between the slit and the screen so that the value of v_1 is 0.5, 1.0, 1.5, and 5.0 (see Figs. 20.19 to 20.22). Discuss the transition to the Fraunhofer region.

20.12 Consider the Fresnel diffraction pattern produced by a plane wave incident normally on a slit of width b. Assume $\lambda = 5 \times 10^{-5}$ cm and $d = 100$ cm. Using Table 20.1, approximately calculate the intensity values (for $b = 0.1$ cm) at $y = 0, \pm 0.05$ cm, ± 0.1 cm. Repeat the analysis for $b = 5$ cm.

[**Ans:** At $y = 0$, $I/I_0 \approx 1.60$; at $y = \pm 0.05$ cm, $I/I_0 \approx 0.356$; at $y = \pm 0.01$ cm, $I/I_0 \approx 0.01685$].

20.13 In Sec. 20.9 we obtained the diffraction pattern of a circular aperture of radius a. Obtain the diffraction pattern of an annular aperture bounded by circles of radii a_1 and a_2 ($> a_1$).

[*Hint:* The integration limits of ρ in Eq. (103) must be a_1 and a_2.]

20.14 Consider a rectangular aperture of dimensions 0.2 mm × 0.3 mm. Obtain the positions of the first few maxima and minima in the Fraunhofer diffraction pattern along directions parallel to the length and breadth of the rectangle. Assume $\lambda = 5 \times 10^{-5}$ cm and that the diffraction pattern is produced at the focal plane of a lens of focal length 20 cm.

[**Ans:** Along the x axis, minima will occur at $x \approx 0.05$, 0.10, 0.15, . . . cm; along the y axis, minima will occur at $y \approx 0.033$, 0.067, 0.1, . . . cm]

20.15 The Fraunhofer diffraction pattern of a circular aperture (of radius 0.5 mm) is observed on the focal plane of a convex lens of focal length 20 cm. Calculate the radii of the first and the second dark rings. Assume $\lambda = 5.5 \times 10^{-5}$ cm.

[**Ans:** 0.13 mm, 0.25 mm]

20.16 In Prob. 20.15, calculate the area of the patch (on focal plane) which will contain 95% of the total energy.

[**Ans:** $\approx 5.55 \times 10^{-3}$ cm^2]

20.17 (a) The output of a He-Ne laser ($\lambda = 6328$ Å) can be assumed to be Gaussian with plane phase front. For $w_0 = 1$ mm and $w_0 = 0.2$ mm, calculate the beam diameter at $z = 20$ m.

(b) Repeat the calculation for $\lambda = 5000$ Å and interpret the results physically.

[**Ans:** (a) 0.83 cm and 4.0 cm]

20.18 A Gaussian beam is coming out of a laser. Assume $\lambda = 6000$ Å and that at $z = 0$, the beam width is 1 mm and the phase front is plane. After traversing 10 m through vacuum, what will be (a) the beam width and (b) the radius of curvature of the phase front?

[**Ans:** $2w \approx 0.77$ cm; $R(z) \approx 1017$ cm]

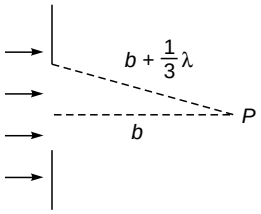


Fig. 20.25

20.19 A plane wave of intensity I_0 is incident normally on a circular aperture as shown in Fig. 20.25. What will be the intensity on the axial point P?

[*Hint:* You may use Eq. (25).]

20.20 Show that a phase variation of the type

$$\exp\left[ikz + \frac{ik(x^2 + y^2)}{2R(z)}\right]$$

represents a diverging spherical wave of radius R.

20.21 Consider a resonator consisting of a plane mirror and a concave mirror of radius of curvature R (see Fig. 20.26). Assume $\lambda = 1$ μm, $R = 100$ cm, and the distance between the two mirrors is 50 cm. Calculate the spot size of the Gaussian beam.

[**Ans:** $w_0 \approx 0.4$ mm]

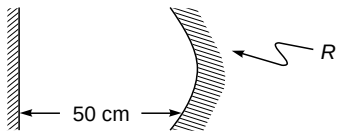


Fig. 20.26

20.22 The output of a semiconductor laser can be approximately described by a Gaussian function with two different widths along the transverse (w_T) and lateral (w_L) directions as

$$\psi(x, y) = A \exp\left(-\frac{x^2}{w_L^2} - \frac{y^2}{w_T^2}\right)$$

where x and y, respectively, axes parallel and perpendicular to the junction plane. Typically $w_T \approx 0.5$ μm and $w_L = 2$ μm. Discuss the far field of this beam (see Fig. 20.27).

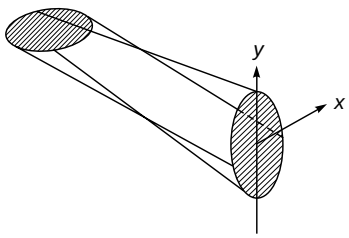


Fig. 20.27

REFERENCES AND SUGGESTED READINGS

1. R. Baierlein, *Newton to Einstein: The Trail of Light*, Cambridge University Press, 1992.
2. P. M. Rinard, "Large Scale Diffraction Patterns from Circular Objects," *American Journal of Physics*, Vol. 44, p. 70, 1976.
3. M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 2000.
4. A. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978.
5. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, Applied Mathematics Series*, Vol. 55, National Bureau of Standards, Washington, 1964.

Chapter Twenty-One

HOLOGRAPHY[1]

The electron microscope was to produce the interference figure between the object beam and the coherent background, that is to say the non-diffracted part of the illuminating beam. This interference pattern I called a *hologram*, from the Greek word *holos*—the whole, because it contained the whole information. The hologram was then reconstructed with light, in an optical system which corrected the aberrations of the electron optics.

—Dennis Gabor in his Nobel lecture[2], December 11, 1971

Important Milestones

1948	*Dennis Gabor discovered the principle of holography.*
1960	*The first successful operation of a laser device was achieved by Theodore Maiman.*
1962	*Off-axis technique of holography was* pioneered *by Leith and Upatnieks.*
1962	*Denisyuk suggested the idea of three-dimensional holograms based on thick photoemulsion layers. His holograms can be reconstructed in ordinary sunlight. These holograms are called Lippmann–Bragg holograms.*
1964	*Leith and Upatnieks pointed out that a multicolor image can be produced by a hologram recorded with three suitably chosen wavelengths.*
1969	*Benton invented "rainbow holography" for display of holograms in white light. This was a vital step to make holography suitable for display applications.*

21.1 INTRODUCTION

A photograph represents a two-dimensional recording of a three-dimensional scene. What is recorded is the intensity distribution that prevailed at the plane of the photograph when it was exposed. The light-sensitive medium is sensitive only to the intensity variations; hence while recording a photograph, the phase distribution which prevailed at the plane of the photograph is lost. Since only the intensity pattern has been recorded, the three-dimensional character (e.g., parallax) of the object scene is lost. Thus one cannot change the perspective of the image in the photograph by viewing it from a different angle, or one cannot refocus any unfocused part of the image in the photograph. Holography is a method invented by Dennis Gabor in 1948, in which one records not only the amplitude but also the phase of the light wave; this is done by using interferometric techniques. Because of this, the image, produced by the technique of holography has a true three-dimensional form. Thus, as with the object, one can change one's position and view a different perspective of the image, or one can focus at different distances. The capability to produce images as true as the object itself is responsible for the wide popularity gained by holography.

[1] A portion of this chapter is based on the unpublished lecture notes of Prof. K. Thyagarajan.

[2] Dennis Gabor received the 1971 Nobel Prize in Physics for discovering the principles of holography; the original paper of Gabor appeared in 1948 (see Ref. 1). Gabor's Nobel lecture entitled "Holography, 1948–1971" is nonmathematical and full of beautiful illustrations; it is reprinted in Ref. 2.

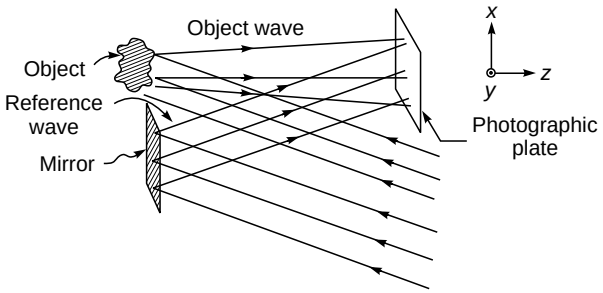


Fig. 21.1 Recording of a hologram.

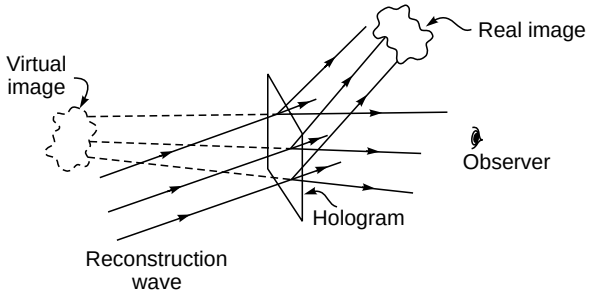


Fig. 21.2 Reconstruction process.

The basic technique in holography is the following: In the recording of the hologram, one superimposes on the object wave another wave called the reference wave, and the photographic plate is made to record the resulting interference pattern (see Fig. 21.1). The reference wave is usually a plane wave. This recorded interference pattern forms the hologram and (as will be shown) contains information about not only the amplitude but also the phase of the object wave. Unlike a photograph, a hologram has little resemblance to the object; in fact, information about the object is coded into the hologram. To view the image, we again illuminate the hologram with another wave, called the reconstruction wave (which in most cases is identical to the reference wave used during the formation of the hologram); this process is termed *reconstruction* (see Fig. 21.2). The reconstruction process leads, in general, to a virtual and a real image of the object scene. The virtual image has all the characteristics of the object, such as parallax. Thus one can move the position of the eye and look behind the objects, or one can focus at different distances. The real image can be photographed without the aid of lenses just by placing a light-sensitive medium at the position where the real image is formed. Figure 21.3(a), (b), and (c) represents the object, its hologram, and the reconstructed image, respectively.

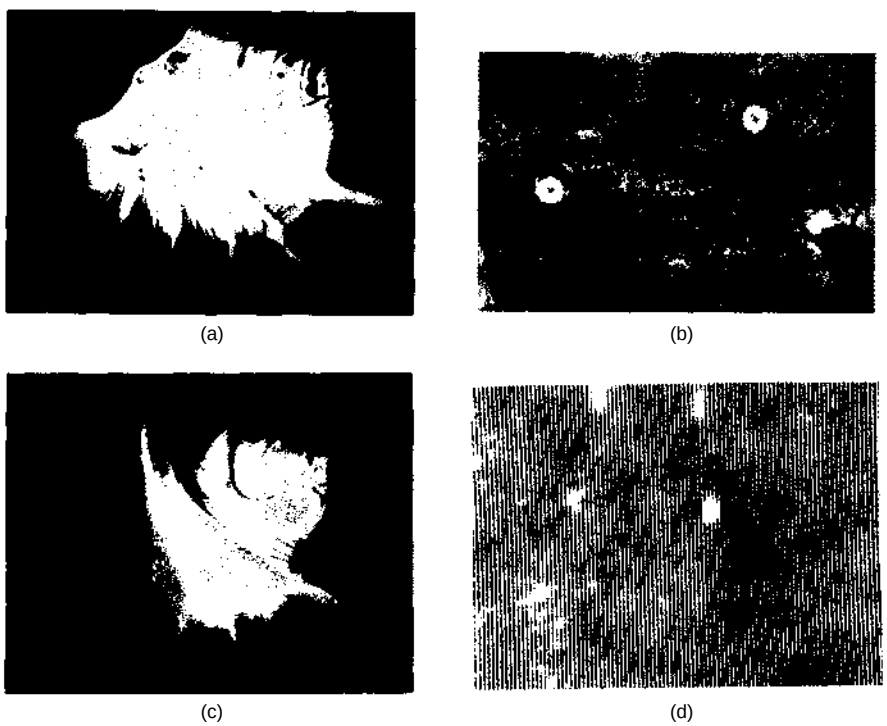


Fig. 21.3 (a) An ordinary photograph of an object. (b) The hologram of the object produced by a method similar to the one as shown in Fig. 21.1. (c) The reconstructed image as seen by an observer. (d) A magnified view of a small portion of the hologram shown in (b) [Photographs courtesy Professor R. S. Sirohi].

21.2 THEORY

If the object is a point scatterer, then the object wave is just $(A/r)\ \cos(kr - \omega t + \phi)$, where r represents the distance of the point of observation from the point scatterer and A represents a constant; $k = 2\pi/\lambda$. Any general object can be thought of as being made up of a large number of points, and the composite wave reflected by the object is the vectorial sum of these. The fundamental problem in holography is the recording of this object wave, in particular, the phase distribution associated with it.

Let us consider the recording process. Let

$$O(x,y) = a(x,y) \cos[\phi(x, y) - \omega t] \quad \textbf{(1)}$$

represents the object wave (which, as mentioned earlier, is due to the superposition of waves from point scatterers on the object) in the plane of the photographic plate which is assumed to be $z = 0$ (see Fig. 21.1). We consider a plane reference wave and assume, for simplicity, that it is propagating in the xz plane inclined at an angle θ with the z direction (see Fig. 21.1). Thus, the field associated with this plane wave is given by

$$\begin{aligned} r(x, y, z) &= A \cos(\mathbf{k} \cdot \mathbf{r} - \omega t) \\ &= A \cos(kx \sin\theta + kz \cos\theta - \omega t) \end{aligned} \quad \textbf{(2)}$$

If $r(x, y)$ represents the field at the plane $z = 0$ due to this reference wave, then one can see that

$$\begin{aligned} r(x, y) &= A \cos(kx \sin\theta - \omega t) \\ &= A \cos(2\pi\alpha x - \omega t) \end{aligned} \quad \textbf{(3)}$$

where $\alpha = \sin\theta/\lambda$ is the spatial frequency (see Sec. 19.9). The above equation represents the field due to a plane wave inclined at an angle θ with the z axis, and as can be seen, the phase varies linearly with x. Notice that there is no y dependence because the plane wave has been assumed to have its propagation vector in the xz plane. Thus the total field at the photographic plate (which is coincident with the plane $z = 0$) is given by

$$u(x, y, t) = a(x, y) \cos[\phi(x, y) - \omega t] + A \cos(2\pi\alpha x - \omega t) \quad \textbf{(4)}$$

The photographic plate responds only to the intensity which is proportional to the time average of $[u(x, y, t)]^2$. Thus, the intensity pattern recorded by the photographic plate is

$$\begin{aligned} I(x, y) &= \langle u^2(x, y, t) \rangle \\ &= \langle \{a(x, y) \cos[\phi(x, y) - \omega t] \\ &\quad + A \cos(2\pi\alpha x - \omega t)\}^2 \rangle \end{aligned} \quad \textbf{(5)}$$

where the angular brackets denote time averaging (see Sec. 17.3). Thus

$$\begin{aligned} I(x, y) &= a^2(x, y) \langle \cos^2[\phi(x, y) - \omega t] \rangle \\ &\quad + A^2 \langle \cos^2(2\pi\alpha x - \omega t) \rangle \\ &\quad + 2a(x, y) A \langle \cos[\phi(x, y) - \omega t] \cos(2\pi\alpha x - \omega t) \rangle \end{aligned} \quad \textbf{(6)}$$

Since

$$\langle \cos^2[\phi(x, y) - \omega t] \rangle = \tfrac{1}{2} = \langle \cos^2(2\pi\alpha x - \omega t) \rangle \quad \textbf{(7)}$$

and

$$\begin{aligned} &\langle \cos[\phi(x, y) - \omega t] \cos(2\pi\alpha x - \omega t) \rangle \\ &\quad = \tfrac{1}{2} \langle \cos[\phi(x, y) + 2\pi\alpha x - 2\omega t] \rangle \\ &\qquad + \tfrac{1}{2} \langle \cos[\phi(x, y) - 2\pi\alpha x] \rangle \\ &\quad = \tfrac{1}{2} \cos[\phi(x, y) - 2\pi\alpha x] \end{aligned} \quad \textbf{(8)}$$

Eq. (6) becomes

$$\begin{aligned} I(x, y) &= \tfrac{1}{2} a^2(x, y) + \tfrac{1}{2} A^2 \\ &\quad + Aa(x, y) \cos[\phi(x, y) - 2\pi\alpha x] \end{aligned} \quad \textbf{(9)}$$

From the above relation it is obvious that the phase information of the object wave, which is contained in $\phi(x, y)$, is recorded in the intensity pattern.

When the photographic plate (which has recorded the above intensity pattern) is developed, one obtains a hologram [see Fig. 21.3(b) and (d)]. The transmittance of the hologram, i.e., the ratio of the transmitted field to the incident field, depends on $I(x, y)$. By a suitable developing process one can obtain a condition under which the amplitude transmittance is linearly related to $I(x, y)$. Thus, in such a case if $R(x, y)$ represents the field of the reconstruction wave at the hologram plane, then the transmitted field is given by

$$\begin{aligned} v(x, y) &= KR(x, y) I(x, y) \\ &= K \left[\tfrac{1}{2} a^2(x, y) + \tfrac{1}{2} A^2 \right] R(x, y) \\ &\quad + KAa(x, y) R(x, y) \cos[\phi(x, y) - 2\pi\alpha x] \end{aligned} \quad \textbf{(10)}$$

where K is a constant. We consider the case when the reconstruction wave is identical to the reference wave $r(x, y)$ (see Fig. 21.2). In such a case we obtain (omitting the constant K)

$$\begin{aligned} v(x, y) &= \left[\tfrac{1}{2} a^2(x, y) + \tfrac{1}{2} A^2 \right] A \cos(2\pi\alpha x - \omega t) \\ &\quad + A^2 a(x, y) \cos(2\pi\alpha x - \omega t) \cos[\phi(x, y) - 2\pi\alpha x] \\ &= \left[\tfrac{1}{2} a^2(x, y) + \tfrac{1}{2} A^2 \right] A \cos(2\pi\alpha x - \omega t) \\ &\quad + \tfrac{1}{2} A^2 a(x, y) \cos[\phi(x, y) - \omega t] \\ &\quad + \tfrac{1}{2} A^2 a(x, y) \cos[4\pi\alpha x - \phi(x, y) - \omega t] \end{aligned} \quad \textbf{(11)}$$

Equation (11) gives the transmitted field in the plane $z = 0$. We consider each of the three terms separately. The first term

is nothing but the reconstruction wave itself whose amplitude is modulated due to the presence of the term $a^2(x, y)$. This part of the total field is traveling in the direction of the reconstructed wave. The second term is identical (within a constant term) to the RHS of Eq. (1) and hence represents the original object wave; this gives rise to a virtual image. Thus the effect of viewing this wave is the same as viewing the object itself. The reconstructed object wave is traveling in the same direction as the original object wave.

To study the last term, we first observe that in addition to the term $4\pi\alpha x$, the phase term $\phi(x, y)$ carries a negative sign. The negative sign represents the fact that the wave has a curvature opposite to that of the object wave. Thus if the object wave is a diverging spherical wave, then the last term represents a converging spherical wave. Thus in contrast to the second term, this wave forms a real image of the object which can be photographed by simply placing a film (see Fig. 21.2).

To determine the effect of the term $4\pi\alpha x$, we consider the case when the object wave is also a plane wave traveling along the z axis. For such a wave $\phi(x, y) = 0$, and the last term represents a plane wave propagating along a direction $\theta' = \sin^{-1}(2 \sin \theta)$. Thus the effect of the term $4\pi\alpha x$ is to rotate the direction of the wave. Hence the last term on the RHS of Eq. (11) represents the conjugate of the object wave propagating along a direction different from that of the reconstruction wave and the object wave, which forms a real image of the object. Since the waves represented by the three terms are propagating along different directions, they separate after traversing a distance and enable the observer to view the virtual image without any disturbance.

A very interesting property possessed by holograms is that even if the hologram is broken up into different fragments, each separate fragment is capable of producing a complete virtual image of the object.[3] This property can be understood from the fact that for a diffusely reflecting object, each point of the object illuminates the complete hologram and consequently each point in the hologram receives waves from the complete object. But the resolution in the image decreases as the size of the fragment decreases. For nondiffusely reflecting objects or for transparencies, one makes use of an additional diffusing screen through which the object is illuminated.

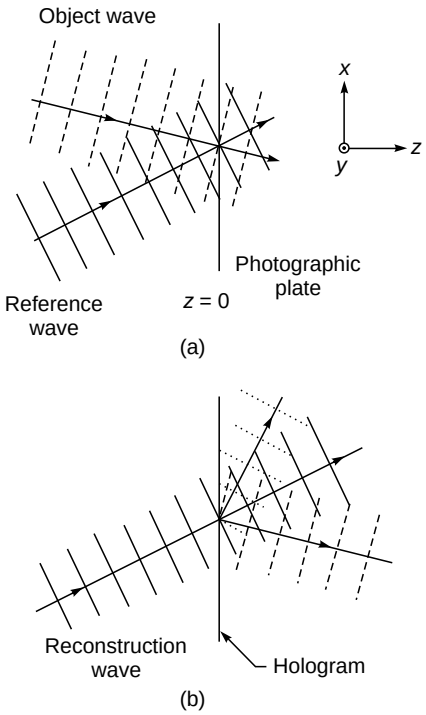


Fig. 21.4 (a) Formation of a hologram when both the object wave and the reference wave are plane waves. (b) Reconstruction of the hologram with another plane wave.

Example 21.1 As an explicit example of the formation and reconstruction of a hologram, we consider the simple case when both the object wave and the reference wave are plane waves [see Fig. 21.4(a)]—a plane object wave corresponds to a single object point lying far away from the hologram. (a) Show that for such a case the hologram consists of a series of Young's interference fringes having an intensity distribution of the $\cos^2$ type. (see also Fig. 14.11). (b) If we reconstruct the hologram with another plane wave [see Fig. 21.4(b)], then show that the transmitted light consists of a zeroth-order plane wave and two first-order plane waves; the two first-order waves correspond to the primary and conjugate waves.

Solution: (a) Consider a plane wave with its propagation vector lying in the xz plane and making an angle θ_1 with the z axis. For such a wave, the field is of the form

$$A_1 \cos[kx \sin\theta_1 + kz \cos\theta_1 - \omega t]$$

If the photographic film is assumed to coincide with the plane $z = 0$, then the field distribution on this plane is given by

$$A_1 \cos(kx \sin\theta_1 - \omega t)$$

Similarly the field (on the plane of the film) due to a plane wave making an angle θ_2 with the z axis, is given by

$$A_2 \cos(kx \sin\theta_2 - \omega t)$$

[3] This property of a hologram exists only when the object is a diffuse scatterer such that the wave from each scattering point of the object reaches all parts of the hologram plate. There are cases where this does not hold good, for example, when a hologram of a transparency is to be recorded.

The resultant intensity distribution is proportional to

$$\langle [A_1 \cos (kx \sin \theta_1 - \omega t) + A_2 \cos (kx \sin \theta_2 - \omega t)]^2 \rangle$$

$$= \frac{1}{2} A_1^2 + \frac{1}{2} A_2^2 + A_1 A_2 \cos [kx (\sin \theta_1 - \sin \theta_2)]$$

$$= \frac{1}{2} (A_1 - A_2)^2 + 2A_1 A_2 \cos^2 \left[\frac{kx}{2} (\sin \theta_1 - \sin \theta_2) \right]$$

For $A_1 = A_2$, the above expression simplifies to

$$2A^2 \cos^2 \left[\frac{kx}{2} (\sin \theta_1 - \sin \theta_2) \right]$$

showing that the intensity remains constant along lines parallel to the y axis with fringe spacing depending on the values of θ_1 and θ_2. Further, the intensity distribution is of the $\cos^2$ type (cf. Fig. 14.11).(b) Before we calculate the transmitted field of the hologram, we first consider a narrow slit of width b being illuminated by a plane wave (see Fig. 21.5). Consider an element ds at a distance s from the center of the slit. Then the amplitude at a far away point P due to this element is proportional to $\sin [k(r - s \sin \theta) - \omega t]\, ds$; here $k = 2\pi/\lambda$ and θ is defined in Fig. 21.5. Thus the total field in the direction θ is given by

$$E \approx A \int_{-b/2}^{+b/2} \sin [k (r - s \sin \theta) - \omega t]\, ds \qquad \textbf{(12)}$$

where A is a constant. The above integral can also be written as

$$E = A \int_{-b/2}^{+b/2} [\sin (kr - \omega t) \cos (ks \sin \theta) - \cos (kr - \omega t) \sin (ks \sin \theta)]\, ds$$

$$= 2A \sin (kr - \omega t) \frac{\sin [(kb/2) \sin \theta]}{k \sin \theta}$$

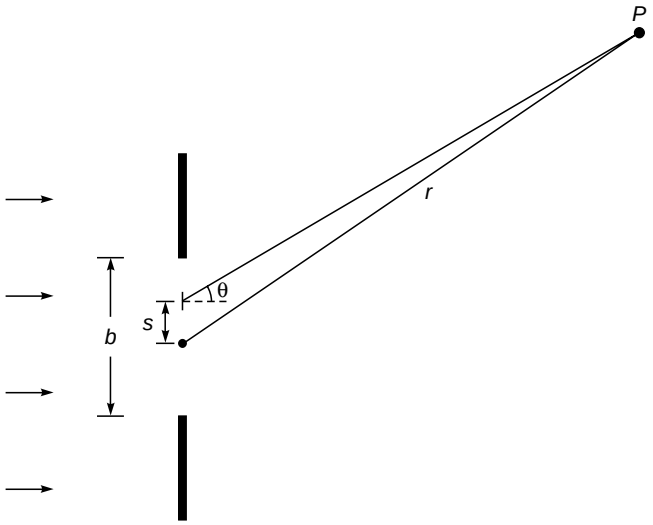


Fig. 21.5 A plane wave incident on a narrow slit of width b.

where the second integral is zero because the integrand is an odd function of s. Thus

$$E = Ab \sin (kr - \omega t) \frac{\sin \beta}{\beta} \qquad \textbf{(13)}$$

where

$$\beta = \frac{1}{2} kb \sin \theta = \frac{\pi b \sin \theta}{\lambda}$$

which is of the same form as obtained in Sec. 18.2. In the present case, the hologram has a $\cos^2 \alpha s$ type of variation in transmittance, and hence the transmitted field will be of the form

$$E = A \int_{-b/2}^{+b/2} \cos^2 \alpha s \sin [kr - ks(\sin\theta - \sin\theta_i) - \omega t]\, ds \qquad \textbf{(14)}$$

where θ_i represents the angle of incidence of the illuminating plane wave. Thus

$$E = \frac{1}{2} \int_{-b/2}^{+b/2} (1 + \cos 2\alpha s) \times \{\sin (kr - \omega t) \cos [ks (\sin \theta - \sin \theta_i)] - \cos (kr - \omega t) \sin [ks (\sin \theta - \sin \theta_i)]\}\, ds$$

$$= \frac{1}{2} A \sin (kr - \omega t) \left[\int_{-b/2}^{+b/2} \cos [ks(\sin \theta - \sin \theta_i)]\, ds + \frac{1}{2} \int_{-b/2}^{+b/2} \cos [ks(\sin \theta - \sin \theta_i + 2\alpha)]\, ds + \frac{1}{2} \int_{-b/2}^{+b/2} \cos [ks(\sin \theta - \sin \theta_i - 2\alpha)]\, ds \right] \qquad \textbf{(15)}$$

The above integrations can easily be carried out. Thus, for example,

$$\int_{-b/2}^{+b/2} \cos [ks (\sin \theta - \sin \theta_i + 2\alpha)] ds$$

$$= \frac{\sin [b (k/2) (\sin \theta - \sin \theta_i + 2\alpha)]}{(k/2)(\sin \theta - \sin \theta_i + 2\alpha)} \qquad \textbf{(16)}$$

which becomes more and more sharply peaked around $\sin \theta = \sin \theta_i - 2\alpha$ as $b \to \infty$, i.e., as the size of the hologram becomes larger. Thus the three integrals in Eq. (15) in the limit of a large value of b give rise to three plane waves propagating along $\sin \theta = \sin \theta_i$, $\sin \theta = \sin \theta_i - 2\alpha$, and $\sin \theta = \sin \theta_i + 2\alpha$, which represent the zeroth-order and two first-order waves.

Example 21.2 Consider the formation of a hologram with a point object and a plane reference wave [see Fig. 14.13(a)]. Choose the z axis to be along the normal from the point source to the plane of the photograph, assumed to be coincident with the plane $z = 0$. For simplicity assume the reference wave to fall normally on the photographic plate. Obtain the interference pattern recorded by the hologram.

Solution: Let the point source be situated at a distance d from the photographic plate. The field at any point $P(x, y, 0)$ on the photographic plate, due to waves emanating from the point object, is given by

$$O(x, y, z = 0, t) = \frac{A}{r}\cos(kr - \omega t) \quad \textbf{(17)}$$

where $r = (x^2 + y^2 + d^2)^{1/2}$ and A represents a constant. A plane wave traveling along a direction parallel to the z axis is given by

$$R(x, y, z, t) = B\cos(kz - \omega t) \quad \textbf{(18)}$$

Hence, the field due to the reference wave at the plane of the photographic plate ($z = 0$) is

$$R(x, y, z = 0, t) = B\cos\omega t \quad \textbf{(19)}$$

Thus, the total field at the plane of the photographic plate is

$$\begin{aligned} T(x, y, t) &= O(x, y, z = 0, t) + R(x, y, z = 0, t) \\ &= \frac{A}{r}\cos(kr - \omega t) + B\cos\omega t \end{aligned} \quad \textbf{(20)}$$

The recorded intensity pattern is

$$\begin{aligned} I(x, y) &= \langle\, |\, T(x, y, t)\,|^2 \,\rangle \\ &= \left\langle \left| \frac{A}{r}\cos(kr - \omega t) + B\cos\omega t \right|^2 \right\rangle \end{aligned} \quad \textbf{(21)}$$

where, as before, angular brackets denote time averaging. Carrying out the above time averaging, we get

$$I(x, y) = \frac{A^2}{2r^2} + \frac{B^2}{2} + \frac{AB}{r}\cos kr \quad \textbf{(22)}$$

If we assume that $d >> x, y$ (which is valid in most practical cases), we can write

$$r = (x^2 + y^2 + d^2)^{1/2} \approx d + \frac{x^2 + y^2}{2d} \quad \textbf{(23)}$$

Thus

$$I(x, y) = \frac{A^2}{2d^2} + \frac{B^2}{2} + \frac{AB}{r}\cos\left[kd + \frac{k}{2d}(x^2 + y^2)\right] \quad \textbf{(24)}$$

The resultant fringe pattern is circular and centered at the origin (see Example 14.7). The hologram thus formed is essentially a zone plate with the transmittance varying sinusoidally in contrast to the Fresnel zone plate [see Fig. 14.13(b) and Sec. 20.3].

21.3 REQUIREMENTS

Since holography is essentially an interference phenomenon, certain coherence requirements have to be met. In Chap. 17 we introduced the notion of coherence length. Thus, if stable interference fringes are to be formed (so that they are recordable), the maximum path difference between the object wave and the reference wave should not exceed the coherence length. Further, the spatial coherence is important so that the waves scattered from different regions of the object could interfere with the reference beam.

During reconstruction, the reconstructed image depends on both the wavelength and the position of the reconstructing source. Hence if the resolution in the reconstructed image has to be good, the source must not be broad and must be emitting a narrow band of wavelengths. It may be worthwhile to mention here that the reconstruction process has associated with it aberrations similar to those in the images formed by lenses. If the reconstruction source is of the same wavelength and is situated at the same relative position with respect to the hologram as the reference source, then the reconstructed image does not suffer from any aberrations.

Another critical requirement in making holograms is stability of the recording arrangement. Thus, the film, the object, and any mirrors used in producing the reference beam must be motionless with respect to one another during exposure. One more requirement which is not so obvious (but is a necessity) is the resolution of the film. Two plane waves making angles $+\theta$ and $-\theta$ with the axis produce an interference pattern with spacing $d = \lambda/(2\sin\theta)$. Assuming $\theta = 15°$ and $\lambda = 6328$ Å (He-Ne laser), one obtains $d = 1.222 \times 10^{-3}$ mm; thus the spatial frequency is 818 lines/mm. Thus the photographic plate should be able to record fringes as close as 0.1222×10^{-4} mm apart. This requires special kinds of material which tend to be exceedingly slow, thus taking the stability requirements even further. Some of the holographic materials are 649F Kodak or 10E 75 or 8E 75 Agfa–Gaevert films and plates.

21.4 SOME APPLICATIONS

The principle of holography finds applications in many diverse fields.[4] The ability to record information about the depth finds application in studying transient microscopic events. Thus, if one has to study some transient phenomenon which occurs in a certain volume, then by using ordinary microscopic techniques it becomes difficult to first locate the position and make observations. If a hologram is recorded of the scene, then the event gets frozen into the hologram and one can focus through the depth of the reconstructed image to study the phenomenon at leisure.

One of the most promising applications of holography lies in the field of interferometry. The ability of the holographic process to release the object wave when reconstructed with a reconstruction wave allows us to perform interference between different waves which exist at different times. Thus, in

[4] See, e.g., Refs. 3 to 12.

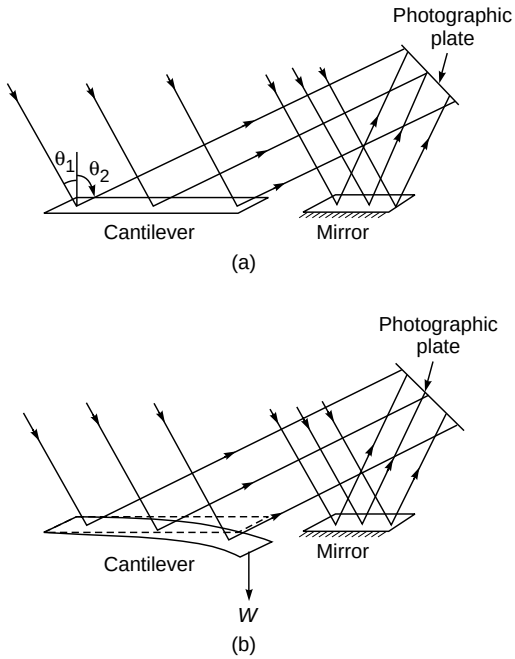


Fig. 21.6 (a) Recording of the unstressed object wave. (b) Recording of the stressed object wave on the same emulsion to produce the doubly exposed hologram.

the technique called double-exposure holographic interferometry, the photographic plate is first partially exposed to the object wave and the reference wave.[5] Then the object is stressed, and the photographic plate is again exposed along with the same reference wave. The photographic plate after development forms the hologram. When this hologram is illuminated with a reconstruction wave, two object waves emerge from the hologram; one corresponds to the unstressed object and the other to the stressed object. Since the object waves themselves have been reconstructed, they interfere and produce interference fringes. These interference fringes are characteristic of the strain suffered by the body. A quantitative study of the fringe pattern produced in the body gives the distribution of strain in the object.

To understand the formation of the fringe pattern, we assume that the deformation of the object has been such as to alter only the phase distribution. Thus, if

$$O(x, y, t) = A(x, y) \cos [\phi(x, y) - \omega t] \qquad \textbf{(25)}$$

represents the object wave (in the hologram plane) when the object is unstressed [see Fig. 21.6(a)] and if $O'(x, y, t)$ represents the object wave when the object is stressed [see Fig. 21.6(b)] then we may write

$$O'(x, y, t) = A(x, y) \cos [\phi'(x, y) - \omega t] \qquad \textbf{(26)}$$

where the phase distribution has been assumed to change from $\phi(x, y)$ to $\phi'(x, y)$. On reconstruction, each of the above two object waves emerges from the hologram, and what is observed is the intensity pattern due to interference of the two waves, given by[6]

$$\begin{aligned} I(x, y) &= \langle \{A(x, y) \cos [\phi(x, y) - \omega t] \\ &\quad + A(x, y) \cos [\phi'(x, y) - \omega t]\}^2 \rangle \\ &= A^2(x, y) + A^2(x, y) \cos [\phi'(x, y) - \phi(x, y)] \end{aligned} \qquad \textbf{(27)}$$

Thus, whenever

$$\phi'(x, y) - \phi(x, y) = 2m\pi \qquad m = 0, 1, 2, \ldots \qquad \textbf{(28)}$$

the two waves interfere constructively, and whenever

$$\phi'(x, y) - \phi(x, y) = (2m + 1)\frac{\pi}{2} \qquad m = 0, 1, 2, \ldots \qquad \textbf{(29)}$$

the two waves interfere destructively. Thus, depending on $[\phi'(x, y) - \phi(x, y)]$, one obtains, on reconstruction, the object superimposed with bright and dark fringes (see Fig. 21.7).

We will consider here a simple application of the above technique in the determination of Young's modulus of a material. If we have a bar fixed at one end and loaded at the other and if it results in a displacement δ of the end of the bar, then we can show that[7]

$$\delta = \frac{WL^3}{3YI} \qquad \textbf{(30)}$$

where W is the load, L is the length of the bar, I is the moment of inertia of cross section which for a rectangular bar of width a and thickness b is given by $I = ab^3/12$, and Y represents Young's modulus of the material of the rod. Thus if we could determine δ for a given load, then Y can be determined from Eq. (30).

We will first determine an expression for $\phi' - \phi$. In Fig. 21.6 we have shown the undisplaced and displaced positions of the cantilever illuminated by a laser light along a direction making an angle θ_1 with the z axis. We observe the cantilever along a direction making an angle θ_2 with the z axis. The phase change when the cantilever undergoes a displacement δ as shown in Fig. 21.6(b) is

$$\begin{aligned} \phi' - \phi &= \frac{2\pi}{\lambda}(\delta\cos\theta_1 + \delta\cos\theta_2) \\ &= \frac{2\pi}{\lambda}\delta\,(\cos\theta_1 + \cos\theta_2) \end{aligned} \qquad \textbf{(31)}$$

[5] This example was provided to the author by Prof. R.S. Sirohi.

[6] The reconstruction process produces other wave components also, but as was observed earlier, these components travel along different directions. Here we are concerned only with the object waves.

[7] See, e.g., Ref. 13, p.75.

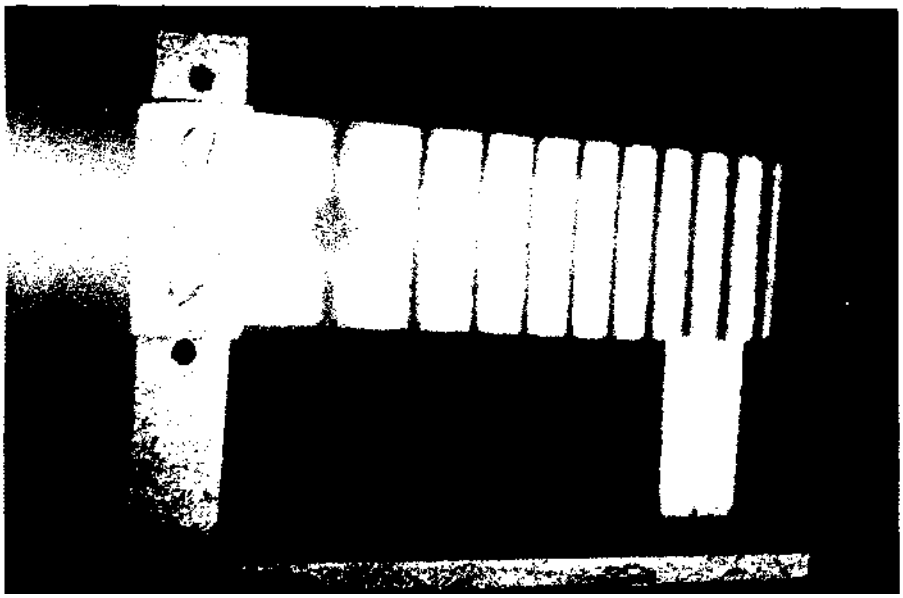


Fig. 21.7 Interference fringes produced in the measurement of Young's modulus using double-exposure interferometry (Photograph courtesy: Prof. R. S. Sirohi).

If there are N fringes over the length L of the cantilever, then since a phase difference of 2π corresponds to one fringe [see Eq. (28)] we can write

$$\frac{2\pi}{\lambda}\,\delta\,(\cos\theta_1 + \cos\theta_2) = N(2\pi)$$

or

$$\delta = \frac{N\lambda}{\cos\theta_1 + \cos\theta_2}$$

Thus by measuring N, θ_1, and θ_2, and knowing λ, δ can be determined. Figure 21.7 shows the reconstruction of a double-exposed hologram of an aluminum strip of width 4 cm, thickness 0.2 cm, and length 12 cm. From the number of fringes formed, one can calculate Young's modulus (see Prob. 21.3).

Summary

- The basic technique in holography is the following: In the recording of the hologram, one superimposes on the object wave another wave called the reference wave, and the photographic plate is made to record the resulting interference pattern. The reference wave is usually a plane wave. This recorded interference pattern forms the hologram and contains information about not only the amplitude but also the phase of the object wave. To view the image, we again illuminate the hologram with another wave, called the reconstruction wave. The reconstruction process leads, in general, to a virtual and a real image of the object scene. The virtual image has all the characteristics of the object such as parallax.
- If the object wave and the reference wave are plane waves, the hologram consists of a series of Young's interference fringes.
- For a point object and a plane reference wave, the hologram is very similar to a zone plate with the transmittance varying sinusoidally in contrast to the Fresnel zone plate.

Problems

21.1 Consider the reconstruction of the hologram as formed in the configuration of Example 21.2, by a plane wave traveling along a direction parallel to the z axis. Show the formation of a virtual and a real image.

21.2 In continuation of Example 21.2, calculate the interference pattern when the incident plane wave makes an angle θ with the z axis (see Fig. 14. 13). Assume $B \approx A/d$.

$$\left[\textbf{Ans}: 4B^2\cos^2\left[k\,d - kx\sin\theta + \frac{k}{2d}(x^2 + y^2)\right]\right]$$

21.3 Figure 21.7 corresponds to the reconstruction of a doubly exposed hologram, the objects corresponding to the unstrained and strained positions of an aluminum bar of width 4 cm, thickness 0.2 cm, and length 12 cm. If the strained position corresponds to a load of 1 g force applied at the end of the bar, calculate Young's modulus of aluminum. Assume $\theta_1 \approx \theta_2 \approx 0$; assume $\lambda = 6328$ Å.

[*Hint: N* represents the number of fringes produced over the length of the cantilever.]

[**Ans:** 0.7×10^{11} N m^{-2}]

REFERENCES AND SUGGESTED READINGS

1. D. Gabor, "A New Microscopic Principle," *Nature*, Vol. 161, p. 777, 1948.
2. K. Thyagarajan and A. K. Ghatak, *Lasers: Theory and Applications*, Plenum Press, New York, 1981. Reprinted by Macmillan India Ltd., New Delhi.
3. J. C. Brown and J. A. Harte, "Holography in the Undergraduate Optics Course," *American Journal of Physics*, Vol. 37, p. 441, 1969.
4. H. J. Caulfield and S. Lu, *The Applications of Holography*, John Wiley & Sons, New York, 1970.
5. R. J. Collier, C. B. Burckhardt, and L. H. Lin, *Optical Holography*, Academic Press, New York, 1971.
6. A. K. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978. Reprinted by Macmillan, New Delhi, 1984.
7. M. P. Givens, "Introduction to Holography," *American Journal of Physics*, Vol. 35, p. 1056, 1967.
8. E. N. Leith and J. Upatnieks, "Photography by Laser," *Scientific American*, Vol. 212, p. 24, June 1965.
9. A. F. Methernal, "Acoustical Holography," *Scientific American*, Vol. 221, p. 36, October 1969.
10. K. S. Pennington, "Advances in Holography," *Scientific American*, Vol. 218, p. 40, February 1968.
11. H. M. Smith, *Principles of Holography*, Wiley Interscience, New York, 1975.
12. D. Venkateshwarulu, "Holography, Theory and Applications," *Journal of Scientific and Industrial Research*, Vol. 29, November 1970.
13. B. L. Worsnop and H. T. Flint, *Advanced Practical Physics for Students*, Asia Publishing House, Bombay, 1951.
14. C. Sakher and Ajoy Ghatak, *Holography*, *Encyclopaedia of Modern Optics*, Eds. R. Guenther, A. Miller, L. Bayvel, and J. Midwinter, Elsevier, 2005.

PART 5
Electromagnetic Character of Light

This part consists of three chapters discussing various aspects of the electromagnetic character of light waves. In Chap. 22, the generation and analysis of various forms of polarized light are discussed followed by a detailed analysis of propagation of electromagnetic waves in anisotropic media including first-principle derivations of wave and ray velocities. Applications such as optical activity and Faraday rotation are also discussed. Chapter 23 is a bit mathematical—starting with Maxwell's equations, various states of polarization are discussed; the wave equation is also derived that led Maxwell to predict the existence of electromagnetic waves. Reflection and refraction of electromagnetic waves by a dielectric interface are discussed in Chap. 24. The results directly explain phenomena such as Brewster's law, total internal reflection, and evanescent waves, and Fabry–Perot transmission resonances.

Chapter Twenty-Two

POLARIZATION AND DOUBLE REFRACTION

As to the other emanation which should produce the irregular refraction, I wished to try what Elliptical waves, or rather spheroidal waves, would do; and these I suppose would spread differently both in the ethereal matter diffused throughout the crystal and in the particles of which it is composed . . .

—Christiaan Huygens

Important Milestones

1669	*Erasmus Bartholinus discovered double refraction in calcite.*
1678	*In the wave theory of light communicated to the Academie des Sciences in Paris, Christiaan Huygens gave the theory of double refraction in calcite, discovered by Bartholinus.*
1809	*Malus showed polarization of light by reflection.*
1811	*David Brewster stated Brewster's law.*
1828	*William Nicol invented the prism which produced polarized light—this prism came to be known as the Nicol prism.*
1929	*Edwin Land, an American scientist and inventor, patented Polaroid, which is the name of a type of synthetic plastic sheet used to polarize light.*

22.1 INTRODUCTION

If we move one end of a string up and down, then a transverse wave is generated [see Fig. 22.1(a)]. Each point of the string executes a sinusoidal oscillation in a straight line (along the x axis), and the wave is, therefore, known as a *linearly polarized wave*. It is also known as a plane polarized wave because the string is always confined to the xz plane. The displacement for such a wave can be written in the form

$$\begin{aligned} x(z,t) &= a\cos(kz - \omega t + \phi_1) \\ y(z,t) &= 0 \end{aligned} \qquad \textbf{(1)}$$

where a represents the amplitude of the wave and ϕ_1 is the phase constant to be determined from the initial condition; the y coordinate of the displacement is always zero. At any instant the displacement will be a cosine curve as shown in Fig. 22.1(a). Further, an arbitrary point $z = z_0$ will execute simple harmonic motion of amplitude a. The string can also be made to vibrate in the yz plane [see Fig. 22.1(b)] for which the displacement is given by

and $$\begin{aligned} y(z,t) &= a\cos(kz - \omega t + \phi_2) \\ x(z,t) &= 0 \end{aligned} \qquad \textbf{(2)}$$

In general, the string can be made to vibrate in any plane containing the z axis. If one rotates the end of the string on the circumference of a circle, then each point of the string will move in a circular path as shown in Fig. 22.2; such a wave is known as a circularly polarized wave, and the corresponding displacement is given by

$$\begin{aligned} x(z,t) &= a\cos(kz - \omega t + \phi) \\ y(z,t) &= a\sin(kz - \omega t + \phi) \end{aligned} \qquad \textbf{(3)}$$

so that $x^2 + y^2$ is a constant ($= a^2$).

We next consider a long narrow slit placed in the path of the string as shown in Fig. 22.3(a). If the length of the slit is along the direction of the displacement, then the entire amplitude will be

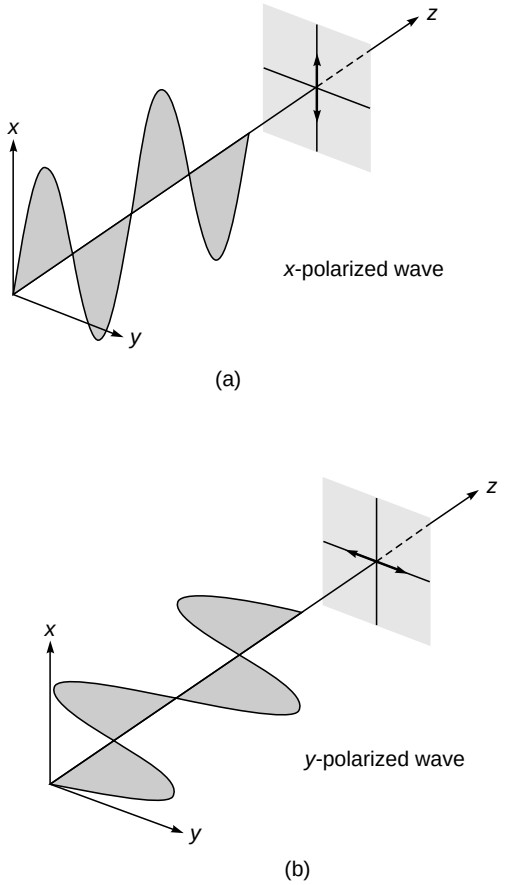


Fig. 22.1 (a) An x-polarized wave on a string with the displacement confined to the xz plane. (b) A y-polarized wave on a string with the displacement confined to the yz plane.

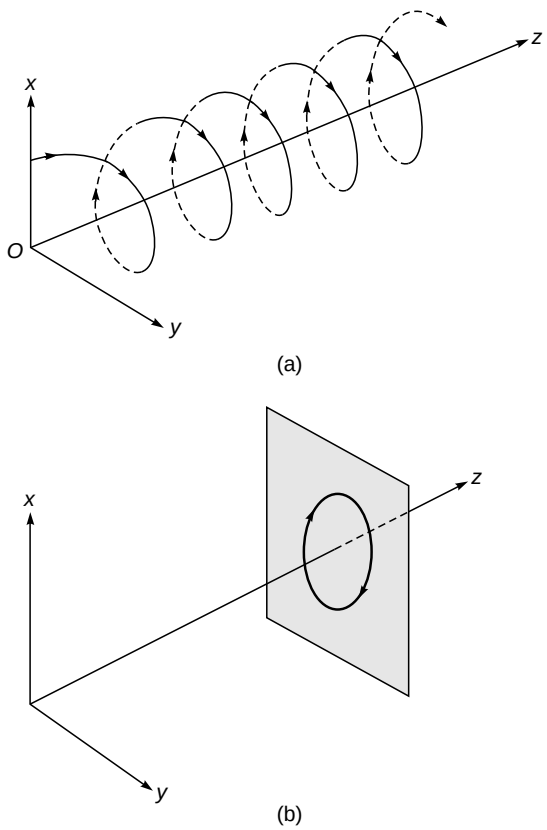


Fig. 22.2 (a) The displacement corresponding to a circularly polarized wave—all points on the string are at the same distance from the z axis. (b) Each point on the string rotates on the circumference of the circle.

transmitted as shown in Fig. 22.3(a). On the other hand, if the slit is at right angles to the direction of the displacement, then almost nothing will be transmitted to the other side of the slit [see Fig. 22.3(b)]. This is so because the slit allows only the component of the displacement, which is along the length of the slit, to pass through. However, if a longitudinal wave were propagating through the string, then the amplitude of the transmitted wave would have been the same for all orientations of the slit. Thus, the change in amplitude of the transmitted wave with the orientation of the slit is due to the transverse character of the wave. Indeed, an experiment which is, in principle, very similar to the experiment discussed above proves the transverse character of light waves. However, before we discuss the experiment with light waves, we must define an unpolarized wave.

We once again consider transverse waves generated at one end of a string. If the plane of vibration is changed in a random manner in very short intervals of time, then such a

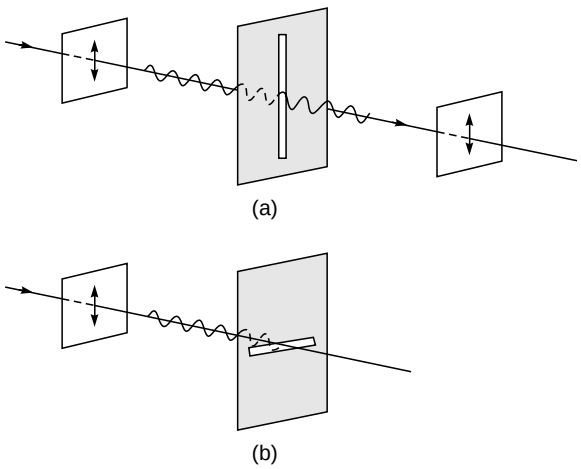


Fig. 22.3 If a linearly polarized transverse wave (propagating on a string) is incident on a long narrow slit, then the slit will allow only the component of the displacement, which is along the length of the slit, to pass through.

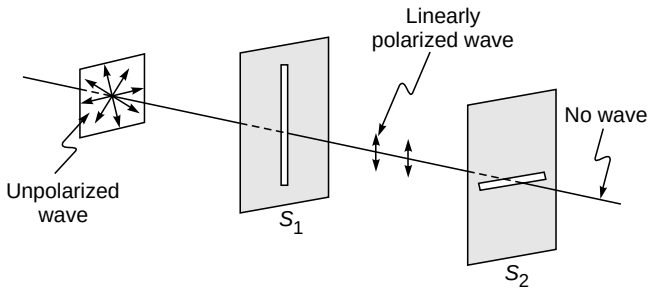


Fig. 22.4 If an unpolarized wave propagating on a string is incident on a long narrow slit S_1, then the transmitted beam is linearly polarized and its amplitude does not depend on the orientation of S_1. If this polarized wave is allowed to pass through another slit S_2, then the intensity of the emerging wave depends on the relative orientation of S_2 with respect to S_1.

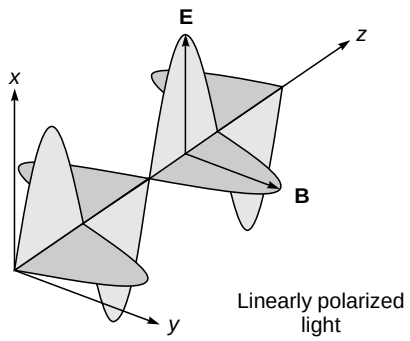


Fig. 22.5 An x-polarized electromagnetic wave propagating in the z direction.

wave is known as an unpolarized wave.[1] If an unpolarized wave falls on a slit S_1 (see Fig. 22.4), then the displacement associated with the transmitted wave is along the length of the slit and a rotation of the slit does not affect the amplitude of the transmitted wave, although the plane of polarization of transmitted wave depends on the orientation of the slit. Thus, the transmitted wave is linearly polarized, and slit S_1 is said to act as a polarizer. If this polarized beam falls on another slit S_2 (see Fig. 22.4), then by rotating slit S_2 we obtain a variation of the transmitted amplitude as discussed earlier; the second slit is said to act as an analyzer.

The transverse character of light waves was known in the early years of the nineteenth century; however, the nature of the displacement associated with a light wave was known only after Maxwell put forward his famous electromagnetic theory. We will discuss the basic electromagnetic theory in Chap. 23 where we will show that associated with a plane electromagnetic wave are an electric field **E** and a magnetic field **B** which are at right angles to each other. For a linearly polarized wave propagating in the z direction, the electric and magnetic fields can be written in the form (see Fig. 22.5)

$$E_x = E_0 \cos(kz - \omega t) \qquad E_y = 0 \qquad E_z = 0 \qquad \textbf{(4)}$$

and

$$B_x = 0 \qquad B_y = B_0 \cos(kz - \omega t) \qquad B_z = 0 \qquad \textbf{(5)}$$

where

$$k = \frac{\omega}{v} = \omega\sqrt{\varepsilon\mu} \qquad \textbf{(6)}$$

and

$$v = \frac{1}{\sqrt{\varepsilon\mu}} \qquad \textbf{(7)}$$

represents the velocity of the waves, and ε and μ are the dielectric permittivity and the magnetic permeability of the medium. Since $E_z = 0$ and $B_z = 0$, the wave is transverse. Equations (4) and (5) also show that **E** and **B** are at right angles to each other and both the vectors are at right angles to the direction of propagation. In fact, the direction of propagation is along the vector $\mathbf{E} \times \mathbf{B}$. The electromagnetic theory also tells us that [see Eq. (33) of Chap. 23]:

$$B_0 = \frac{1}{v} E_0 \qquad \textbf{(8)}$$

Let us next consider an ordinary light beam falling on a Polaroid P_1 as shown in Fig. 22.6(a); a Polaroid is a plastic-like material used for producing polarized light—it will be discussed in detail in the next section. In general, an ordinary light beam (such as the one coming from a sodium lamp or from the sun) is unpolarized; i.e., the electric vector (in a plane transverse to the direction of propagation) keeps changing its direction in a random manner [see Fig. 22.7(a)]. When such a beam is incident on a Polaroid, the emergent light is linearly polarized with its electric vector oscillating in a particular direction as shown in Fig. 22.6(a) [see also Fig. 22.7 (b)]. The direction of the electric vector of the emergent beam will depend on the orientation of the Polaroid. As will be shown in Sec. 22.2, the component of E along a particular direction gets absorbed by the Polaroid, and the component at right angles to it passes through. The direction of the electric vector of the emergent wave is usually called the pass axis of the Polaroid. Returning to Fig. 22.6(a),

[1] By a short interval, we imply times which are short compared to the detection time; however, for the wave to be characterized with a certain frequency ν, this time has to be much greater than $1/\nu$, so that in the short interval it executes a large number of oscillations (see also Sec. 17.1).

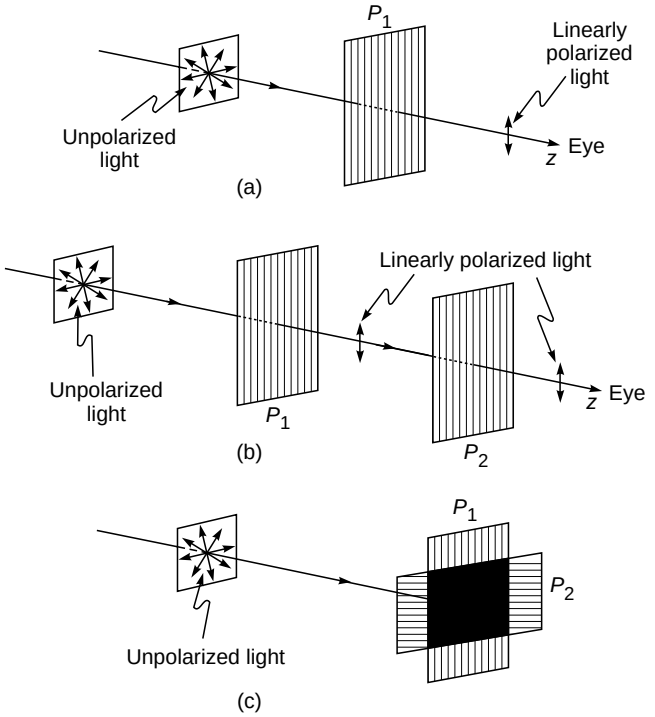


Fig. 22.6 If an ordinary light beam is allowed to fall on a Polaroid, then the emerging beam will be linearly polarized; and if we place another Polaroid P_2, then the intensity of the transmitted light will depend on the relative orientation of P_2 with respect to P_1.

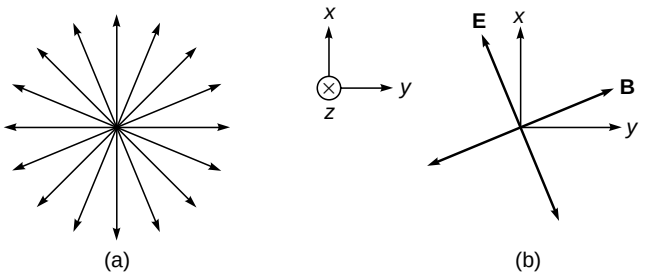


Fig. 22.7 (a) For an unpolarized wave propagating in the $+z$ direction, the electric vector (which lies in the xy plane) continues to change its direction in a random manner. (b) For a linearly polarized wave, the electric (or the magnetic) vector oscillates along a particular direction.

if the position of the eye is as shown in the figure, then one will observe no variation of intensity if the Polaroid is rotated about the z axis. However, if we place another Polaroid P_2 [see Fig. 22.6(b)], then by rotating the Polaroid P_2 (about the z axis) we will observe variation of intensity, when the two Polaroids are parallel, maximum light will pass through the second Polaroid [see Fig. 22.6 (b)] and when the two Polaroids are perpendicular to each other, no light will pass through the second Polaroid [see Fig. 22.6 (c)]. A similar phenomenon will also be observed if instead of rotating the Polaroid P_2 we rotate P_1. On the basis of our earlier discussions, this phenomenon proves the transverse character of light; i.e., the displacement associated with a light wave is at right angles to the direction of propagation of the wave. The Polaroid P_1 acts as a polarizer, and the transmitted beam is linearly polarized. The second Polaroid acts as an analyzer.

22.2 PRODUCTION OF POLARIZED LIGHT

In this section we will discuss various methods for the production of linearly polarized light waves.

22.2.1 The Wire Grid Polarizer and the Polaroid

The physics behind the working of the wire grid polarizer is probably the easiest to understand. It essentially consists of a large number of thin copper wires placed parallel to one another as shown in Fig. 22.8. When an unpolarized electromagnetic wave is incident on it, then the component of the electric vector along the length of the wire is absorbed. This is so because the electric field does work on the electrons inside the thin wires, and the energy associated with the electric field is lost in the Joule heating of the wires. On the other hand, since the wires are assumed to be very thin, the component of the electric vector along the x axis passes through without much attenuation. Thus the emergent wave is linearly polarized with the electric vector along the x axis. However, for the system to be effective (i.e., for the E_y component to be almost completely attenuated) the spacing between the wires should be $\lesssim \lambda$. Clearly, the fabrication of such a polarizer for a 3 cm microwave is relatively easy because the spacing has to be $\lesssim 3$ cm. On the other hand,

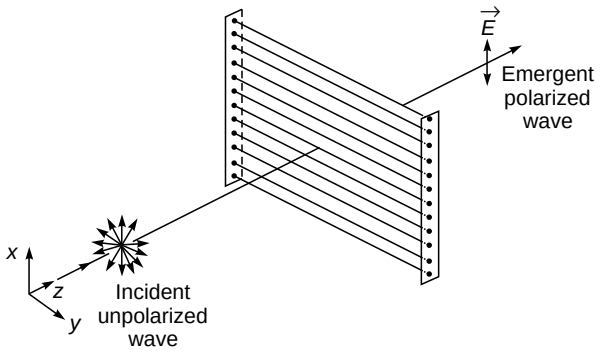


Fig. 22.8 The wire grid polarizer.

since the light waves are associated with a very small wavelength ($\sim 5 \times 10^{-5}$ cm), the fabrication of a polarizer in which the wires are placed at distances $\lesssim 5 \times 10^{-5}$ cm is extremely difficult. Nevertheless, Bird and Parrish did succeed in putting about 30,000 wires in about 1 in; for more details see Ref. 1. The details of the procedure for making this wire grating are also discussed in this book. The original work of Bird and Parrish was published in 1950 (see Ref. 2).

As already pointed out, it is extremely difficult to fabricate a wire grid polarizer which would be effective for visible light. However, instead of long, thin wires, one may employ long chain polymer molecules that contain atoms (such as iodine) which provide high conductivity along the length of the chain. These long chain molecules are aligned so that they are almost parallel to one another. Because of the high conductivity provided by the iodine atoms, the electric field parallel to the molecules gets absorbed. A sheet containing such long chain polymer molecules (which are aligned parallel to one another) is known as a *Polaroid*. When a light beam is incident on such a Polaroid, the molecules (aligned parallel to one another) absorb the component of electric field which is parallel to the direction of alignment because of the high conductivity provided by the iodine atoms; the component perpendicular to it passes through. Thus the aligned conducting molecules act similar to the wires in the wire grid polarizer, and since the spacing between two adjacent long chain molecules is small compared to the optical wavelength, the Polaroid is usually very effective in producing linearly polarized light. The aligning of the long chain conducting molecules is not very difficult; experimental details of producing the polarizer are given in Ref. 1.

22.2.2 Polarization by Reflection

Let us consider the incidence of a plane wave on a dielectric. We assume that the electric vector associated with the incident wave lies in the plane of incidence as shown in Fig. 22.9. It will be shown in Sec. 24.2 that if the angle of incidence θ is such that

$$\theta = \theta_p = \tan^{-1}\left(\frac{n_2}{n_1}\right) \qquad \textbf{(9)}$$

then the reflection coefficient is zero. Thus, if an unpolarized beam is incident at this angle, then the reflected beam will be linearly polarized with its electric vector perpendicular to the plane of incidence (see Fig. 22.10). Equation (9) is referred to as *Brewster's law,* and at this angle of incidence, the reflected and the transmitted rays are at right angles to one another; the angle θ_p is known as the polarizing angle or the Brewster angle.

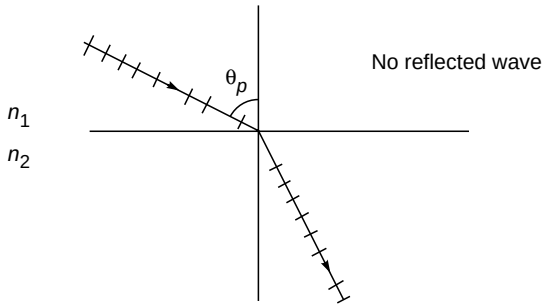


Fig. 22.9 If a linearly polarized wave (with its **E** in the plane of incidence) is incident on the interface of two dielectrics with the angle of incidence equal to θ_p [= $\tan^{-1}(n_2/n_1)$], then the reflection coefficient is zero.

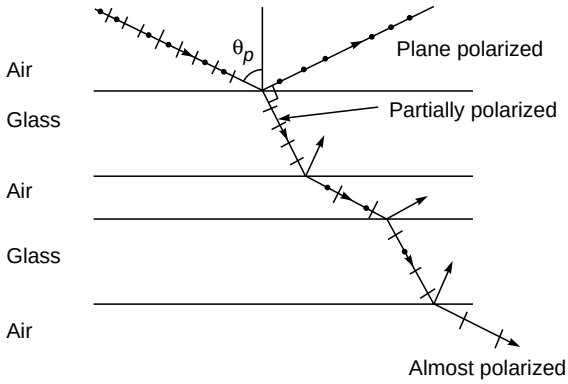


Fig. 22.10 If an unpolarized beam is incident with an angle of incidence equal to θ_p, the reflected beam is plane polarized whose electric vector is perpendicular to the plane of incidence. The transmitted beam is partially polarized, and if this beam is made to undergo several reflections, then the emergent beam is almost plane polarized with its electric vector in the plane of incidence.

For the air-glass interface, $n_1 = 1$ and $n_2 \approx 1.5$, giving $\theta_p \approx 57^0$. The transmitted beam is partially polarized, and if one uses a large number of reflecting surfaces, one obtains an almost plane polarized transmitted beam (see Fig. 22.10).

For the air-water interface, $n_1 \approx 1$ and $n_2 \approx 1.33$ and the polarizing angle $\theta_p \approx 53°$. Thus if the sunlight is incident on the sea at an angle close to the polarizing angle, then the reflected light will be almost polarized. If we now view through a rotating Polaroid, the sea will appear more transparent when the Polaroid blocks the reflected light. Figure 22.11 shows sunlight incident on a water surface at an angle close to the polarizing angle so that the reflected light is almost polarized. If the Polaroid allows the (almost polarized) reflected beam

(a) (b)

Fig. 22.11 If the sunlight is incident on the water surface at an angle close to the polarizing angle, then the reflected light will be almost polarized. (a) If the Polaroid allows the (almost polarized) reflected beam to pass through, we see the glare from the water surface. (b) The glare can be blocked by using a vertical polarizer, and one can see inside the water. Figure adapted from the website www.polarization.com/water/water.html © J. Alcoz, 2001; used with permission of Dr. Alcoz. Color photographs appear in the insert at the back of the book.

to pass through, we see the glare from water surface [see Fig. 22.11(a)]; the glare can be blocked by using a vertical polarizer, and one can see the inside of the water [see Fig. 22.11(b) and Fig. 28 in the insert at the back of the book].

22.2.3 Polarization by Double Refraction

In Secs. 22.5 and 22.12 we will discuss the phenomenon of double refraction and will show that when an unpolarized beam enters an anisotropic crystal, it splits up into two beams, each being characterized by a certain state of polarization. If, by some method, we could eliminate one of the beams, then we would obtain a linearly polarized beam.

A simple method for eliminating one of the beams is through selective absorption; this property of selective absorption is known as dichroism. A crystal such as tourmaline has different coefficients of absorption for the two linearly polarized beams into which the incident beam splits up. Consequently, one of the beams gets absorbed quickly, and the other component passes through without much attenuation. Thus, if an unpolarized beam is passed through a tourmaline crystal, the emergent beam will be linearly polarized (see Fig. 22.12).

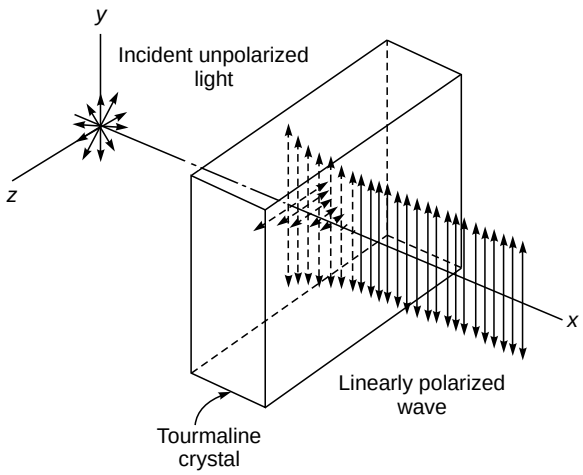


Fig. 22.12 When an unpolarized beam enters a dichroic crystal such as tourmaline, it splits up into two linearly polarized components. One of the components gets absorbed quickly, and the other component passes through without much attenuation [Adapted from Ref. 3; used with permission].

Another method for eliminating one of the polarized beams is through total internal reflection. We will show in Secs. 22.5 and 22.12 that the two beams have different velocities, and as such the corresponding refractive indices will be different. If one can sandwich a layer of a material whose refractive index lies between the two, then for one of the beams, the incidence will be at a rarer medium and for the other it will be at a denser medium. This principle is used in a Nicol prism which consists of a calcite crystal cut in such a way that for the beam, for which the sandwiched material is a rarer medium, the angle of incidence is greater than the critical angle. Thus this particular beam will be eliminated by total internal reflection. Figure 22.13 shows a properly cut calcite crystal in which a layer of Canada balsam has been introduced so that the ordinary ray undergoes total internal reflection. The extraordinary component passes through, and the beam emerging from the crystal is linearly polarized.

22.2.4 Polarization by Scattering

If an unpolarized beam is allowed to fall on a gas, then the beam scattered at 90° to the incident beam is linearly polarized. This follows from the fact that the waves propagating

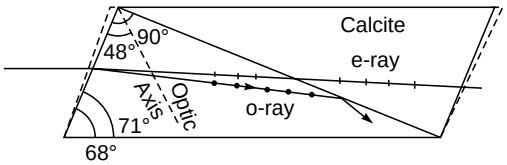
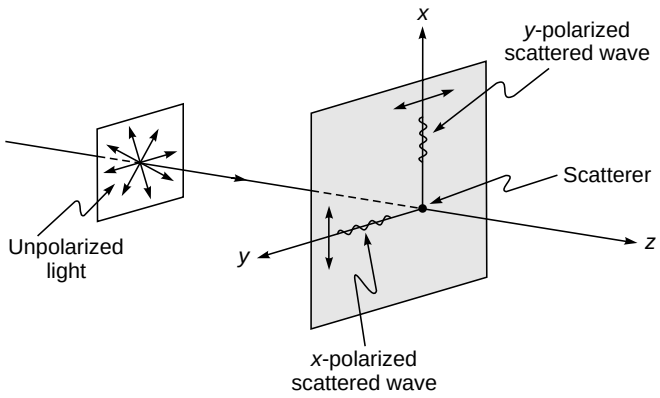


Fig. 22.13 The Nicol prism. The dashed outline corresponds to the natural crystal which is cut in such a way that the ordinary ray undergoes total internal reflection at the Canada balsam layer.



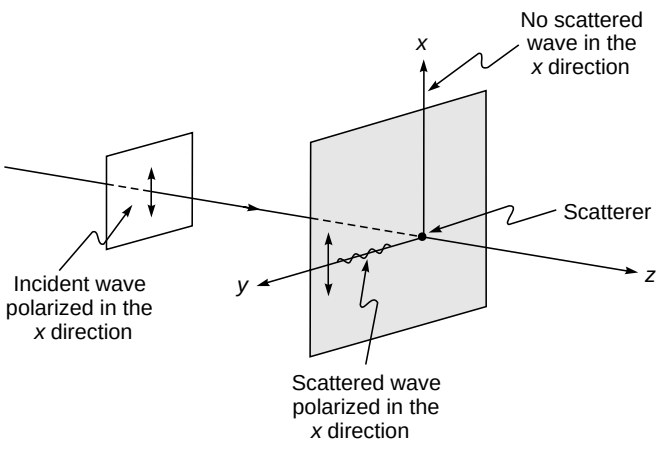


Fig. 22.14 (a) If the electromagnetic wave is propagating along the z direction, then the scattered wave along any direction that is perpendicular to the z axis will be linearly polarized. (b) If a linearly polarized wave (with its **E** oscillating along the x direction) is incident on a dipole, then there will be no scattered wave in the x direction.

in the y direction are produced by the x component of the dipole oscillations (see Fig. 22.14). The y component of the dipole oscillations will produce no field in the y direction (see Sec. 23.4.1). Indeed, it was through scattering experiments that Barkla could establish the transverse character of X-rays. Clearly, if the incident beam is linearly polarized with its electric vector along the x direction, then there will be no scattered light along the x axis. As such, one can carry out an analysis of a scattered wave by allowing it to undergo a further scattering [see Fig. 22.14(b)].

As discussed in Sec. 7.6, the blue color of the sky is due to Rayleigh scattering of sunlight by molecules in our atmosphere. When the Sun is about to set, if we look vertically upward, light will have a high degree of polarization; this is so because the angle of scattering will be very close to 90°. If we view the blue sky (which is vertically above us) with a rotating Polaroid, we will observe considerable variation of intensity.

22.3 MALUS' LAW

Let us consider a polarizer P_1 which has a pass axis parallel to the x axis (see Fig. 22.15); i.e., if an unpolarized beam propagating in the z direction is incident on the polarizer, then the electric vector associated with the emergent wave will oscillate along the x axis. Note that if the polarizer is a Polaroid, then for the pass axis to be along the x direction, the long chain molecules must be aligned along the y axis. We next consider the incidence of the x-polarized beam on the Polaroid P_2 whose pass axis makes an angle θ with the x axis (see Fig. 22.15). If the amplitude of the incident electric field is E_0, then the amplitude of the wave emerging from the Polaroid P_2 will be $E_0 \cos\theta$, and thus the intensity of the emerging beam will be given by

$$I = I_0 \cos^2\theta \qquad \textbf{(10)}$$

where I_0 represents the intensity of the emergent beam when the pass axis of P_2 is also along the x axis (i.e., when $\theta = 0$). Equation (10) represents *Malus' law.* Thus, if a linearly

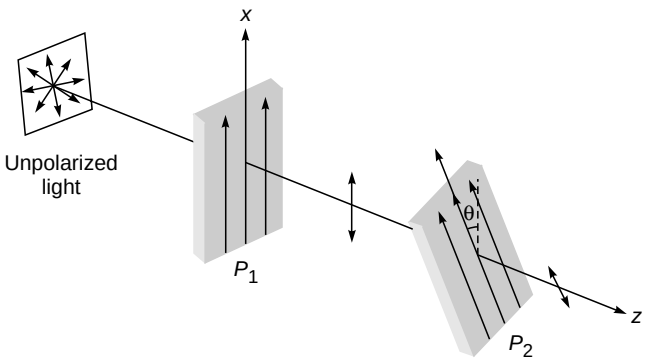


Fig. 22.15 An unpolarized light beam gets x-polarized after passing through the polaroid P_1, the pass axis of the second polaroid P_2 makes an angle θ with the x axis. The intensity of the emerging beam will vary as $\cos^2\theta$.

polarized beam is incident on a Polaroid and if the Polaroid is rotated about the z axis, then the intensity of the emergent wave will vary according to the above law. For example, if the Polaroid P_2 shown in Fig. 22.15 is rotated in the clockwise direction, then the intensity will increase until the pass axis is parallel to the x axis; a further rotation will result in a decrease in intensity until the pass axis is parallel to the y axis, where the intensity will be almost zero. If we further rotate it, it will pass through a maximum and again a minimum before it reaches its original position.

22.4 SUPERPOSITION OF TWO DISTURBANCES

Let us consider the propagation of two linearly polarized electromagnetic waves (both propagating along the z axis) with their electric vectors oscillating along the x axis. The electric fields associated with the waves can be written in the form

$$\mathbf{E}_1 = \hat{\mathbf{x}}\, a_1 \cos(kz - \omega t + \theta_1) \quad \textbf{(11)}$$

$$\mathbf{E}_2 = \hat{\mathbf{x}}\, a_2 \cos(kz - \omega t + \theta_2) \quad \textbf{(12)}$$

where a_1 and a_2 represent the amplitudes of the waves, $\hat{\mathbf{x}}$ represents the unit vector along the x axis, and θ_1 and θ_2 are phase constants. The resultant of these two waves is given by

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 \quad \textbf{(13)}$$

which can always be written in the form

$$\mathbf{E} = \hat{\mathbf{x}}\, a \cos(kz - \omega t + \theta) \quad \textbf{(14)}$$

where

$$a = [a_1^2 + a_2^2 + 2a_1a_2 \cos(\theta_1 - \theta_2)]^{1/2} \quad \textbf{(15)}$$

represents the amplitude of the wave. Equation (14) tells us that the resultant is also a linearly polarized wave with its electric vector oscillating along the same axis.

We next consider the superposition of two linearly polarized electromagnetic waves (both propagating along the z axis) but with their electric vectors oscillating along two mutually perpendicular directions. Thus, we may have

$$\mathbf{E}_1 = \hat{\mathbf{x}}\, a_1 \cos(kz - \omega t) \quad \textbf{(16)}$$

$$\mathbf{E}_2 = \hat{\mathbf{y}}\, a_2 \cos(kz - \omega t + \theta) \quad \textbf{(17)}$$

For $\theta = n\pi$, the resultant will also be a linearly polarized wave with its electric vector oscillating along a direction making a certain angle with the x axis; this angle will depend on the relative values of a_1 and a_2.

To find the state of polarization of the resultant field, we consider the time variation of the resultant electric field at an arbitrary plane perpendicular to the z axis which we may, without any loss of generality, assume to be $z = 0$. If E_x and E_y represent the x and y components of the resultant field $\mathbf{E}$ $(= \mathbf{E}_1 + \mathbf{E}_2)$, then

$$E_x = a_1 \cos \omega t \quad \textbf{(18)}$$

and

$$E_y = a_2 \cos(\omega t - \theta) \quad \textbf{(19)}$$

where we have used Eqs. (16) and (17) with $z = 0$. For $\theta = n\pi$, the above equations simplify to

$$E_x = a_1 \cos \omega t$$

and

$$E_y = (-1)^n a_2 \cos \omega t \quad \textbf{(20)}$$

from which we obtain

$$\frac{E_y}{E_x} = \pm \frac{a_2}{a_1} \qquad \text{(independent of } t\text{)} \quad \textbf{(21)}$$

where the upper and lower signs correspond to n even and n odd, respectively. In the E_xE_y plane, Eq. (21) represents a straight line; the angle ϕ that this line makes with the E_x axis depends on the ratio a_2/a_1. In fact

$$\phi = \tan^{-1}\left(\pm \frac{a_2}{a_1}\right) \quad \textbf{(22)}$$

The condition $\theta = n\pi$ implies that the two vibrations are either in phase ($n = 0, 2, 4, \ldots$) or out of phase ($n = 1, 3, 5, \ldots$). Thus, the superposition of two linearly polarized electromagnetic waves with their electric fields at right angles to each other and oscillating in phase is again a linearly polarized wave with its electric vector, in general, oscillating in a direction which is different from the fields of either of the two waves. Figure 22.16 is a plot of the resultant field corresponding to Eq. (20) for various values of a_2/a_1. The tip of the electric vector oscillates (with angular frequency ω) along the thick lines shown in the figure. The equation of the straight line is given by Eq. (21).

For $\theta \neq n\pi$ ($n = 0, 1, 2, \ldots$), the resultant electric vector does not, in general, oscillate along a straight line. We first consider the simple case corresponding to $\theta = \pi/2$ with $a_1 = a_2$. Thus,

$$E_x = a_1 \cos \omega t \quad \textbf{(23)}$$

$$E_y = a_1 \sin \omega t \quad \textbf{(24)}$$

If we plot the time variation of the resultant electric vector whose x and y components are given by Eqs. (23) and (24), we find that the tip of the electric vector rotates on the circumference of a circle (of radius a_1) in the counterclockwise direction [see Fig. 22.17(c)], and the propagation is in the $+z$ direction which is coming out of the page. Such a wave is known as a right circularly polarized wave (usually

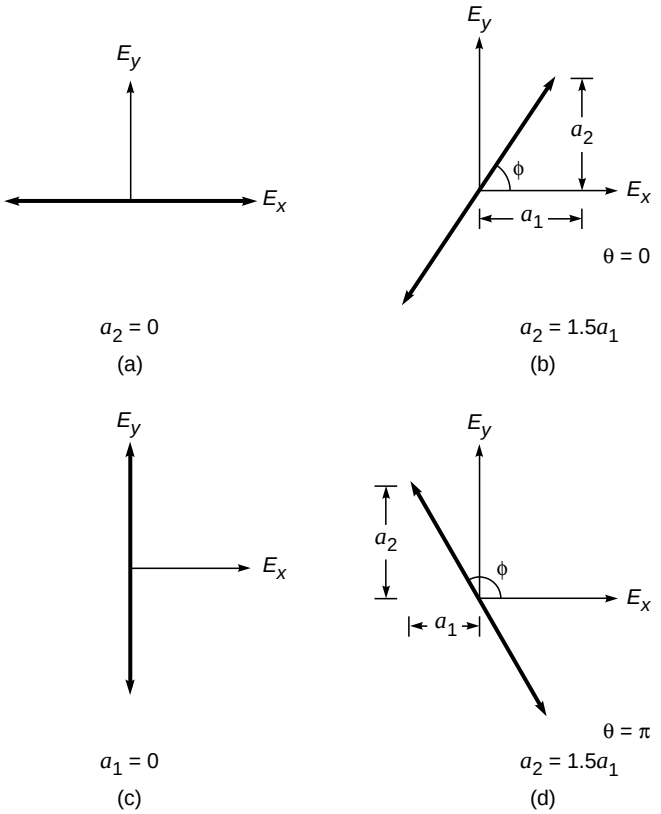


Fig. 22.16 The superposition of two linearly polarized waves with their electric fields oscillating in phase along the x axis and the y axis. The resultant is again a linearly polarized wave with its electric vector oscillating in a direction making an angle ϕ with the x axis.

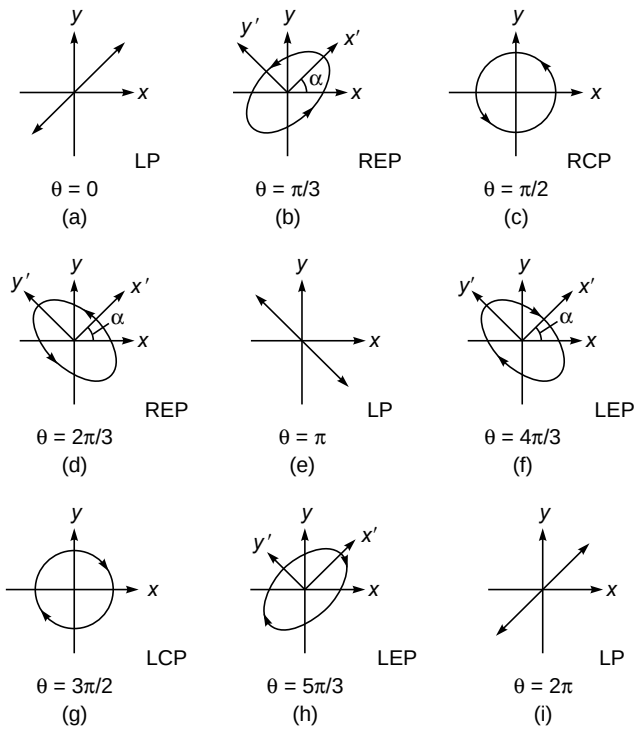


Fig. 22.17 States of polarization for various values of θ corresponding to $a_1 = a_2$ [see Eqs. (18) and (19)]. For example, (c) and (g) correspond to right circularly and left circularly polarized light, respectively; similarly, (b) and (d) correspond to right elliptically polarized (REP) light, and (f) and (h) correspond to left elliptically polarized (LEP) light. The propagation is out of the page.

abbreviated as a RCP wave).[2] That the tip of the resultant electric vector should lie on the circumference of a circle is also obvious from the fact that

$$E_x^2 + E_y^2 = a_1^2 \qquad \text{(independent of } t\text{)}$$

For $\theta = 3\pi/2$, we would have had

$$E_x = a_1 \cos \omega t \tag{25}$$

$$E_y = -a_1 \sin \omega t \tag{26}$$

which would also represent a circularly polarized wave; however, the electric vector will rotate in the clockwise direction [see Fig. 22.17(g)]. Such a wave is known as a left circularly polarized wave (usually abbreviated as a LCP wave).

For $\theta \neq m\pi/2$ ($m = 0, 1, 2, \ldots$), the tip of the electric vector rotates on the circumference of an ellipse (see Fig. 22.17 which corresponds to various values of θ). As can be seen from the figure, this ellipse will degenerate into a straight line or a circle when θ becomes an even or an odd multiple of $\pi/2$. In general, when $a_1 \neq a_2$, one obtains an elliptically polarized wave which degenerates into a straight line for $\theta = 0, \pi, 2\pi, \ldots$, etc. We will show this mathematically in Sec. 22.4.1.

The different states of polarization are a characteristic of any transverse wave. For example, as discussed in Sec. 22.1, if we move a stretched string up and down, we generate a linearly polarized wave with its displacement confined to the vertical plane. Similarly, we may generate a linearly polarized wave with its displacement confined to the horizontal plane [see Fig. 22.1(b)]. Further, we may rotate the end of the string on the circumference of a circle (or an ellipse) to produce a circularly polarized (or an elliptically polarized) wave; similar

[2] Our convention for labeling left and right circularly polarized light is consistent with the one used by Ref. 4, but in some books the opposite convention is used.

to the case of an electromagnetic wave, one may produce an elliptically polarized wave by allowing two linearly polarized waves to propagate through the string. For such a wave, the particles of the string actually move on the circumference of a circle (or an ellipse). On the other hand, for an elliptically polarized electromagnetic wave, it is the electric (or the magnetic) field which changes its magnitude and direction at a particular point; the presence of these fields can be felt by their interaction with a charged particle. In particular, for a circularly polarized wave, the magnitude of the field remains the same; the direction changes with an angular frequency ω. On the other hand, for a linearly polarized wave, the direction of the field does not change; it is the magnitude which keeps on oscillating about the zero value with the angular frequency of the wave.

22.4.1 The Mathematical Analysis

In this section, we will show that Eqs. (18) and (19) represent an elliptically polarized wave. We rewrite Eqs. (18) and (19) as

$$E_x = a_1 \cos \omega t$$
$$E_y = a_2 \cos (\omega t - \theta)$$

We assume that the major axis of the ellipse is along the x' or the y' axes and that the x' axis makes an angle α with the x axis [see Fig. 22.17(b)]; i.e.,

$$E_x' = E_1 \cos(\omega t - \phi) \quad \textbf{(28)}$$

and

$$E_y' = E_2 \sin(\omega t - \phi) \quad \textbf{(29)}$$

Obviously,

$$\left(\frac{E_x'}{E_1}\right)^2 + \left(\frac{E_y'}{E_2}\right)^2 = 1$$

which represents the equation of an ellipse. Now, for the rotated coordinates

$$E_x = E_x' \cos \alpha - E_y' \sin \alpha$$
$$E_y = E_x' \sin \alpha + E_y' \cos \alpha$$

If we multiply the first equation by $\cos \alpha$ and the second equation by $\sin \alpha$ and add, we get

$$E_x' = E_x \cos \alpha + E_y \sin \alpha \quad \textbf{(30)}$$

Similarly,

$$E_y' = -E_x \sin \alpha + E_y \cos \alpha \quad \textbf{(31)}$$

Substituting Eqs. (18), (19), (28), and (29) into Eqs. (30) and (31), we get

$$E_1 \cos(\omega t - \phi) = a_1 \cos \omega t \cos \alpha + a_2 \cos(\omega t - \theta) \sin \alpha$$
$$E_2 \sin(\omega t - \phi) = -a_1 \cos \omega t \sin \alpha + a_2 \cos(\omega t - \theta) \cos \alpha$$

The above equations have to be valid at all times; thus, we equate the coefficients of $\cos \omega t$ and $\sin \omega t$ on both sides of the equation to obtain

$$E_1 \cos \phi = a_1 \cos \alpha + a_2 \cos \theta \sin \alpha$$
$$E_1 \sin \phi = a_2 \sin \theta \sin \alpha$$

and

$$-E_2 \sin \phi = -a_1 \sin \alpha + a_2 \cos \theta \cos \alpha$$
$$E_2 \cos \phi = a_2 \sin \theta \cos \alpha$$

If we square the four equations above and add, we get

$$E_1^2 + E_2^2 = a_1^2 + a_2^2$$

which is to be expected because the total intensity of both beams should be equal. Further,

$$\frac{E_2}{E_1} = \frac{a_2 \sin \theta \cos \alpha}{a_1 \cos \alpha + a_2 \cos \theta \sin \alpha} = \frac{a_1 \sin \alpha - a_2 \cos \theta \cos \alpha}{a_2 \sin \theta \sin \alpha} \quad \textbf{(32)}$$

Thus,

$$a_2^2 \sin^2 \theta \sin \alpha \cos \alpha = a_1^2 \sin \alpha \cos \alpha - a_2^2 \cos^2\theta \sin \alpha \cos \alpha - a_1 a_2 \cos \theta (\cos^2\alpha - \sin^2\alpha)$$

Simple manipulations give

$$\tan 2\alpha = \frac{2 a_1 a_2 \cos \theta}{a_1^2 - a_2^2} \quad \textbf{(33)}$$

We consider some simple examples.

For $a_1 = a_2$ $\quad 2\alpha = \dfrac{\pi}{2} \Rightarrow \alpha = \dfrac{\pi}{4}$ **(34)**

implying that the major (or minor) axis of the ellipse makes 45° with the x axis [see Fig. 22.17(b)]. Further,

$$\frac{E_2}{E_1} = \frac{\sin \theta}{1 + \cos \theta} = \tan \frac{\theta}{2} \quad \textbf{(35)}$$

Thus, for $a_1 = a_2$ and for

$$\theta = \frac{\pi}{3}, \frac{\pi}{2}, \frac{2\pi}{3}, \frac{4\pi}{3}, \frac{3\pi}{2}, \frac{5\pi}{3}$$

$$\frac{E_2}{E_1} = +0.577, 1, 1.732, -1.732, -1, -0.577$$

which correspond to REP, RCP, REP, LEP, LCP, and LEP, respectively, as shown in Fig. 22.17. For example, for $\theta = 4\pi/3$

$$E_x' = E_1 \cos(\omega t - \phi)$$
$$E_y' = -1.732\, E_1 \sin(\omega t - \phi)$$

Thus, the major axis of the ellipse is along the y' axis. To determine the state of polarization, without any loss of generality,

we may choose $t = 0$ at the instant so that ϕ may be assumed to be zero:

$$E_x' = E_1 \cos \omega t$$
$$E_y' = -1.732 E_1 \sin \omega t$$

Thus at

$$t = 0 \qquad E_x' = E_1 \qquad E_y' = 0$$
$$t = \frac{\pi}{2\omega} \qquad E_x' = 0 \qquad E_y' = -1.732 E_1$$
$$t = \frac{\pi}{\omega} \qquad E_x' = -E_1 \qquad E_y' = 0$$

etc., and the electric vector will rotate in the clockwise direction as shown in Fig. 22.17(f).

22.5 THE PHENOMENON OF DOUBLE REFRACTION

When an unpolarized light beam is incident normally on a calcite crystal, it would in general, split up into two linearly polarized beams as shown in Fig. 22.18(a). The beam which travels undeviated is known as the ordinary ray

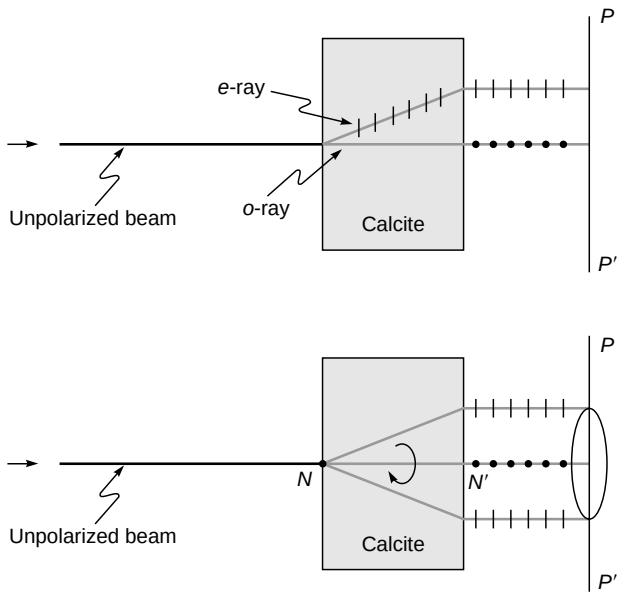


Fig. 22.18 (a) When an unpolarized light beam is incident normally on a calcite crystal, it would in general, split up into two linearly polarized beams. (b) If we rotate the crystal about *NN*,′ then the *e*-ray will rotate about *NN*′.

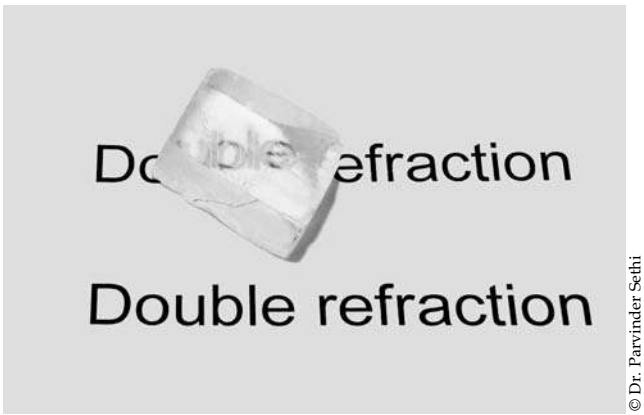


Fig. 22.19 A calcite crystal showing double refraction.

(usually abbreviated as the *o*-ray) and obeys Snell's laws of refraction On the other hand, the second beam, which in general does not obey Snell's laws, is known as the extraordinary ray (usually abbreviated as the *e*-ray). The appearance of two beams is due to the phenomenon of double refraction, and a crystal such as calcite is usually referred to as a *double-refracting* crystal (see Fig. 22.19). If we put a Polaroid *PP*′ behind the calcite crystal and rotate the Polaroid (about *NN*′), then for two positions of the Polaroid (when the pass axis is perpendicular to the plane of the paper) the *e*-ray will be completely blocked and only the *o*-ray will pass through. On the other hand, when the pass axis of the Polaroid is in the plane of the paper (i.e., along the line *PP*′), then the *o*-ray will be completely blocked and only the *e*-ray will pass through. Further, if we rotate the crystal about *NN*,′ then the *e*-ray will rotate about the axis [see Fig. 22.18(b)].

In Sec. 22.13 we will show that whereas the velocity of the ordinary ray is the same in all directions, the velocity of the extraordinary ray is different in different directions; a substance (such as as calcite, quartz) which exhibits different properties in different directions is called an anisotropic substance. Along a particular direction (fixed in the crystal), the two velocities are equal; this direction is known as the optic axis of the crystal. In a crystal such as calcite, the two rays have the same speed only along one direction (which is the optic axis); such crystals are known as uniaxial crystals.[3] The velocities of the ordinary and the extraordinary rays are given by the following equations [see Eq. (123)]:

$$v_{ro} = \frac{c}{n_o} \qquad \text{ordinary ray} \qquad \textbf{(36)}$$

$$\frac{1}{v_{re}^2} = \frac{\sin^2\theta}{(c/n_e)^2} + \frac{\cos^2\theta}{(c/n_o)^2} \qquad \text{extraordinary ray} \qquad \textbf{(37)}$$

[3] In general, there may be two directions along which the two rays have the same speed; such crystals are known as biaxial crystals. The analysis of biaxial crystals is quite difficult; interested readers may look up Refs. 5 and 6.

where n_o and n_e are constants of the crystal and θ is the angle that the ray makes with the optic axis; we have assumed the optic axis to be parallel to the z axis. Thus, c/n_o and c/n_e are the velocities of the extraordinary ray when it propagates parallel and perpendicular to the optic axis. Now, the equation of an ellipse (in the zx plane) is given by

$$\frac{z^2}{a^2}+\frac{x^2}{b^2}=1 \qquad \textbf{(38)}$$

If (ρ, θ) represents the polar coordinates, then $z = \rho\cos\theta$ and $x = \rho\sin\theta$, and the equation of the ellipse can be written in the form

$$\frac{1}{\rho^2}=\frac{\cos^2\theta}{a^2}+\frac{\sin^2\theta}{b^2} \qquad \textbf{(39)}$$

In three dimensions, Eq. (39) will represent an ellipsoid of revolution with the optic axis as the axis of revolution. Thus if we plot v_{re} as a function of θ, we obtain an ellipsoid of revolution; on the other hand, since v_{ro} is independent of θ, if we plot v_{ro} (as a function of θ), we obtain a sphere. Along the optic axis, $\theta = 0$ and

$$v_{ro}=v_{re}=\frac{c}{n_o}$$

We next consider the value of v_{re} perpendicular to the optic axis (i.e., for $\theta = \pi/2$). For a negative crystal $n_e < n_o$ and

$$v_{re}\left(\theta=\frac{\pi}{2}\right)=\frac{c}{n_e}>v_{ro} \qquad \textbf{(40)}$$

Thus the minor axis will be along the optic axis, and the ellipsoid of revolution will lie outside the sphere [see Fig. 22.20(a)]. On the other hand, for a positive crystal $n_e > n_o$ and

$$v_{re}\left(\theta=\frac{\pi}{2}\right)=\frac{c}{n_e}<v_{ro} \qquad \textbf{(41)}$$

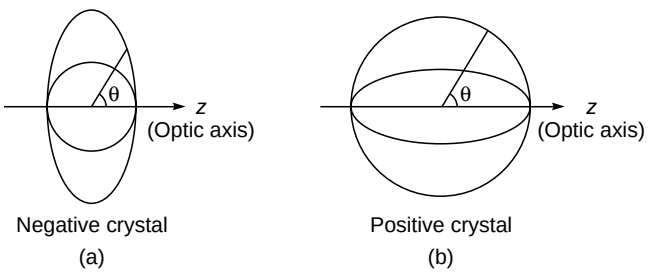


Fig. 22.20 (a) In a negative crystal, the ellipsoid of revolution (which corresponds to the extra ordinary ray) lies outside the sphere; the sphere corresponds to the ordinary ray. (b) In a positive crystal, the ellipsoid of revolution (which corresponds to the extraordinary ray) lies inside the sphere.

The major axis will now be along the optic axis, and the ellipsoid of revolution will lie inside the sphere [see Fig. 22.20(b)]. The ellipsoid of revolution and the sphere are known as the *ray velocity surfaces*.

We next consider an unpolarized plane wave incident on a calcite crystal. The plane wave splits up into two plane waves. One is referred to as the ordinary wave (usually abbreviated as the *o*-wave), and the other is referred to as the extraordinary wave (usually abbreviated as the *e*-wave). For both waves, the space and time dependence of the vectors **E**, **D**, **B**, and **H** can be assumed to be of the form

$$e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$$

where **k** denotes the propagation vector and represents the direction normal to the phase fronts. In general, the **k** vector for the *o*- and *e*-waves will be different. In Sec. 22.12 we will show that

1. Both ordinary and extraordinary waves are linearly polarized.
2. $\mathbf{D}\cdot\mathbf{k}=0$ for both *o*- and *e*-waves **(42)**
   Thus **D** is always at right angles to **k**, and for this reason the direction of **D** is chosen as the direction of "vibrations."
3. If we assume the z axis to be parallel to the optic axis then
   $\mathbf{D}\cdot\hat{\mathbf{z}}=0$ (and $\mathbf{D}\cdot\mathbf{k}=0$) for the *o*-wave **(43)**
   Thus for the *o*-wave, the **D** vector is at right angles to the optic axis as well as to **k**.
4. On the other hand, for the *e*-wave, **D** lies in the plane containing **k** and the optic axis, and of course,
   $$\mathbf{D}\cdot\mathbf{k}=0 \qquad \textbf{(44)}$$

Using the recipe given above, we will consider the refraction of a plane electromagnetic wave incident on a negative crystal such as calcite; a similar analysis can be carried out for positive crystals.

22.5.1 Normal Incidence

We first assume a plane wave incident normally on a uniaxial crystal as shown in Fig. 22.21. Without loss of generality, we can always choose the optic axis to lie on the plane of the paper. The direction of the optic axis is shown as a dashed line in Fig. 22.21. To determine the ordinary ray, with point B as the center, we draw a sphere of radius c/n_o. Similarly, we draw another sphere (of the same radius) from point D. The common tangent plane to these spheres is shown as OO', which represents the wave front corresponding to the ordinary refracted ray. The dots show the direction of vibrations (i.e., direction of **D**) which are perpendicular to **k** and to the optic axis [see Eq. (43)].

To determine the extraordinary ray, we draw an ellipse (centered at point B) with its minor axis ($= c/n_o$) along the optic

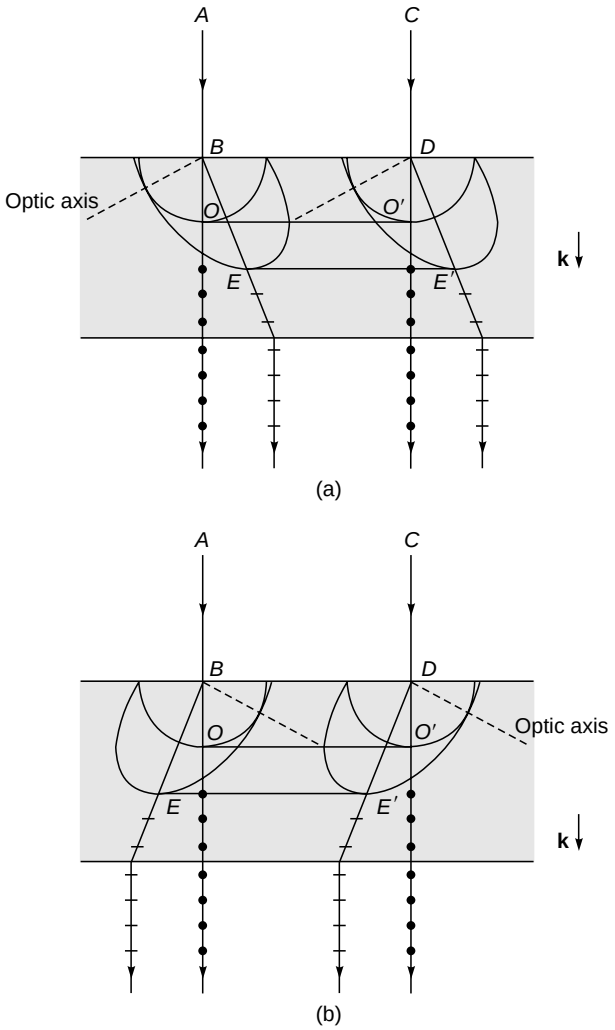


Fig. 22.21 The refraction of a plane wave incident normally on a negative crystal whose optic axis is along the dashed line.

axis and with its major axis equal to c/n_e. The ellipsoid of revolution is obtained by rotating the ellipse about the optic axis. Similarly, we draw another ellipsoid of revolution from point *D*. The common tangent plane to these ellipsoids (which will be perpendicular to **k**) is shown as *EE′* in Fig. 22.21.

If we join point *B* to the point of contact *O*, then corresponding to the incident ray *AB*, the direction of the ordinary ray will be along *BO*. Similarly, if we join point *B* to the point of contact *E* (between the ellipsoid of revolution and the tangent plane *EE′*), then corresponding to the incident ray *AB*, the direction of the extraordinary ray will be along *BE*.

The direction of **k** is the *same* for both *o* and *e*-waves, i.e., both are along *BO*. However, if we have a narrow beam incident as *AB*, then whereas the ordinary ray will propagate along *BO*, the extraordinary ray will propagate in a different direction *BE*; this is also explicitly shown in Fig. 22.18(a).

Obviously, if we have a different direction of the optic axis [see Fig, 22.21(b)], then although the direction of the ordinary ray will remain the same, the extraordinary ray will propagate in a different direction. Thus if a ray is incident normally on a calcite crystal, and if the crystal is rotated about the normal, then the optic axis and the extraordinary ray will also rotate (about the normal) on the periphery of a cone; each time the ray will lie in the plane containing the normal and the optic axis [see Fig. 22.18(b)].

The ray refractive index corresponding to the extraordinary ray n_{re} will be given by

$$n_{re} = \frac{c}{v_{re}} = (n_o^2 \cos^2\theta + n_e^2 \sin^2\theta)^{1/2} \qquad \textbf{(45)}$$

If one starts with Eq. (45) and uses Fermat's principle to obtain the refracted ray, the results will be consistent with the ones obtained in this section (see Sec. 3.5).

Now, as mentioned earlier, the direction of vibrations for the ordinary ray is normal to the optic axis and the vector **k**; as such, the directions of these vibrations in this case will be normal to the plane of the paper and are shown as dots in Fig. 22.21. Similarly, since the direction of vibrations for the extraordinary ray is perpendicular to **k** and lies in the plane containing the extraordinary ray and the optic axis, they are along the small straight lines drawn on the extraordinary ray in Fig. 22.21. Thus, an incident ray will split up into two rays propagating in different directions, and when they leave the crystal, we will obtain two linearly polarized beams.

In the above case, we have assumed the optic axis to make an arbitrary angle α with the normal to the surface. In the special cases of $\alpha = 0$ and $\alpha = \pi/2$, the ordinary and the extraordinary rays travel along the same directions as shown in Fig. 22.22(a), (b), and (c). Figure 22.22(b) corresponds to the case when the optic axis is normal to the plane of the paper; and as such, the section of the extraordinary wave front in the plane of the paper will be a circle. Once again, both the ordinary and the extraordinary rays travel along the same direction. Figure 22.22(a) and (b) corresponds to the same configuration; in both cases the optic axis is parallel to the surface. The figures represent two different cross sections of the same set of spherical and ellipsoidal wave fronts.

Now, corresponding to Fig. 22.22(a) and (b), if the incident wave is polarized perpendicular to the optic axis, it will propagate as an *o*-wave with velocity c/n_o. On the other hand, if the incident wave is polarized parallel to the optic axis, it will propagate as an *e*-wave with velocity c/n_e. In Fig. 22.22(c) the optic axis is normal to the surface, and both waves will travel with the same velocity.

In the configuration shown in Fig. 22.22(a) and (b), although both waves travel in the same direction, they propagate with different velocities. This phenomenon is

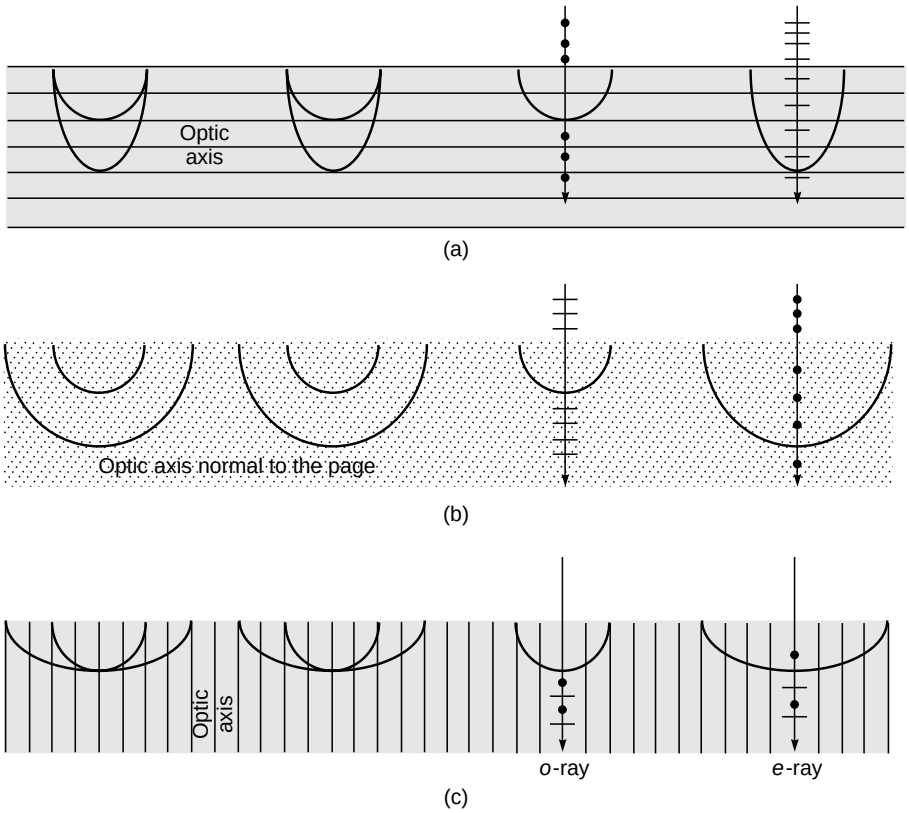


Fig. 22.22 Propagation of a plane wave incident normally on a negative uniaxial crystal. In (a) and (c) the optic axis is shown as parallel straight lines, and in (b) the optic axis is perpendicular to the plane of the figure and is shown as dots. In each case, the extraordinary and the ordinary rays travel in the same direction.

used in the fabrication of quarter and half wave plates (see Sec. 22.6). On the other hand, in the configuration shown in Fig. 22.22(c), both the waves not only travel in the same direction but also propagate with the same velocity.

22.5.2 Oblique Incidence

We next consider the case of a plane wave incident obliquely on a negative uniaxial crystal [see Fig. 22.23(a)]. Once again we use Huygens' principle to determine the shape of the refracted wave fronts. Let *BD* represent the incident wave front. If the time taken for the disturbance to reach point *F* from *D* is *t*, then with *B* as center we draw a sphere of radius $(c/n_o)t$ and an ellipsoid of revolution of semiminor and semimajor axes $(c/n_o)t$, and $(c/n_e)t$, respectively; the semiminor axis is along the optic axis. From point *F* we draw tangent planes *FO* and *FE* to the sphere and the ellipsoid of revolution, respectively. These planes represent the refracted wave fronts corresponding to the ordinary and the extraordinary rays, respectively. If the points of contact are *O* and *E*, then the ordinary and extraordinary refracted rays will propagate along *BO* and *BE*, respectively; this can also be shown using Fermat's principle (see Sec. 3.5). The directions of vibration of these rays are shown by dots and small lines, respectively, and are obtained by using the general rules discussed earlier. The shape of the refracted wave fronts corresponding to the particular case of $\alpha = 0$ and $\alpha = \pi/2$ can be obtained very easily.

Figure 22.23(b) corresponds to the case when the optic axis is normal to the plane of incidence. The sections of both the wave fronts will be circles; consequently, the extraordinary ray will also satisfy Snell's law, and we will have

$$\frac{\sin i}{\sin r} = n_e \qquad \text{for } e\text{-ray when optic axis is normal to plane of incidence} \tag{46}$$

Of course, for the ordinary ray we will *always* have

$$\frac{\sin i}{\sin r} = n_o \tag{47}$$

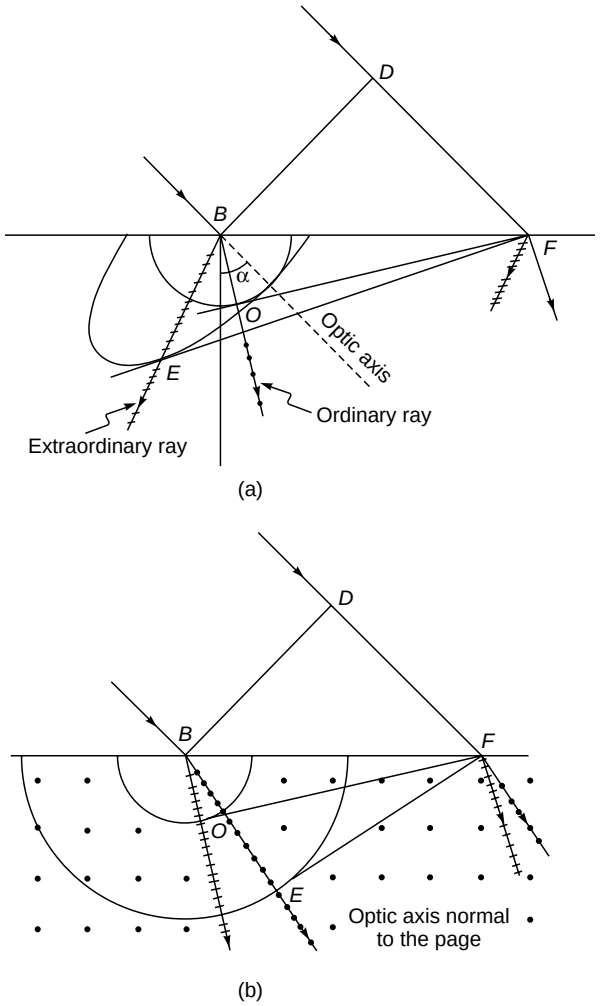


Fig. 22.23 Refraction of a plane wave incident obliquely on a negative uniaxial crystal. In (a), the direction of the optic axis is along the dashed line. In (b), the optic axis is perpendicular to the plane of the paper.

22.6 INTERFERENCE OF POLARIZED LIGHT: QUARTER WAVE PLATES AND HALF WAVE PLATES

In Sec.22.5 we considered how a plane wave (incident on a doubly refracting crystal) splits up into two waves, each characterized by a certain state of polarization. The direction of vibration associated with the ordinary and extraordinary waves is obtained by using the recipe given by Eqs. (43) and (44). In this section, we will consider the normal incidence of a plane polarized beam on a calcite crystal whose optic axis is parallel to the surface of the crystal as shown in Fig. 22.24. We will study the state of polarization of the beam emerging from the crystal. We will assume the z axis to be along the optic axis. Now, as discussed in Sec. 22.5, if the incident beam is y-polarized, the beam will propagate as an ordinary wave and the extraordinary wave will be absent. Similarly, if the incident beam is z-polarized, the beam will propagate as an extraordinary wave and the ordinary wave will be absent. For any other state of polarization of the incident beam, both the extraordinary and the ordinary components will be present. For a negative crystal such as calcite $n_e < n_o$, and the e-wave will travel faster than the o-wave; this is shown by putting s (slow) and f (fast) inside the parentheses in Fig. 22.24.

Let the electric vector (of amplitude E_0) associated with the incident polarized beam make an angle ϕ with the z axis; in Fig. 22.24, ϕ has been shown to be equal to 45°—but for the time being we will keep our analysis general and assume ϕ to be an arbitrary angle. Such a beam can be assumed to be a superposition of two linearly polarized beams (vibrating in

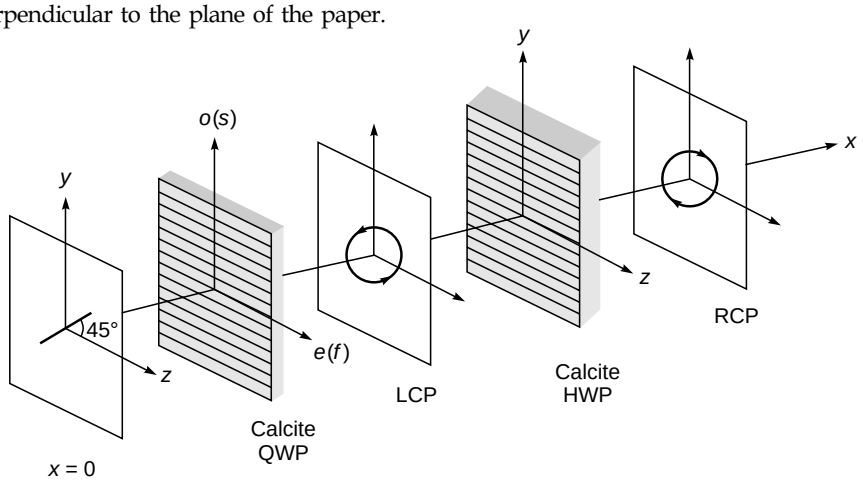


Fig. 22.24 A linearly polarized beam making an angle 45° with the z axis gets converted to a LCP after propagating through a calcite QWP; further, a LCP gets converted to a RCP after propagating through a calcite HWP. The optic axis in the QWP and HWP is along the z direction as shown by lines parallel to the z axis.

phase), polarized along the y and z directions with amplitudes $E_0 \sin\phi$ and $E_0 \cos\phi$, respectively. The z component (whose amplitude is $E_0 \cos\phi$) passes through as an extraordinary beam propagating with wave velocity c/n_e. The y component (whose amplitude is $E_0 \sin\phi$) passes through as an ordinary beam propagating with wave velocity c/n_o. Since $n_e \neq n_o$, the two beams will propagate with different velocities; as such, when they come out of the crystal, they will not be in phase. Consequently, the emergent beam (which will be a superposition of these two beams) will be, in general, elliptically polarized.

Let the plane $x = 0$ represent the surface of the crystal on which the beam is incident. The y and z components of the incident beam can be written in the form

$$E_y = E_0 \sin\phi \cos(kx - \omega t)$$
$$E_z = E_0 \cos\phi \cos(kx - \omega t) \qquad \textbf{(48)}$$

where k ($= \omega/c$) represents the free space wave number. Thus, at $x = 0$, we have

$$E_y(x = 0) = E_0 \sin\phi \cos\omega t$$
$$E_z(x = 0) = E_0 \cos\phi \cos\omega t$$

Inside the crystal, the two components will be given by

$$E_y = E_0 \sin\phi \cos(n_o kx - \omega t) \qquad \text{ordinary wave}$$
$$E_z = E_0 \cos\phi \cos(n_e kx - \omega t) \qquad \text{extraordinary wave}$$

If the thickness of the crystal is d, then at the emerging surface, we have

$$E_y = E_0 \sin\phi \cos(\omega t - \theta_o)$$
$$E_z = E_0 \cos\phi \cos(\omega t - \theta_e)$$

where $\theta_o = n_o kd$ and $\theta_e = n_e kd$. By appropriately choosing the instant $t = 0$, the components may be rewritten as

$$E_y = E_0 \sin\phi \cos(\omega t - \theta)$$
$$E_z = E_0 \cos\phi \cos\omega t \qquad \textbf{(49)}$$

where

$$\theta = \theta_o - \theta_e = kd(n_o - n_e) = \frac{\omega}{c}(n_o - n_e)d \qquad \textbf{(50)}$$

represents the phase difference between the ordinary and the extraordinary beams. Clearly, if the thickness of the crystal is such that $\theta = 2\pi, 4\pi, \ldots$, the emergent beam will have the same state of polarization as the incident beam. Now, if the thickness d of the crystal is such that $\theta = \pi/2$, the crystal is said to be a quarter wave plate (usually abbreviated as QWP)—a phase difference of $\pi/2$ implies a path difference of a quarter of a wavelength. On the other hand, if the thickness of the crystal is such that $\theta = \pi$, the crystal is said to be a half wave plate (usually abbreviated as HWP). For a general value of θ, the state of polarization was discussed in Sec. 22.4.1.

As an example, let us consider the case when $\phi = \pi/4$ and $\theta = \pi/2$; i.e., the y and z components of the incident wave have equal amplitudes, and the crystal introduces a phase difference of $\pi/2$ (see Fig. 22.24). Thus, for the emergent beam we have

$$E_y = \frac{E_0}{\sqrt{2}} \sin\omega t \qquad E_z = \frac{E_0}{\sqrt{2}} \cos\omega t \qquad \textbf{(51)}$$

which represents a circularly polarized wave because

$$E_y^2 + E_z^2 = \frac{E_0^2}{\sqrt{2}}$$

To determine the direction of rotation of the electric vector, we note that at $t = 0$,

$$E_y = 0 \qquad E_z = \frac{E_0}{\sqrt{2}}$$

and at $t = \Delta t$,

$$E_y \approx \frac{E_0}{\sqrt{2}} \omega \Delta t \quad E_z \approx \frac{E_0}{\sqrt{2}}$$

The above equations show that as time increases, the electric vector rotates in the counterclockwise direction, and hence the beam is left circularly polarized as shown in Fig. 22.24. To introduce a phase difference of $\pi/2$, the thickness of the crystal should have a value given by

$$d = \frac{c}{\omega(n_o - n_e)} \frac{\pi}{2} = \frac{1}{4} \frac{\lambda_0}{n_o - n_e} \qquad \textbf{(52)}$$

where λ_0 is the free space wavelength. For calcite,

$$n_o = 1.65836 \qquad n_e = 1.48641$$

which correspond to $\lambda_0 = 5893$ Å at 18°C. Substituting these values, we obtain

$$d = \frac{5893 \times 10^{-8}}{4 \times 0.17195} \text{ cm} \approx 0.000857 \text{ mm}$$

Thus a calcite QWP (at $\lambda_0 = 5893$ Å) will have a thickness of 0.000857 mm and will have its optic axis parallel to the surface; such a QWP will introduce a phase difference of $\pi/2$ between the ordinary and extraordinary components at $\lambda_0 = 5893$ Å. If the thickness is an odd multiple of the above quantity, i.e., if

$$d = (2m + 1) \frac{\lambda_0}{4(n_o - n_e)} \qquad m = 0, 1, 2, \ldots \qquad \textbf{(53)}$$

then in the example considered above (i.e., when $\phi = \pi/4$) it can be easily shown that the emergent beam will be left circularly polarized for $m = 0, 2, 4, \ldots$ and right circularly polarized for $m = 1, 3, 4, \ldots$. The y-polarized o-wave in calcite has a smaller wave velocity ($= c/n_o$), and hence it is referred to as a slow wave and shown as $o(s)$ in Figs. 22.24 and 22.25;

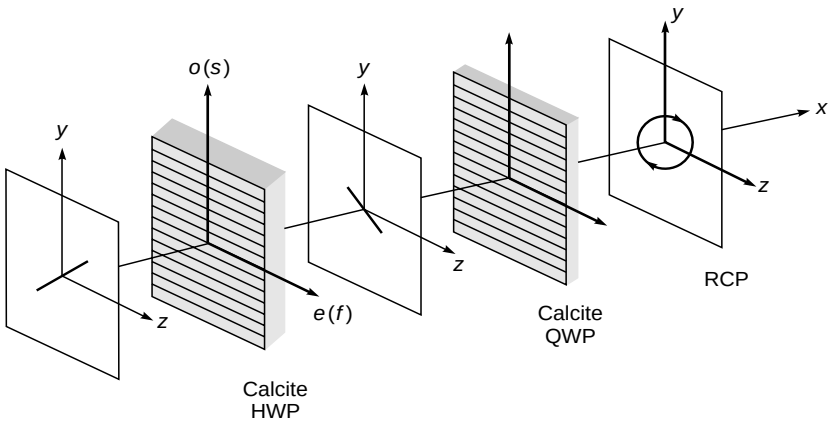


Fig. 22.25 If the linearly polarized beam making an angle 45° with the z axis is incident on a HWP, the plane of polarization gets rotated by 90°; this beam gets converted to a RCP after propagating through a calcite QWP. The optic axis in the HWP and QWP is along the z direction as shown by lines parallel to the z axis.

similarly, the extraordinary wave is the fast wave (in calcite), hence shown as $e(f)$.

We next consider the case when the linearly polarized beam (with $\phi = \pi/4$) is incident on a HWP so that $\theta = \pi$; i.e., the y and z components of the incident wave have equal amplitudes, and the crystal introduces a phase difference of π (see Fig. 22.25). Thus, for the emergent beam we have

$$E_y = -\frac{E_0}{\sqrt{2}} \cos \omega t \qquad E_z = \frac{E_0}{\sqrt{2}} \cos \omega t$$

which represents a linearly polarized wave with the direction of polarization making an angle of 135° with the z axis (see Fig. 22.25). If we now pass this beam through a calcite QWP, the emergent beam will be right circularly polarized as shown in Fig. 22.25. On the other hand, if a left circularly polarized is incident normally on a calcite HWP, the emergent beam will be right circularly polarized as shown in Fig. 22.24.

Thus, for a HWP the thickness (for a negative crystal) is given by

$$d = (2m+1)\,\frac{\lambda_0}{2(n_o - n_e)}$$

If the crystal thickness is such that if $\theta \neq \pi/2, \pi, 3\pi/2, 2\pi, \ldots$, the emergent beam will be elliptically polarized; similar to that shown in Fig. 22.17 (of course, there the propagation was along the z axis, and here it is along the x axis).

For a positive crystal (such as quartz), $n_e > n_o$ and Eqs. (49) should be written in the form

$$\begin{aligned} E_y &= E_0 \sin\phi \cos(\omega t + \theta') \\ E_z &= E_0 \cos\phi \cos \omega t \end{aligned} \qquad \textbf{(54)}$$

where

$$\theta' = \frac{\omega}{c}\, d\,(n_e - n_o)$$

For a quarter wave plate,

$$d = (2m+1)\frac{\lambda_0}{4(n_e - n_o)} \qquad m = 0, 1, 2, \ldots$$

Thus, if in Fig. 22.24 the calcite QWP is replaced by a quartz QWP, the emergent beam will be right circularly polarized.

Example 22.1 A left circularly polarized beam ($\lambda_0 = 5893$ Å) is incident normally on a calcite crystal (with its optic axis cut parallel to the surface) of thickness 0.005141 mm. What will be the state of polarization of the emergent beam?

Solution: The electric field for the incident beam at $x = 0$ is

$$E_y = \frac{E_0}{\sqrt{2}} \sin \omega t \qquad E_z = \frac{E_0}{\sqrt{2}} \cos \omega t \qquad \textbf{(55)}$$

Now

$$\theta = \frac{(n_o - n_e)d \times 2\pi}{\lambda_0}$$

$$= \frac{0.17195 \times 0.005141 \times 2\pi}{5893 \times 10^{-7}} \approx 3\pi$$

Thus the emergent wave will be [cf. Eq. (49)]

$$E_y = \frac{E_0}{\sqrt{2}} \sin(\omega t - 3\pi)$$

$$= -\frac{E_0}{\sqrt{2}} \sin \omega t \qquad E_z = \frac{E_0}{\sqrt{2}} \cos \omega t$$

which represents a right circularly polarized beam.

Example 22.2 A left circularly polarized beam (λ_0 = 5893 Å) is incident on a quartz crystal (with its optic axis cut parallel to the surface) of thickness 0.025 mm. Determine the state of polarization of the emergent beam. Assume n_o and n_e to be 1.54425 and 1.55336, respectively.

Solution: As in Example 22.1, the electric field for the incident beam at $x = 0$ is given by Eq. (55). Further,

$$\theta' = (n_e - n_o)\frac{2\pi}{\lambda_0} d = 2\pi \frac{0.00911 \times 0.025}{5893 \times 10^{-7}} \approx 0.77\pi$$

Thus the emergent beam will be

$$E_y = \frac{E_0}{\sqrt{2}} \cos(\omega t + 0.77\pi) \qquad E_z = \frac{E_0}{\sqrt{2}} \cos(\omega t)$$

which will represent a right elliptically polarized light beam.

22.7 ANALYSIS OF POLARIZED LIGHT

In the previous sections we have seen that a plane wave can be characterized by different states of polarizations, which may be any one of the following:

- Linearly polarized
- Circularly polarized
- Elliptically polarized
- Unpolarized
- Mixture of linearly polarized and unpolarized
- Mixture of circularly polarized and unpolarized
- Mixture of elliptically polarized and unpolarized light

To the naked eye, all the states of polarizations will appear to be the same. In this section, we will discuss the procedure for determining the state of polarization of a light beam.

If we introduce a Polaroid in the path of the beam and rotate it about the direction of propagation, then one of the following three possibilities can occur:

1. If there is complete extinction at two positions of the polarizer, then the beam is linearly polarized.
2. If there is no variation of intensity, then the beam is unpolarized or circularly polarized or a mixture of unpolarized and circularly polarized light. We now put a quarter wave plate on the path of the beam followed by the rotating Polaroid. If there is no variation of intensity, then the incident beam is unpolarized. If there is complete extinction at two positions, then the beam is circularly polarized (this is so because a quarter wave plate will transform a circularly polarized light into a linearly polarized light). If there is a variation of intensity (without complete extinction), then the beam is a mixture of unpolarized and circularly polarized light.
3. If there is a variation of intensity (without complete extinction), then the beam is elliptically polarized or a mixture of linearly polarized and unpolarized or a mixture of elliptically polarized and unpolarized light. We now put a quarter wave plate in front of the Polaroid with its optic axis parallel to the pass axis of the Polaroid at the position of maximum intensity. The elliptically Polarized light will transform to a linearly polarized light. Thus, if one obtains two positions of the Polaroid where complete extinction occurs, then the original beam is elliptically polarized. If complete extinction does not occur and the position of maximum intensity occurs at the same orientation as before, the beam is a mixture of unpolarized and linearly polarized light. Finally, if the position of maximum intensity occurs at a different orientation of the Polaroid, the beam is a mixture of elliptically polarized and unpolarized light.

22.8 OPTICAL ACTIVITY

When a linearly polarized light beam propagates through an "optically active" medium such as sugar solution, then as the beam propagates, its plane of polarization rotates. This rotation is directly proportional to the distance traversed by the beam and also to the concentration of sugar in the solution. Indeed, by measuring the angle by which the plane of polarization is rotated, one can accurately determine the concentration of sugar in the solution.

The rotation of the plane of polarization is due to the fact that the "modes" of the optically active substance are left circularly polarized (LCP) and right circularly polarized (RCP) which propagate with slightly different velocities (see Sec. 22.16). By modes we imply that if a LCP light beam is incident on the substance, then it will propagate as a LCP beam; similarly, a RCP light beam will propagate as a RCP beam but with a slightly different velocity. On the other hand, if a linearly polarized light beam is incident, then we must express the linear polarization as a superposition of a RCP and a LCP beam and then consider the independent propagation of the two beams. We illustrate through an example.

We consider a RCP beam propagating in the +z direction

$$\begin{aligned} E_x^r &= E_0 \cos(k_r z - \omega t) \\ E_y^r &= -E_0 \sin(k_r z - \omega t) \end{aligned} \qquad \textbf{(56)}$$

where $k_r = (\omega/c)\, n_r$ and the superscript and subscript r signify that we are considering a RCP beam. Similarly, a LCP beam (of the same amplitude) propagating in the +z direction

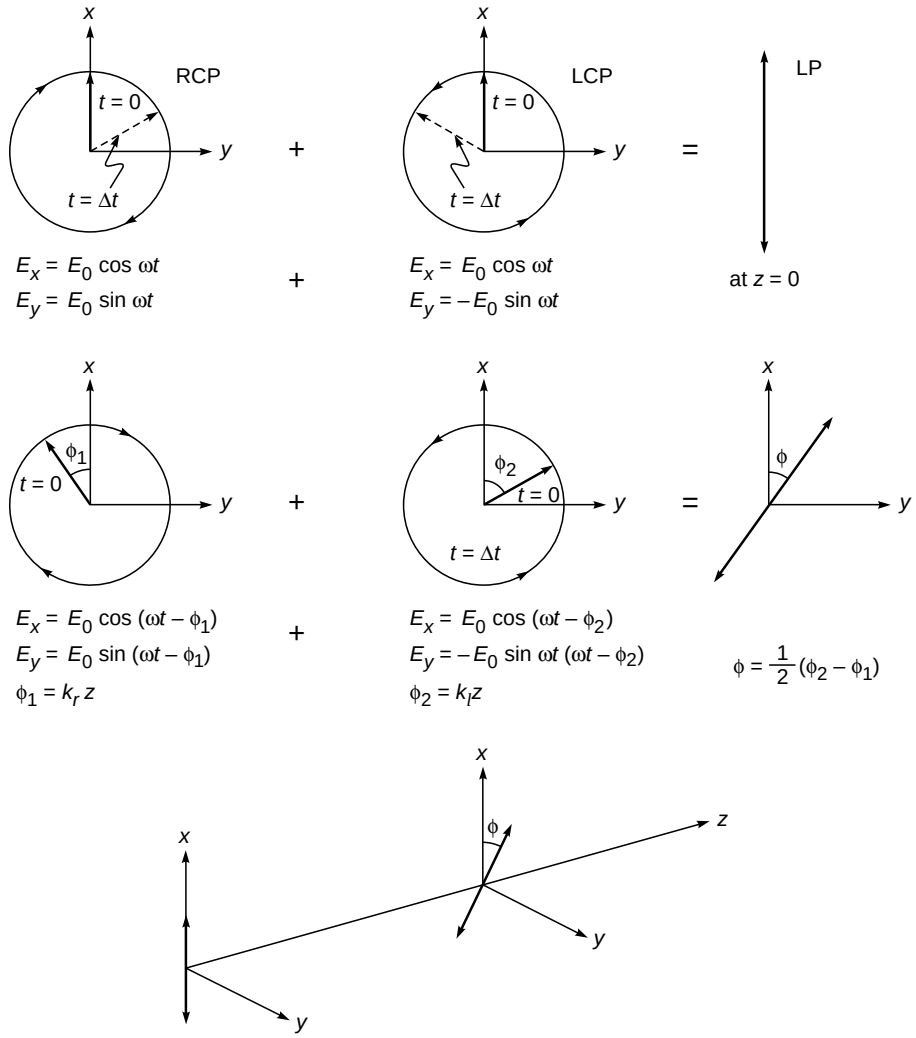


Fig. 22.26 The "clockwise" rotation of a linearly polarized wave as it propagates through a "right-handed" optically active medium.

can be described by the following equations:

$$E_x^l = E_0 \cos(k_l z - \omega t)$$
$$E_y^l = E_0 \sin(k_l z - \omega t) \qquad \textbf{(57)}$$

where $k_l = (\omega/c)\, n_l$; n_r and n_l are the refractive indices corresponding to the RCP and LCP beams, respectively. If we assume the simultaneous propagation of the two beams, then the x and y components of the resultant fields are given by

$$E_x = E_0 [\cos(k_r z - \omega t) + \cos(k_l z - \omega t)]$$

or

$$E_x = 2E_0 \cos\left[\tfrac{1}{2}(k_l - k_r)z\right] \cos[\omega t - \theta(z)]$$

Similarly

$$E_y = 2E_0 \sin\left[\tfrac{1}{2}(k_l - k_r)z\right] \cos[\omega t - \theta(z)]$$

where

$$\theta(z) = \tfrac{1}{2}(k_l + k_r)z$$

Thus the resultant wave is *always* linearly polarized with the plane of polarization rotating with z. If the direction of the oscillating electric vector makes an angle ϕ with the x axis, then (see Fig. 22.26)

$$\phi(z) = \tfrac{1}{2}(k_l - k_r)z$$

or

$$\phi(z) = \frac{\pi}{\lambda_0}(n_l - n_r)\, z = \frac{\omega}{2c}(n_l - n_r)z \qquad \textbf{(58)}$$

where λ_0 is the free space wavelength. Now, if

$n_l > n_r \Leftrightarrow$ the optically active substance is said to be right-handed or dextrorotatory

$n_r > n_l \Leftrightarrow$ the optically active substance is said to be left-handed or levorotatory

For example, for turpentine,

$$\phi = +37° \qquad \text{for } z = 10 \text{ cm}$$

As mentioned earlier, we observe optical activity even in a sugar solution, and this is due to the helical structure of sugar molecules. Determination of the concentration of sugar solutions by measuring the rotation of the plane of polarization is a widely used method in industry. Note that if $n_l = n_r$ (as is indeed the case in an isotropic substance), then $\phi(z) = 0$ and a linearly polarized beam remains linearly polarized along the same direction. Optical activity is also exhibited in crystals. For example, for a linearly polarized light propagating along the optic axis of a quartz crystal,[4] the plane of polarization gets rotated. Indeed

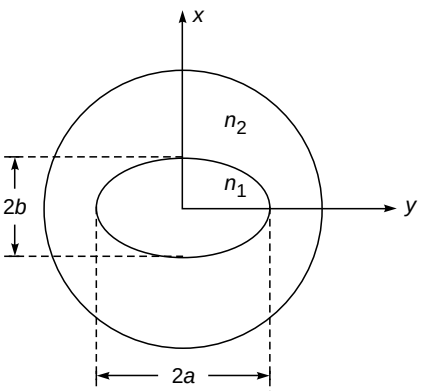
$$|n_l - n_r| \approx 7 \times 10^{-5}$$

$$\Rightarrow \quad \phi \approx \frac{7}{60}\pi = 21°$$

for $z = 0.1$ cm at $\lambda_0 = 6000$ Å

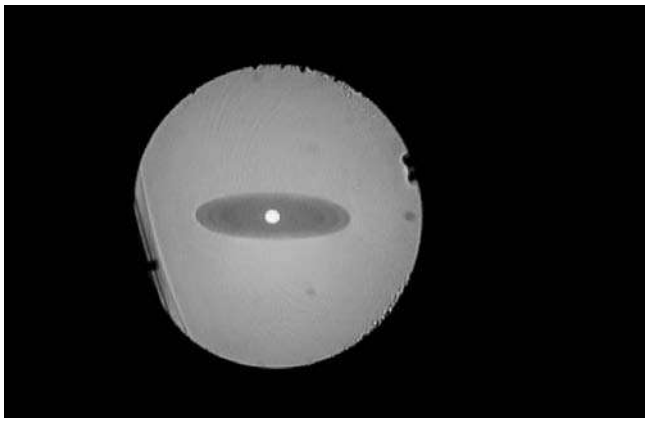
22.9 CHANGE IN THE STATE OF POLARIZATION OF A LIGHT BEAM PROPAGATING THROUGH AN ELLIPTIC CORE SINGLE-MODE OPTICAL FIBER

A very interesting phenomenon is the propagation of polarized light through an elliptic core optical fiber. We will have a brief discussion on optical fibers in Chaps. 27 and 29; it will suffice here to say that in an ordinary optical fiber we have a cylindrical core (of circular cross section) cladded with a medium of slightly lower refractive index. The guidance of the light beam takes place through the phenomenon of total internal reflection (see Figs. 27.2 and 27.7). Because of the circular symmetry of the problem, the incident beam can have any state of polarization[5] which will be maintained as the beam propagates through the fiber. Now, if we have an elliptic core fiber [see Fig. 22.27(a)], then the modes of the fiber

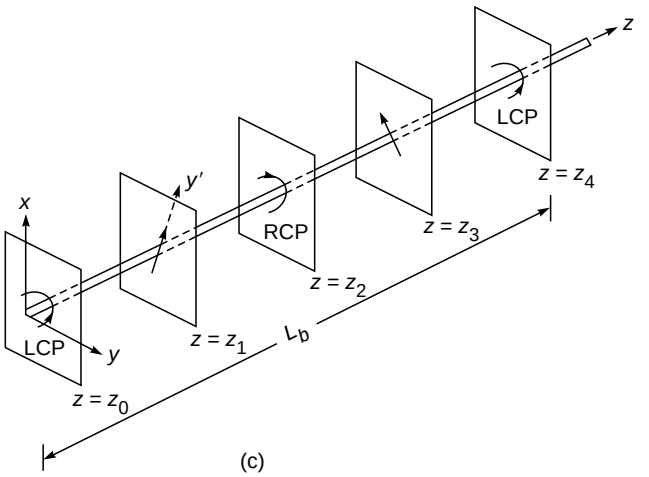


Elliptic core optical fiber

(a)



(b)



(c)

Fig. 22.27 (a) The transverse cross section of an elliptic core fiber; the modes are (approximately) *x*-polarized and *y*-polarized. (b) The actual cross-sectional view of the perform of an elliptic core fiber fabricated at CGCRI, Kolkata [photograph courtesy Dr. Shyamal Bhadra, CGCRI, Kolkata]. (c) Propagation of a left circularly polarized beam incident on an elliptic core fiber. If we view along the *y′* axis, then dark spots will be observed at $z = z_1, 5z_1, 9z_1, \ldots$.

[4] When a wave propagates along the optic axis of a quartz crystal, it is, strictly speaking, not like calcite. The modes are *not* linearly polarized; they are RCP and LCP propagating with slightly different velocities.

[5] We are considering here a single-mode fiber so that no matter what the incident transverse field distribution is, it soon "settles down" to the transverse field distribution of the fundamental mode which propagates with the velocity ω/β_0. This velocity is independent of the state of polarization (SOP) of the incident beam.

are (approximately) x and y polarized; i.e., if an x-polarized beam is incident on the fiber, it will propagate without any change in the state of polarization with a certain phase velocity ω/β_x. Similarly, a y-polarized beam will propagate as a y-polarized beam with a slightly different velocity ω/β_y. Now, let a circularly polarized beam be incident on the input face of the fiber at $z = 0$. Then we must resolve the incident beam into x- and y-polarized beams propagating with slightly different velocities. Thus

$$\mathbf{E}(x, y, z) = \psi(x, y)\,[\,\hat{\mathbf{x}}\cos(\beta_x z - \omega t) + \hat{\mathbf{y}}\sin(\beta_y z - \omega t)] \tag{59}$$

where $\psi(x, y)$ is the transverse field distribution of the fundamental mode which is assumed to be (approximately) the same for both x and y polarizations (see Sec. 28.5). If $\beta_x = \beta_y$, as is indeed true for circular core fibers, the beam will remain circularly polarized for all values of z. Now, at $z = 0$,

$$\begin{aligned} E_x &= \psi(x, y)\cos\omega t \\ E_y &= -\psi(x, y)\sin\omega t \end{aligned} \tag{60}$$

which represents a left circularly polarized wave [see Fig. 22.27 (b)]. For

$$z = z_1 = \frac{\pi}{2(\beta_y - \beta_x)} \tag{61}$$

i.e., for $\beta_y z_1 = \beta_x z_1 + \pi/2$,

$$\begin{aligned} E_x &= \psi(x, y)\cos(\phi_1 - \omega t) \\ &= +\psi(x, y)\cos(\omega t - \phi_1) \\ E_y &= \psi(x, y)\sin\left(\phi_1 + \frac{\pi}{2} - \omega t\right) \\ &= +\psi(x, y)\cos(\omega t - \phi_1) \end{aligned}$$

where

$$\phi_1 = \beta_x z_1$$

which represents a linearly polarized wave [see Fig. 22.27(c)]; we assume the direction of the **E** vector to be along the y' axis. Similarly, at

$$\begin{aligned} z &= z_2 = \frac{\pi}{\beta_y - \beta_x} = 2z_1 \\ E_x &= \psi(x, y)\cos(\phi_2 - \omega t) \\ &= \psi(x, y)\cos(\omega t - \phi_2) \\ E_y &= \psi(x, y)\sin(\phi_2 + \pi - \omega t) \\ &= \psi(x, y)\sin(\omega t - \phi_2) \end{aligned} \tag{62}$$

where

$$\phi_2 = \beta_x z_2$$

and the wave will be right circularly polarized [see Fig. 22.27(c)]. At

$$z = z_3 = \frac{3\pi}{2\beta_x - \beta_y} = 3z_1$$

we have

$$\begin{aligned} E_x &= \psi(x, y)\cos(\phi_3 - \omega t) \\ &= \psi(x, y)\cos(\omega t - \phi_3) \\ E_y &= \psi(x, y)\sin\left(\phi_3 + \frac{3\pi}{2} - \omega t\right) \\ &= -\psi(x, y)\cos(\omega t - \phi_3) \end{aligned}$$

where

$$\phi_3 = \beta_x z_3$$

Thus the wave is again linearly polarized, but now the direction of the oscillating electric field is at right angles to the field at $z = z_1$. In a similar manner, we can easily continue to determine the SOP of the propagating beam. Thus at $z = 5z_1$, $9z_1$, $13z_1$, . . . the SOP will be the same as at $z = z_1$, and at $z = 7z_1$, $11z_1$, $15z_1$, . . . the SOP will be the same as at $z = 3z_1$. Similarly at $z = 4z_1$, $8z_1$, $12z_1$, . . . the beam will be LCP, and at $z = 2z_1$, $6z_1$, $10z_1$, . . . the beam will be RCP.

Now, let the fiber be rotated in such a way that the y' axis is along the vertical line (the x' and the z axes are assumed to lie in the horizontal plane). Thus if we put our eyes vertically above the fiber and view vertically down, then the regions $z = z_1$, $5z_1$, $9z_1$, . . . will appear dark (see Fig. 22.28). This is so because in these regions the electric field is oscillating in the y' direction (which is the vertical direction), and we know that if the dipole oscillates along the y' direction, there is no radiation emitted in that particular direction (see Figs. 22.14 and 23.4). Thus by measuring the distance between two consecutive black spots ($= 4z_1$) we can calculate z_1 and hence $\beta_y - \beta_x$. Furthermore, by moving the eyes to the horizontal plane, i.e., viewing along the x' axis, we see the regions $z = z_1$, $5z_1$, $9z_1$, . . . appear bright and the regions $z = 3z_1$, $7z_1$, $11z_1$, . . . appear dark. Thus the experiment allows us to understand not only the changing SOP of a beam propagating through a birefringent fiber, but also the radiation pattern of an oscillating dipole.

As a numerical example, we consider an elliptical core fiber for which

$$2a = 2.14\,\mu\text{m} \qquad 2b = 8.85\,\mu\text{m}$$
$$n_1 = 1.535 \qquad n_2 = 1.47$$

[see Fig. 22.27(a)]. For such a fiber operating at $\lambda_0 = 6328$ Å ($k_0 \approx 9.929 \times 10^4\ \text{cm}^{-1}$),

$$\frac{\beta_x}{k_0} \approx 1.506845 \quad \text{and} \quad \frac{\beta_y}{k_0} \approx 1.507716$$

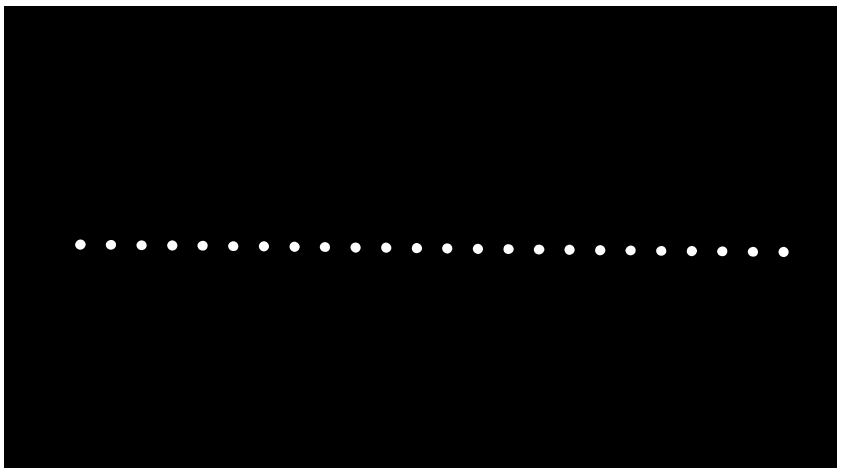


Fig. 22.28 Schematic of the intensity variation as seen from the top (or side) of an elliptic core fiber when a circularly polarized beam is incident on it. The actual photograph from an experiment by Andrew Corporation is given in Ref. 13.

The quantity

$$L_b = \frac{2\pi}{\Delta\beta} = \frac{2\pi}{\beta_y - \beta_x} \approx 0.727 \text{ mm}$$

is known as the coupling length.

22.10 WOLLASTON PRISM

A Wollaston prism is used to produce two linearly polarized beams. It consists of two similar prisms (of, say, calcite) with the optic axis of the first prism parallel to the surface and the optic axis of the second prism parallel to the edge of the prism as shown in Fig. 22.29. Let us first consider the incidence of a *z*-polarized beam as shown in Fig. 22.29(a). The beam will propagate as an *o*-ray in the first prism (because the vibrations are perpendicular to the optic axis) and will see the refractive index n_o. When this beam enters the second prism, it will become an *e*-ray and will see the refractive index n_e. For calcite $n_o > n_e$ and therefore the ray will bend away from the normal. Since the optic axis is normal to the plane of incidence, the refracted ray will obey Snell's laws [see Fig. 22.23(b)], and the angle of refraction will be given by

$$n_o \sin 20° = n_e \sin r_1$$

where we have assumed the angle of the prism to be 20° (see Fig. 22.29). Assuming $n_o \approx 1.658$ and $n_e \approx 1.486$, we readily get

$$r_1 \approx 22.43°$$

Thus the angle of incidence at the second surface is $i_1 = 22.43° - 20° = 2.43°$. The output angle θ_1 is given by $n_e \sin 2.43° = \sin \theta_1 \Rightarrow \theta_1 = 3.61°$.

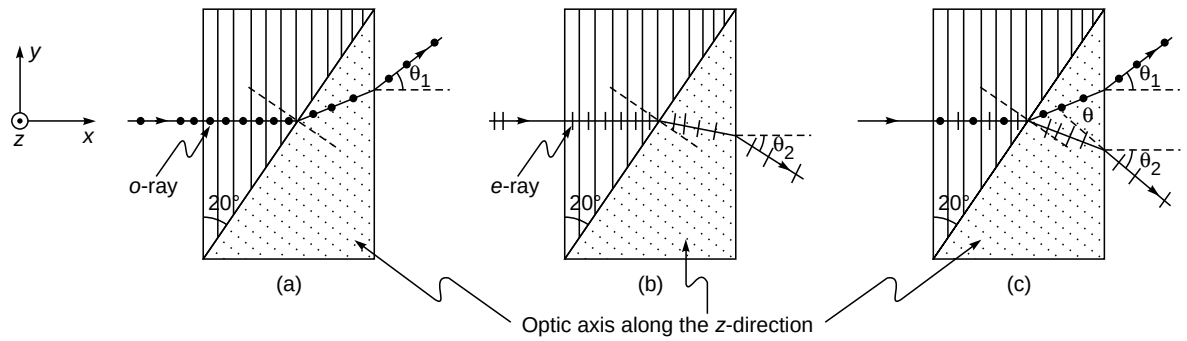


Fig. 22.29 A Wollaston prism. The optic axis of the first prism is along the *y* axis, and the optic axis of the second prism is along the *z* axis. (a) If the incident beam is *z*-polarized, it will propagate as an *o*-wave in the first prism and an *e*-wave in the second prism. (b) If the incident beam is *y*-polarized, it will propagate as an *e*-wave in the first prism and an *o*-wave in the second prism. (c) For an unpolarized beam incident normally, there will be two linearly polarized beams propagating in different directions. The ray paths correspond to prisms being of calcite.

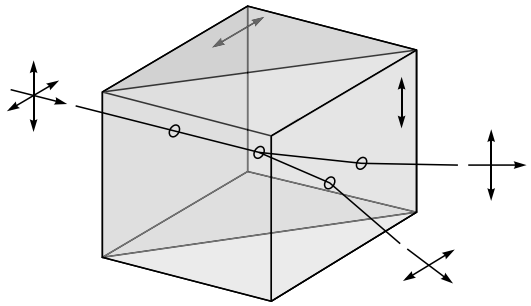


Fig. 22.30 Schematic of an actual Wollaston prism. The prism separates an unpolarized light beam into two linearly polarized beams. It typically consists of two properly oriented calcite prisms (so that the optic axes are perpendicular to each other), cemented together typically with Canada balsam. A commercially available Wollaston prism has divergence angles from 15° to about 45°.

We next consider the incidence of a *y*-polarized beam as shown in Fig. 22.29(b). The beam will propagate as an *e*-ray in the first prism and as an *o*-ray in the second prism. The angle of refraction is now given by

$$n_e \sin 20 = n_o \sin r_2 \quad \Rightarrow \quad r_2 \approx 17.85°$$

Thus the angle of incidence at the second interface is

$$i_2 = 20° - 17.85° = 2.15°$$

The output angle θ_2 is given by

$$n_o \sin 2.15° = \sin \theta_2 \quad \Rightarrow \quad \theta_2 \approx 3.57°$$

Thus, if an unpolarized beam is incident on the Wollaston prism, the angular separation between the two orthogonally polarized beams is $\theta = \theta_1 + \theta_2 \approx 7.18°$; see also Fig. 22.30 and Fig. 30 in the insert at the back of the book.

22.11 ROCHON PRISM

We next consider the Rochon prism which consists of two similar prisms of (say) calcite; the optic axis of the first prism is normal to the face of the prism while the optic axis of the second prism is parallel to the edge as shown in Fig. 22.31. Now, in the first prism both beams will *see* the same refractive index n_o; this follows from the fact that the ordinary and extraordinary waves travel with the same velocity (= c/n_o) along the optic axis of the crystal.

When the beam enters the second crystal, the ordinary ray (whose **D** is normal to the optic axis) will *see* the same refractive index and go undeviated as shown in Fig. 22.31. On the other hand, the extraordinary ray (whose **D** is along the optic axis) will *see* the refractive index n_e and will bend

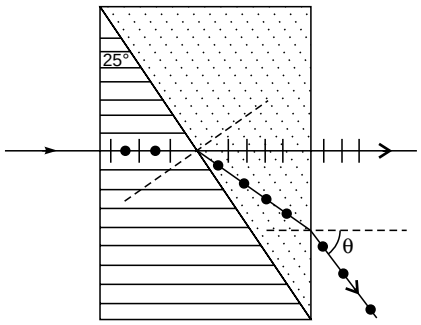


Fig. 22.31 Production of two orthogonally polarized beams by a Rochon prism.

away from the normal. We assume the angle of the prism to be 25°. The angle of refraction will be determined from

$$n_o \sin 25° = n_e \sin r$$

Thus
$$\sin r = \frac{n_o}{n_e} \sin 25°$$

$$= \frac{1.658}{1.486} \times 0.423 \approx 0.472$$

$$\Rightarrow \qquad r = 28.2°$$

Therefore the angle of incidence at the second surface will be 28.2° – 25° = 3.2°. The emerging angle will be given by

$$\sin \theta = n_e \sin (3.2°) \approx 0.083$$

$$\Rightarrow \quad \theta \approx 4.8°$$

22.12 PLANE WAVE PROPAGATION IN ANISOTROPIC MEDIA

In this section, we discuss the plane wave solutions of Maxwell's equations in an anisotropic medium and prove the various assumptions made in Sec. 22.5. The difference between an isotropic and an anisotropic medium lies in the relationship between the displacement vector **D** and the electric vector **E**; the displacement vector **D** is defined in Sec. 23.9. In an isotropic medium, **D** is in the same direction as **E**, and one can write

$$\mathbf{D} = \varepsilon \mathbf{E} \tag{63}$$

where ε is the dielectric permittivity of the medium. On the other hand, in an anisotropic medium **D** is not, in general, in the direction of **E**, and the relation between **D** and **E** can be written in the form

$$\begin{aligned} D_x &= \varepsilon_{xx} E_x + \varepsilon_{xy} E_y + \varepsilon_{xz} E_z \\ D_y &= \varepsilon_{yx} E_x + \varepsilon_{yy} E_y + \varepsilon_{yz} E_z \\ D_z &= \varepsilon_{zx} E_x + \varepsilon_{zy} E_y + \varepsilon_{zz} E_z \end{aligned} \tag{64}$$

where ε_{xx}, ε_{xy}, ... are constants. One can show that[6]

$$\varepsilon_{xy} = \varepsilon_{yx} \qquad \varepsilon_{xz} = \varepsilon_{zx}$$
$$\text{and} \qquad \varepsilon_{yz} = \varepsilon_{zy} \tag{65}$$

Further, one can always choose a coordinate system (i.e., one can always choose appropriately the directions of x, y, and z axes inside the crystal) such that

$$\begin{aligned} D_x &= \varepsilon_x E_x \\ D_y &= \varepsilon_y E_y \\ D_z &= \varepsilon_z E_z \end{aligned} \tag{66}$$

This coordinate system is known as the principal axis system, and the quantities ε_x, ε_y, and ε_z are known as the principal dielectric permittivities of the medium. If

$$\varepsilon_x \neq \varepsilon_y \neq \varepsilon_z \qquad \text{biaxial} \tag{67}$$

we have a biaxial medium and the quantities

$$n_x = \sqrt{\frac{\varepsilon_x}{\varepsilon_0}} \qquad n_y = \sqrt{\frac{\varepsilon_y}{\varepsilon_0}} \qquad n_z = \sqrt{\frac{\varepsilon_z}{\varepsilon_0}} \tag{68}$$

are said to be the principal refractive indices of the medium; in the above equation ε_0 represents the dielectric permittivity of free space ($= 8.8542 \times 10^{-12}$ C^2 N^{-1} m^{-2}). If

$$\varepsilon_x = \varepsilon_y \neq \varepsilon_z \qquad \text{uniaxial} \tag{69}$$

we have a uniaxial medium with the z axis representing the optic axis of the medium. The quantities

$$\begin{aligned} n_o &= \sqrt{\frac{\varepsilon_x}{\varepsilon_0}} = \sqrt{\frac{\varepsilon_y}{\varepsilon_0}} \\ n_e &= n_z = \sqrt{\frac{\varepsilon_z}{\varepsilon_0}} \end{aligned} \tag{70}$$

are known as ordinary and extraordinary refractive indices; typical values for some uniaxial crystals are given in Table 22.1. For a uniaxial medium, since $\varepsilon_x = \varepsilon_y$, the x and y directions can be arbitrarily chosen as long as they are perpendicular to the optic axis; i.e., *any* two mutually perpendicular axes (which are also perpendicular to the z axis) can be taken as the principal axes of the medium.[7] On the other hand, if

$$\varepsilon_x = \varepsilon_y = \varepsilon_z \qquad \text{isotropic} \tag{71}$$

Table 22.1 Ordinary and extraordinary refractive Indices for some uniaxial crystals (Table adapted from Refs. 6 and 7).

Name of the crystal	*Wavelength*	n_o	n_e
Calcite	4046 Å	1.68134	1.49694
	5890 Å	1.65835	1.48640
	7065 Å	1.65207	1.48359
Quartz	5890 Å	1.54424	1.55335
Lithium niobate	6000 Å	2.2967	2.2082
KDP	6328 Å	1.50737	1.46685
ADP	6328 Å	1.52166	1.47685

then we have an isotropic medium and can choose any three mutually perpendicular axes as the principal axis system. We will assume the anisotropic medium to be nonmagnetic so that

$$\mathbf{B} = \mu_0 \mathbf{H}$$

where μ_0 is the free space magnetic permeability.

Let us consider the propagation of a plane electromagnetic wave; for such a wave the vectors **E**, **H**, **D**, and **B** are proportional to exp $[i(\mathbf{k} \cdot \mathbf{r} - \omega t)]$. Thus

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \qquad \mathbf{H} = \mathbf{H}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \\ \mathbf{D} &= \mathbf{D}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \qquad \mathbf{B} = \mathbf{B}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \end{aligned} \tag{72}$$

where the vectors $\mathbf{E}_0$, $\mathbf{H}_0$, $\mathbf{D}_0$, and $\mathbf{B}_0$ are independent of space and time; **k** represents the propagation vector of the wave, and ω is the angular frequency. The wave velocity v_w (also known as the phase velocity) and the wave refractive index n_w are defined through

$$v_w = \frac{\omega}{k} = \frac{c}{n_w} \tag{73}$$

Thus

$$|\mathbf{k}| = k = \frac{\omega}{c} n_w \tag{74}$$

In this section, it is our objective to determine the possible values of n_w when a plane wave propagates through an anisotropic dielectric. Now, in a dielectric medium

$$\text{div } \mathbf{D} = 0 \tag{75}$$

[6] See, e.g., Ref. 5.

[7] This follows from the fact that for a uniaxial medium

$$D_x = \varepsilon_x E_x \qquad \text{and} \qquad D_y = \varepsilon_y E_y = \varepsilon_x E_y$$

Now, if we rotate the x and y axes (about the z axis) by an angle θ and call the rotated axes x' and y', then

$$\begin{aligned} D_{x'} &= D_x \cos\theta + D_y \sin\theta = \varepsilon_x (E_x \cos\theta + E_y \sin\theta) \\ &= \varepsilon_x E_{x'} \end{aligned}$$

Similarly $D_{y'} = \varepsilon_x E_{y'}$, implying that the x' and y' axes can also be chosen as principal axes.

or $$\frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z} = 0$$

For a plane wave given by Eq. (72) the above equation becomes

$$i(k_x D_x + k_y D_y + k_z D_z) = 0$$

or

$$\mathbf{D} \cdot \mathbf{k} = 0 \qquad \textbf{(76)}$$

implying that **D** is *always* at right angles to **k** [see Eq. (44)]. Similarly since in a nonmagnetic medium div **H** = 0,

H will always be at right angles to **k**. **(77)**

Now, in the absence of any currents (i.e., **J** = 0) Maxwell's curl equations [see Eqs. (7) and (8) of Chap. 23] become

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = i\omega \mathbf{B} = i\omega\mu_0 \mathbf{H} \qquad \textbf{(78)}$$

and

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} = -i\omega \mathbf{D} \qquad \textbf{(79)}$$

where we have assumed the medium to be nonmagnetic (i.e., $\mathbf{B} = \mu_0 \mathbf{H}$). Now, if

$$\mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}$$

then

$$\begin{aligned}(\nabla \times \mathbf{E})_x &= \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \\ &= (ik_y E_{0z} - ik_z E_{0y}) e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \\ &= i(k_y E_z - k_z E_y) = i(\mathbf{k} \times \mathbf{E})_x\end{aligned}$$

Thus

$$\nabla \times \mathbf{E} = i(\mathbf{k} \times \mathbf{E}) = i\omega\mu_0 \mathbf{H}$$

$$\Rightarrow \quad \mathbf{H} = \frac{1}{\omega\mu_0}(\mathbf{k} \times \mathbf{E}) \qquad \textbf{(80)}$$

and

$$\nabla \times \mathbf{H} = i(\mathbf{k} \times \mathbf{H}) = -i\omega \mathbf{D}$$

$$\Rightarrow \quad \mathbf{D} = \frac{1}{\omega}(\mathbf{H} \times \mathbf{k}) \qquad \textbf{(81)}$$

Equations (80) and (81) show that

H is at right angles to **k**, **E,** and **D** **(82)**

implying

k, **E,** and **D** will always be in the same plane

Further [see Eq. (76)]

D is at right angles to **k** **(83)**

Substituting for **H** in Eq. (81), we get

$$\mathbf{D} = \frac{1}{\omega^2 \mu_0}[(\mathbf{k} \cdot \mathbf{k})\mathbf{E} - (\mathbf{k} \cdot \mathbf{E})\mathbf{k}] \qquad \textbf{(84)}$$

where we have used the vector identity

$$(\mathbf{A} \times \mathbf{B}) \times \mathbf{C} = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{B} \cdot \mathbf{C})\mathbf{A}$$

Thus

$$\begin{aligned}\mathbf{D} &= \frac{k^2}{\omega^2 \mu_0}[\mathbf{E} - (\hat{\boldsymbol{\kappa}} \cdot \mathbf{E})\hat{\boldsymbol{\kappa}}] \\ &= \frac{n_w^2}{c^2 \mu_0}[\mathbf{E} - (\hat{\boldsymbol{\kappa}} \cdot \mathbf{E})\hat{\boldsymbol{\kappa}}] \qquad \textbf{(85)}\end{aligned}$$

where

$$\hat{\boldsymbol{\kappa}} = \frac{\mathbf{k}}{k} \qquad \textbf{(86)}$$

represents the unit vector along **k** (see Fig. 22.32). Since

$$D_x = \varepsilon_x E_x = \varepsilon_0 n_x^2 E_x$$

we have for the *x* component of Eq. (85)

$$\frac{\varepsilon_0 \mu_0 c^2 n_x^2}{n_w^2} E_x = E_x - \kappa_x(\kappa_x E_x + \kappa_y E_y + \kappa_z E_z)$$

Since $c^2 = 1/(\varepsilon_0 \mu_0)$, we have

$$\left(\frac{n_x^2}{n_w^2} - \kappa_y^2 - \kappa_z^2\right) E_x + \kappa_x \kappa_y E_y + \kappa_x \kappa_z E_z = 0 \qquad \textbf{(87)}$$

where we have used the relation $\kappa_x^2 + \kappa_y^2 + \kappa_z^2 = 1$ (since $\hat{\boldsymbol{\kappa}}$ is a unit vector). Similarly,

$$\kappa_x \kappa_y E_x + \left(\frac{n_y^2}{n_w^2} - \kappa_x^2 - \kappa_z^2\right) E_y + \kappa_y \kappa_z E_z = 0 \qquad \textbf{(88)}$$

$$\kappa_x \kappa_z E_x + \kappa_y \kappa_z E_y + \left(\frac{n_z^2}{n_w^2} - \kappa_x^2 - \kappa_y^2\right) E_z = 0 \qquad \textbf{(89)}$$

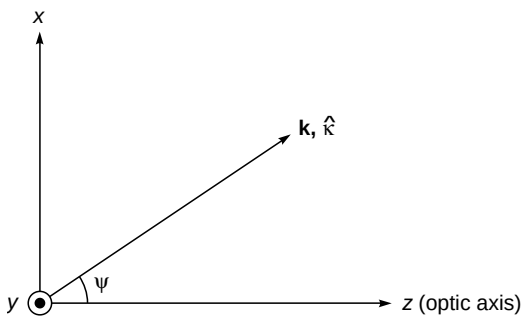


Fig. 22.32 In uniaxial crystals, we can always choose the *y* axis in such a way that $\kappa_y = 0$; the optic axis is assumed to be in the *z* direction. If ψ is the angle that k makes with the optic axis, then $\kappa_x = \kappa \sin\psi$ and $\kappa_z = \kappa \cos\psi$.

Since the above equations form a set of three homogeneous equations, for nontrivial solutions, we must have

$$\begin{vmatrix} \frac{n_x^2}{n_w^2} - \kappa_y^2 - \kappa_z^2 & \kappa_x \kappa_y & \kappa_x \kappa_z \\ \kappa_x \kappa_y & \frac{n_y^2}{n_w^2} - \kappa_x^2 - \kappa_z^2 & \kappa_y \kappa_z \\ \kappa_x \kappa_z & \kappa_y \kappa_z & \frac{n_z^2}{n_w^2} - \kappa_x^2 - \kappa_y^2 \end{vmatrix} = 0 \qquad \textbf{(90)}$$

We should remember that we still do not know the possible values of n_w. Indeed, for a given direction of propagation (i.e., for given values of κ_x, κ_y, and κ_z) the solution of Eq. (90) gives us the two allowed values of n_w. From Eq. (90) it appears as if we will have a cubic equation in n_w^2 which would give us three roots of n_w^2; however, the coefficient of n_w^6 will always be zero and hence there will be *always* two roots. We illustrate the general procedure by considering propagation through a uniaxial medium.

22.12.1 Propagation in Uniaxial Crystals

In this section, we completely restrict ourselves to uniaxial crystals for which

$$n_x = n_y = n_o \qquad \text{and} \qquad n_z = n_e \qquad \textbf{(91)}$$

As discussed earlier, for a uniaxial crystal, the x and y directions can be arbitrarily chosen as long as they are perpendicular to the optic axis. Now, for a wave propagating along *any* direction **k**, we choose our y axis in such a way that it is at right angles to **k**; i.e., the y axis is normal to the plane defined by **k** and the z axis; obviously, the x axis will lie in the same plane (see Fig. 22.32). Thus we may write

$$\kappa_x = \sin\psi \qquad \kappa_y = 0 \qquad \kappa_z = \cos\psi$$

where ψ is the angle that the **k** vector makes with the optic axis (see Fig. 22.32). Equations (87) to (89) therefore become

$$\left(\frac{n_o^2}{n_w^2} - \cos^2\psi\right) E_x + \sin\psi\cos\psi\, E_z = 0 \qquad \textbf{(92)}$$

$$\left(\frac{n_o^2}{n_w^2} - 1\right) E_y = 0 \qquad \textbf{(93)}$$

and

$$\sin\psi\cos\psi\, E_x + \left(\frac{n_e^2}{n_w^2} - \sin^2\psi\right) E_z = 0 \qquad \textbf{(94)}$$

Once again we have a set of three homogeneous equations, and for nontrivial solutions the determinant must be zero. However, since two equations involve only E_x and E_z and one equation involves only E_y, we have the following two independent solutions:

First Solution: We assume $E_y \neq 0$; then $E_x = 0 = E_z$. From Eq. (93) we obtain the solution

$$n_w = n_{wo} = n_o \qquad \text{ordinary wave} \qquad \textbf{(95)}$$

The corresponding wave velocity is

$$v_w = v_{wo} = \frac{c}{n_o} \qquad y\text{-polarized } o\text{-wave} \qquad \textbf{(96)}$$

Since the wave velocity is independent of the direction of the wave, it is referred to as the ordinary wave (usually abbreviated as the *o*-wave) and hence the subscript *o* on n_w and v_w. Further, for the *o*-wave, the **D** vector (and the **E** vector) is y-polarized. Thus, *for the o-wave, the* **D** *vector* (*and the* **E** *vector*) *are perpendicular to the plane containing the* **k** *vector and the optic axis* (see Fig. 22.33). This was the recipe that was given through Eq. (43).

Second Solution: The second solution of Eqs. (92) to (94) will correspond to

$$E_y = 0 \qquad E_x, E_z \neq 0 \qquad \textbf{(97)}$$

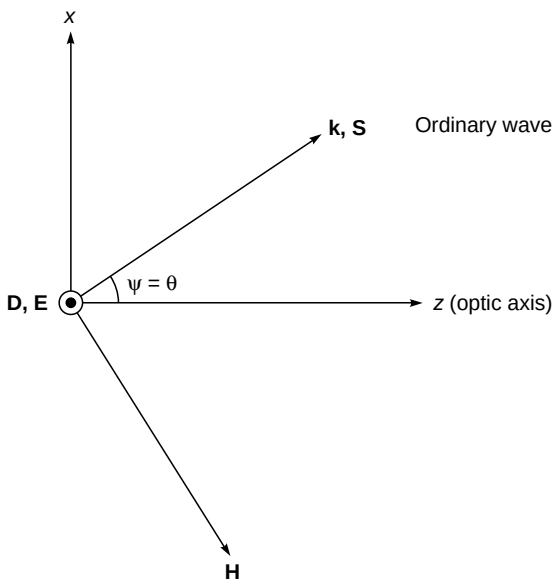


Fig. 22.33 For the ordinary wave (in uniaxial crystals), the **D** and **E** vectors are in the y direction; **k** and **S** are in the same direction in the xz plane, and **H** also lies in the xz plane.

We use Eqs. (92) to (94) to obtain

$$\frac{E_z}{E_x} = -\frac{n_o^2/n_w^2 - \cos^2\psi}{\sin\psi\cos\psi} = -\frac{\sin\psi\cos\psi}{n_e^2/n_w^2 - \sin^2\psi}$$

Simple manipulations give

$$\frac{1}{n_w^2} = \frac{1}{n_{we}^2} = \frac{\cos^2\psi}{n_o^2} + \frac{\sin^2\psi}{n_e^2} \qquad \textbf{(98)}$$

where the subscript *e* refers to the fact that the wave refractive index corresponds to the extraordinary wave. The corresponding wave velocity is given by

$$v_{we}^2 = \frac{c^2}{n_{we}^2} = \frac{c^2}{n_o^2}\cos^2\psi + \frac{c^2}{n_e^2}\sin^2\psi \qquad \textbf{(99)}$$

Since the wave velocity is dependent on the direction of the wave, it is referred to as the extraordinary wave and hence the subscript *e*. Of course, for the extraordinary wave, we must have

$$D_y = \varepsilon_y E_y = 0$$

From the above equation and Eq. (81), it follows that the displacement vector **D** of the wave is normal to the *y* axis and also to **k**, implying that *the displacement vector* **D** *associated with the extraordinary wave lies in the plane containing the propagation vector* **k** *and the optic axis and is normal to* **k** (see Fig. 22.34). This was the recipe given through Eq. (44). Figure 22.34 also shows the Poynting vector **S** (= **E** × **H**) which represents the direction of energy propagation (i.e., the direction of the *e*-ray). The small dashes on the extraordinary ray in Fig. 22.21(a) and (b) represent the directions of the **D** vector. Let ϕ and θ represent the angles that the **S** vector makes with the **k** vector and the optic axis, respectively (see Fig. 22.33). To determine the angle ϕ, we note that

$$\frac{\varepsilon_z E_z}{\varepsilon_x E_x} = \frac{D_z}{D_x} = -\tan\psi$$

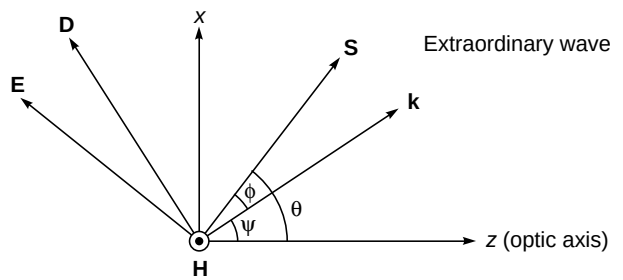


Fig. 22.34 For the extraordinary wave (in uniaxial crystals), **E**, **D**, **S**, and **k** vectors lie in the *xz* plane and **H** will be in the *y* direction. **S** is at right angles to **E** and **H**; **D** is at right angles to **k** and **H**.

and since

$$\frac{E_z}{E_x} = -\tan(\phi + \psi) \qquad \textbf{(100)}$$

we get

$$\frac{n_e^2}{n_o^2}\tan(\phi + \psi) = \tan\psi$$

or

$$\phi = \tan^{-1}\left(\frac{n_o^2}{n_e^2}\tan\psi\right) - \psi$$

Obviously, for negative crystals $n_o > n_e$ and ϕ will be positive, implying that ray direction is further away from the optic axis as shown in Fig. 22.34.

Conversely, for positive crystals $n_o < n_e$ and ϕ will be negative, implying that the ray direction will be toward the optic axis.

Example 22.3 We consider calcite for which (at λ = 5893 Å and 18ºC)

$$n_o = 1.65835 \qquad n_e = 1.48640$$

If we consider **k** making an angle of 30º to the optic axis, then ψ = 30° and elementary calculations give ϕ = 5.7°.

22.13 RAY VELOCITY AND RAY REFRACTIVE INDEX

The direction of energy propagation (or the ray propagation) is along the Poynting vector **S** which is given by

$$\mathbf{S} = \mathbf{E} \times \mathbf{H} \qquad \textbf{(101)}$$

Thus, since the plane containing the vectors **k**, **E**, and **D** is normal to **H**, the Poynting vector **S** will also lie in the plane containing the vectors **k**, **E**, and **D** (see Figs. 22.33 and 22.34). For the extraordinary wave, the direction of the propagation of the wave $\hat{\boldsymbol{\kappa}}$ is not along the direction of energy propagation $\hat{\mathbf{s}}$, where $\hat{\mathbf{s}}$ is the unit vector along **S**. The ray velocity (or the energy transmission velocity) v_r is defined as

$$v_r = \frac{S}{u} \qquad \textbf{(102)}$$

where u is the energy density. Now,

$$u = \frac{1}{2}(\mathbf{D}\cdot\mathbf{E} + \mathbf{B}\cdot\mathbf{H})$$

$$= \frac{1}{2}(\mathbf{D}\cdot\mathbf{E} + \mu_0\,\mathbf{H}\cdot\mathbf{H}) \qquad \textbf{(103)}$$

[see Eq. (58) of Chap. 23]. Substituting for **H** and **D** from Eqs. (80) and (81), we obtain

$$u = \frac{1}{2\omega}[(\mathbf{H}\times\mathbf{k})\cdot\mathbf{E} + (\mathbf{k}\times\mathbf{E})\cdot\mathbf{H}]$$

$$= \frac{1}{2\omega}[\mathbf{k}\cdot(\mathbf{E}\times\mathbf{H}) + \mathbf{k}\cdot(\mathbf{E}\times\mathbf{H})]$$

$$= \frac{1}{\omega}\,\mathbf{k}\cdot\mathbf{S} \qquad \textbf{(104)}$$

Thus Eq. (102) becomes

$$v_r = \frac{\omega S}{\mathbf{k}\cdot\mathbf{S}} = \frac{\omega}{k\cos\phi} = \frac{v_w}{\cos\phi} \qquad \textbf{(105)}$$

where ϕ is the angle between $\hat{\boldsymbol{\kappa}}$ and $\hat{\mathbf{s}}$ (see Fig. 22.34). The ray refractive index n_r is defined as

$$n_r = \frac{c}{v_r} = \frac{c}{v_w}\cos\phi = n_w\cos\phi \qquad \textbf{(106)}$$

To express **E** in terms of **D**, we refer to Fig. 22.34 and write

$$\mathbf{D} = (\mathbf{D}\cdot\hat{\mathbf{e}})\,\hat{\mathbf{e}} + (\mathbf{D}\cdot\hat{\mathbf{s}})\,\hat{\mathbf{s}}$$

where $\hat{\mathbf{e}}$ is a unit vector along the direction of the electric field **E**. Thus

$$\mathbf{D} - (\mathbf{D}\cdot\hat{\mathbf{s}})\,\hat{\mathbf{s}} = (\mathbf{D}\cdot\hat{\mathbf{e}})\,\hat{\mathbf{e}} = (D\cos\phi)\,\frac{\mathbf{E}}{E} \qquad \textbf{(107)}$$

Similarly,

$$\mathbf{E} = (\mathbf{E}\cdot\hat{\mathbf{d}})\,\hat{\mathbf{d}} + (\mathbf{E}\cdot\hat{\boldsymbol{\kappa}})\,\hat{\boldsymbol{\kappa}} \qquad \textbf{(108)}$$

where $\hat{\mathbf{d}}$ represents a unit vector along the displacement vector **D** (see Fig. 22.34). If we now substitute for $\mathbf{E} - (\mathbf{E}\cdot\hat{\boldsymbol{\kappa}})\,\hat{\boldsymbol{\kappa}}$ in Eq. (85), we get

$$\mathbf{D} = \frac{n_w^2}{\mu_0 c^2}(\mathbf{E}\cdot\hat{\mathbf{d}})\,\hat{\mathbf{d}}$$

or

$$D = \frac{n_w^2}{\mu_0 c^2}E\cos\phi \qquad \textbf{(109)}$$

Substituting in Eq. (107), we get

$$\mathbf{D} - (\mathbf{D}\cdot\hat{\mathbf{s}})\,\hat{\mathbf{s}} = \frac{n_w^2}{\mu_0 c^2}\cos^2\phi\,\mathbf{E} = \frac{n_r^2}{\mu_0 c^2}\mathbf{E}$$

where, in the last step, we used Eq. (106). Taking the *x* component of the above equation (where *x* represents the direction of one of the principal axes), we obtain

$$D_x - (D_x s_x + D_y s_y + D_z s_z)\,s_x = \frac{n_r^2}{\mu_0 c^2}E_x = \frac{n_r^2}{\mu_0 c^2 \varepsilon_x}D_x$$

If we use the relations

$$n_x^2 = \frac{\varepsilon_x}{\varepsilon_0} \qquad c^2 = \frac{1}{\varepsilon_0\mu_0} \qquad \text{and} \qquad s_x^2 + s_y^2 + s_z^2 = 1$$

we get

$$\left(\frac{n_r^2}{n_x^2} - s_y^2 - s_z^2\right)D_x + s_x s_y D_y + s_x s_z D_z = 0 \qquad \textbf{(110)}$$

Similarly,

$$s_x s_y D_x + \left(\frac{n_r^2}{n_y^2} - s_x^2 - s_z^2\right)D_y + s_z s_y D_z = 0 \qquad \textbf{(111)}$$

$$s_x s_z D_x + s_z s_y D_y + \left(\frac{n_r^2}{n_z^2} - s_x^2 - s_y^2\right)D_z = 0 \qquad \textbf{(112)}$$

As in the previous section, the above set of equations forms a set of three homogeneous equations. For nontrivial solutions, we must have

$$\begin{vmatrix} \dfrac{n_r^2}{n_x^2} - s_y^2 - s_z^2 & s_x s_y & s_x s_z \\ s_x s_y & \dfrac{n_r^2}{n_y^2} - s_x^2 - s_z^2 & s_z s_y \\ s_x s_z & s_z s_y & \dfrac{n_r^2}{n_z^2} - s_x^2 - s_y^2 \end{vmatrix} = 0 \qquad \textbf{(113)}$$

We still do not know the possible values of n_r. Indeed for a given ray direction (i.e., for given values of s_x, s_y, and s_z) the solution of the above equation gives the two allowed values of n_r and hence two possible values of the ray velocities. We illustrate this by considering propagation through uniaxial media.

22.13.1 Ray Propagation in Uniaxial Crystals

We next consider a uniaxial crystal with its optic axis along the *z* direction. Thus

$$n_x = n_y = n_o \qquad \text{and} \qquad n_z = n_e \qquad \textbf{(114)}$$

As discussed in the previous section, *x* and *y* directions can be arbitrarily chosen as long as they are perpendicular to the *z* axis. We choose the *y* axis in such a way that the ray propagates in the *xz* plane, making an angle θ with the *z* axis (see Fig. 22.34); thus

$$s_x = \sin\theta \qquad s_y = 0 \quad \text{and} \qquad s_z = \cos\theta \qquad \textbf{(115)}$$

and Eqs. (110) to (112) become

$$\left(\frac{n_r^2}{n_o^2} - \cos^2\theta\right)D_x + \sin\theta\cos\theta\,D_z = 0 \qquad \textbf{(116)}$$

$$\left(\frac{n_r^2}{n_o^2} - 1\right) D_y = 0 \qquad \textbf{(117)}$$

$$\sin\theta \cos\theta\, D_x + \left(\frac{n_r^2}{n_e^2} - \sin^2\theta\right) D_z = 0 \qquad \textbf{(118)}$$

Obviously, one of the roots is given by

$$n_r = n_{ro} = n_o$$

with $D_x = 0 = D_z$ y-polarized **(119)**

The corresponding ray velocity is given by

$$v_r = v_{ro} = \frac{c}{n_{ro}} = \frac{c}{n_o} \qquad \text{ordinary ray} \qquad \textbf{(120)}$$

Since the ray velocity is independent of the direction of the ray, it is referred to as the ordinary ray and hence the subscript o on v_r and n_r.

To obtain the other solution, we use Eqs. (116) and (118) to get

$$\frac{D_z}{D_x} = -\frac{n_r^2/n_o^2 - \cos^2\theta}{\sin\theta\cos\theta} = -\frac{\sin\theta\cos\theta}{n_r^2/n_e^2 - \sin^2\theta}$$

and obviously,

$$D_y = 0$$

Simple manipulations give

$$n_r^2 = n_{re}^2 = n_o^2 \cos^2\theta + n_e^2 \sin^2\theta$$

extraordinary ray **(121)**

with

$$\frac{D_z/n_e^2}{D_x/n_o^2} = \frac{E_z}{E_x} = -\tan\theta \qquad (D_y = 0) \qquad \textbf{(122)}$$

The corresponding ray velocity is given by [cf. Eq. (37)]

$$\frac{1}{v_r^2} = \frac{1}{v_{re}^2} = \frac{n_{re}^2}{c^2} = \frac{\cos^2\theta}{c^2/n_o^2} + \frac{\sin^2\theta}{c^2/n_e^2} \qquad \textbf{(123)}$$

which corresponds to the extraordinary ray and hence the subscript e on v_r and n_r. As discussed in Sec. 22.5, the above equation represents an ellipse, and if we rotate it about the z axis (i.e., the optic axis), we get an ellipsoid of revolution. These *ray velocity surfaces* are used in constructing Huygens' secondary wavelets while discussing propagation in uniaxial crystals. For example, in Fig. 22.21(a) we have a plane wave incident normally. The extraordinary wave also propagates in a direction which is normal to the surface. However, the extraordinary rays travel in the directions BE and DE' with EE' representing the wave front for the extraordinary wave. Returning to Eq. (120), we obtain (see Fig. 22.34)

$$\tan\theta = -\frac{D_z/n_e^2}{D_x/n_o^2} = \frac{n_o^2}{n_e^2}\tan\psi \qquad \textbf{(124)}$$

Thus when the wave propagates along a direction which makes an angle ψ with the optic axis, the ray will propagate along the direction

$$\theta = \tan^{-1}\left[\frac{n_o^2}{n_e^2}\tan\psi\right] \qquad \textbf{(125)}$$

As an example, for calcite

$$n_o = 1.65836 \qquad n_e = 1.48641 \qquad \text{with } \psi = 30°$$

we obtain $\theta \approx 35.7°$. Thus the ray direction is farther away from the optic axis, consistent with what is shown in Fig. 22.21. Note that $\theta = \phi + \psi$ (see Example 22.3).

22.14 JONES' CALCULUS

Through Jones' calculus, it becomes quite straightforward to determine the polarization state of the beam emerging from a polarizer or a phase retarder (such as a QWP or a HWP). We illustrate this through some simple examples. We use exponential notation; for example, a y-polarized beam (propagating in the x direction) is described by

$$\mathbf{E} = \hat{\mathbf{y}}\, E_0 \cos(kx - \omega t) = \hat{\mathbf{y}}\ \text{Re}\,(E_0 e^{i(kx-\omega t)}) \qquad \textbf{(126)}$$

Such a wave is represented by the vector

$$|\, y > = \begin{pmatrix} 1 \\ 0 \end{pmatrix} E_0 \qquad \textbf{(127)}$$

Similarly, a z-polarized wave is given by

$$|\, z > = \begin{pmatrix} 0 \\ 1 \end{pmatrix} E_0 \qquad \textbf{(128)}$$

Now, for an RCP (propagating in the x direction) we may write

$$E_y = E_0 \cos(kx - \omega t)$$

$$E_z = -E_0 \sin(kx - \omega t)$$

which in exponential notation can be written as

$$\mathbf{E} = \hat{\mathbf{y}}\, e^{i(kx - \omega t)} + \hat{\mathbf{z}}\, e^{i(kx - \omega t + \pi/2)}$$

Thus neglecting the phase factor,

$$|\,\text{RCP}> = \begin{pmatrix} 1 \\ e^{i\pi/2} \end{pmatrix} E_0 = \begin{pmatrix} 1 \\ i \end{pmatrix} E_0 \qquad \textbf{(129)}$$

Similarly

$$|\,\text{LCP}> = \begin{pmatrix} 1 \\ -i \end{pmatrix} E_0 \qquad \textbf{(130)}$$

Let us next consider a phase retarder such as a QWP or a HWP or even an elliptic core fiber. As discussed in earlier sections, the modes of such a device are linearly polarized along the fast and slow axes, as shown in Fig. 22.24. The electric fields along these directions are denoted by E_f and E_s; the subscripts *f* and *s* denote the fast and slow axes, respectively. As an example, we consider a calcite QWP for which $n_o > n_e$. The extraordinary wave is *z*-polarized (i.e., along the optic axis), and its velocity $(= c/n_e)$ is more than the velocity of the *o*-wave $(= c/n_e)$. Thus (for calcite) the *fast axis* is along the *z* direction, and the *slow axis* is along the *y* direction as shown in Fig. 22.24.

The slow and fast components are the *modes* of the retardation plate; i.e., after propagating through the retardation plates (of thickness *d*), the fields are given by

$$E_s' = e^{ik_s d}E_s$$
$$E_f' = e^{ik_f d}E_f$$

where

$$k_s = \frac{2\pi}{\lambda_0}n_s \qquad \text{and} \qquad k_f = \frac{2\pi}{\lambda_0}\,n_f$$

(For calcite $n_s = n_o = 1.65836$ and $n_f = n_e = 1.48641$ at $\lambda_0 =$ 5893 Å.) Since only the relative phase difference is of interest, we may write

$$E_s' = e^{i\Phi}E_s$$
$$E_f' = E_f$$

where

$$\Phi = \frac{2\pi}{\lambda_0}(n_s - n_f)d$$

is the phase difference introduced by the phase retarder. The calcite plate is therefore represented by the following matrix:

$$\begin{pmatrix} e^{i\Phi} & 0 \\ 0 & 1 \end{pmatrix}$$

Thus we may write

$$\begin{pmatrix} E_s' \\ E_f' \end{pmatrix} = \begin{pmatrix} e^{i\Phi} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} E_s \\ E_f \end{pmatrix}$$

As mentioned earlier, *y* and *z* axes are the slow and fast axes, respectively. Thus

$$\begin{pmatrix} E_y' \\ E_z' \end{pmatrix} = \begin{pmatrix} e^{i\Phi} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} E_y \\ E_z \end{pmatrix}$$

For a *y*-polarized wave

$$E_y = E_0 \qquad E_z = 0$$

and a *y*-polarized wave will remain *y*-polarized (except for change of phase). Similarly, a *z*-polarized wave will remain *z*-polarized—these are, of course, the *modes* of the phase retarder. Now, for a QWP, $\Phi = \pi/2$, giving

$$\begin{pmatrix} E_y' \\ E_z' \end{pmatrix} = \begin{pmatrix} iE_y \\ E_z \end{pmatrix}$$

For a linearly polarized beam with **E** at π/4 with the *y* and *z* axes,

$$E_y = \frac{1}{\sqrt{2}}E_0 \qquad \text{and} \qquad E_z = \frac{1}{\sqrt{2}}E_0$$

Thus

$$\begin{pmatrix} E_y' \\ E_z' \end{pmatrix} = \begin{pmatrix} i \\ 1 \end{pmatrix}\frac{E_0}{\sqrt{2}}$$

which is a LCP. If this LCP is incident on a similarly oriented calcite QWP, the output beam will be

$$\begin{pmatrix} E_y'' \\ E_z'' \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} i \\ 1 \end{pmatrix}\frac{E_0}{\sqrt{2}} = \frac{E_0}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

which represents a LP beam with its **E** oscillating at right angles to the direction of the incident polarization. Note that two QWPs put together makes a HWP.

Let us next consider the incidence of a LEP given by

$$E_y = \frac{E_0}{2}\cos(kx - \omega t)$$
$$E_z = \frac{\sqrt{3}}{2}E_0\sin(kx - \omega t)$$
$$= \frac{\sqrt{3}}{2}E_0\,\mathrm{Re}\,e^{i\left(kx-\omega t-\frac{\pi}{2}\right)}$$

Thus the beam coming out of the QWP is given by

$$\begin{pmatrix} E_y' \\ E_z' \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & 1 \end{pmatrix}E_0\begin{pmatrix} \frac{1}{2} \\ -i\sqrt{3}/2 \end{pmatrix}$$

which is LP with its **E** making an angle of 120° with the *z* axis. Thus the Jones matrices for a QWP would be

$$\begin{pmatrix} i & 0 \\ 0 & 1 \end{pmatrix}$$

when the fast axis is horizontal (see Fig. 22.24) and

$$\begin{pmatrix} -i & 0 \\ 0 & 1 \end{pmatrix}$$

when the slow axis is horizontal. Similarly,

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

will represent the Jones matrix for the *y* polarizer and *z* polarizer respectively. It is left as an exercise to show that the matrix

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

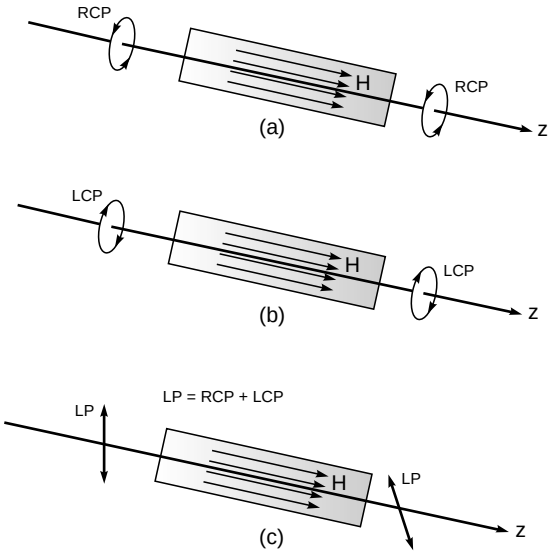


Fig. 22.35 An electromagnetic wave is propagating through a dielectric. If we apply a static magnet field along the direction of propagation of the wave, the modes are now Right Circularly Polarized (RCP) and Left Circularly Polarized (LCP). (a) Thus a right circularly polarized wave will propagate as a right circularly polarized wave with a particular velocity, and (b) a left circularly polarized wave will also propagate as a left circularly polarized wave but with a slightly different velocity. (c) If a linearly polarized wave is incident, the direction of the electric vector will get rotated.

will correspond to a polarization rotator; thus if $E_y/E_z = \tan\alpha$ then $E'_y/E'_z = \tan(\alpha + \theta)$.

The use of Jones' matrices makes it very straightforward to consider more complicated cases such as two QWPs, with their axes at an angle.

22.15 FARADAY ROTATION

Consider a linearly polarized light propagating through a medium. If a magnetic field is applied along the direction of propagation of the polarized wave, then the plane of polarization gets rotated—this rotation is usually referred to as *Faraday rotation* after the famous physicist Michael Faraday who discovered this phenomenon in 1845. In the presence of a (longitudinal) magnetic field, the modes of propagation are the left circularly polarized (LCP) wave and the right circularly polarized (RCP) wave (see Fig. 22.35 and Sec. 22.17). Thus the situation is somewhat similar to the phenomenon of optical activity discussed in Sec. 22.8. The angle θ by which the plane of polarization rotates is given by the empirical formula

$$\theta = VHl$$

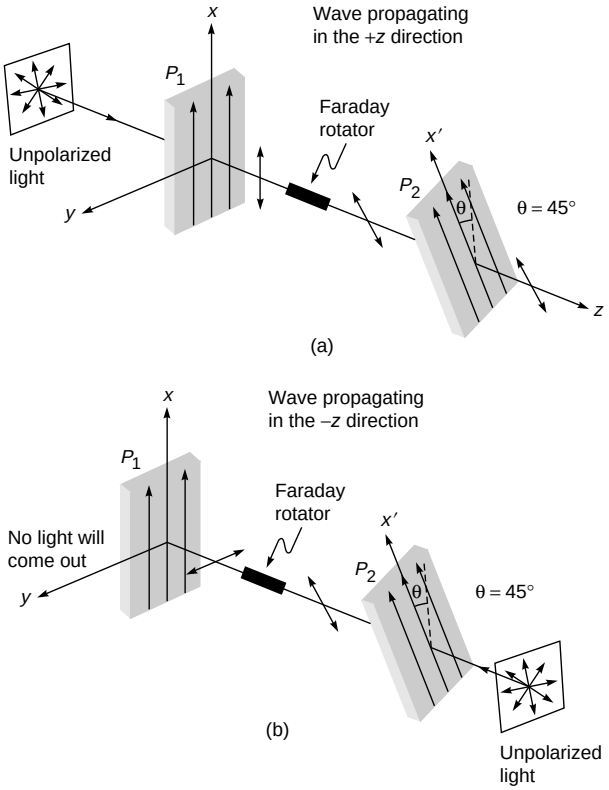


Fig. 22.36 P_1 and P_2 are two linear polarizers with pass axes at 45° to each other. (a) The light beam incident from the left gets polarized along the x direction. The x-polarized light passes through the Faraday rotator which rotates the state of polarization by 45°, which is along the pass axis of the second polarizer P_2; thus the light passes through. (b) For a light beam incident from the right (with arbitrary state of polarization), it will get polarized along the x' direction. The x'-polarized light passes through the Faraday rotator which will rotate the state of polarization by 45°. Thus the beam coming out of the Faraday rotator is polarized along the y direction and is perpendicular to the pass axis of the second polarizer P_1—thus no light will pass through P_1.

where H is the applied magnetic field, l is the length of the medium, and V is called the Verdet constant. For silica $V \simeq 2.64 \times 10^{-4}$ deg/A $\simeq 4.6 \times 10^{-6}$ rad/A.

22.15.1 THE FARADAY ISOLATOR

One of the very important applications of Faraday rotation is in the construction of the device known as the Faraday isolator [see Figs. 22.36(a) and (b)]. Faraday isolators allow light to pass through only in one direction and are extensively used to avoid optical feedback. In Fig. 22.36, P_1 and P_2 are

two linear polarizers with pass axes at 45° to each other. The Faraday rotator is chosen to give a 45° rotation. The light beam incident from the left gets polarized along the x direction. The x-polarized light passes through the Faraday rotator which rotates the state of polarization by 45°. Thus the beam coming out of the Faraday rotator is polarized along the x' direction and is along the pass axis of the second polarizer P_2 [see Fig. 22.36(a)]. Thus the light passes through, and for a good isolator the transmission can be very high. Now, if an unpolarized light beam is incident from the right, it will get polarized along the x' direction. The x'-polarized light passes through the Faraday rotator which will further rotate the state of polarization by 45°. Thus the beam coming out of the Faraday rotator is polarized along the y direction and is perpendicular to the pass axis of the second polarizer P_1—thus no light will pass through P_1 [see Fig. 22.36(b)]. Note that if the magnetic field is along the $+z$ direction, then for the wave propagating along the $+z$ direction [see Fig. 22.36(a)] the rotation will be in the clockwise direction. On the other hand, for the wave propagating along the $-z$ direction [see Fig. 22.36(b)], the magnetic field is opposite to the direction of propagation and the Faraday rotation will be in the counterclockwise direction.

In the wavelength region 0.7 to 1.1 μm one often uses terbium-doped borosilicate glass. Faraday isolators are extensively used in many fiber-optic devices, and in the wavelength range of 1.3 to 1.55 μm (which is the wavelength range of interest in fiber-optic communication systems) one often uses YIG (yttrium iron garnet) crystals.

22.15.2 LARGE CURRENT MEASUREMENT USING FARADAY ROTATION

The Faraday rotation has a very important application in measuring large currents using single-mode optical fibers. We consider a large length of a single-mode fiber wound in many turns in the form of a loop around a current-carrying conductor (see Fig. 22.37 and the corresponding figure (Fig. 45) in the insert at the back of the book. If a current I is passing through the conductor, then by Ampere's law

$$\int \mathbf{H} \cdot d\mathbf{l} = NI$$

where N represents the number of loops of the fiber around the conductor. Thus if a linearly polarized light is incident on the fiber, then its plane of polarization will get rotated by the angle

$$\theta = VNI$$

The rotation θ does not depend on the shape of the loop. As an example, for I = 200 A and N = 50, θ ≈ 0.26°. The light

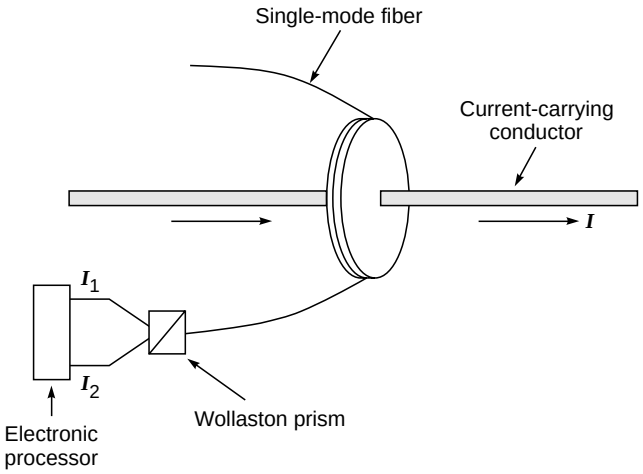


Fig. 22.37 A single-mode fiber wound helically around a current-carrying conductor. The rotation of the plane of polarization is detected by passing the light through a Wollaston prism and then an electronic processor.

from the fiber is allowed to fall on a Wollaston prism, and the outputs are measured separately; the Faraday rotation θ is given by

$$\theta = \text{constant} \frac{I_1 - I_2}{I_1 + I_2}$$

where I_1, and I_2 are the currents in the electronic processor due to the two beams coming out of the Wollaston prism. Figure 22.38 shows an actual variation of the output with the current passing through the conductor. Such a setup can be used to measure very high currents (~ 10,000 A).

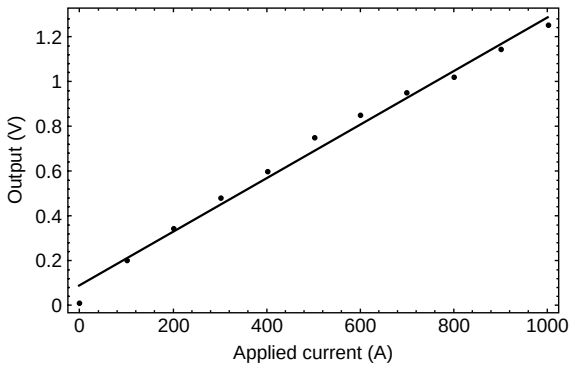


Fig. 22.38 A typical variation of the output signal with current [Figure kindly provided by Dr. Partha-sarathi Palai].

22.16 THEORY OF OPTICAL ACTIVITY

As mentioned earlier, in an isotropic dielectric, the **D** vector is in the same direction as **E** and we have

$$\mathbf{D} = \varepsilon\mathbf{E} = \varepsilon_0 n^2 \mathbf{E} \tag{131}$$

where ε_0 (= 8.854×10^{-12} m k s units) is the permittivity of free space and $n\left(=\sqrt{\varepsilon/\varepsilon_0}\right)$ is the refractive index of the medium. Now, if we dissolve cane sugar in water, the medium is still isotropic; however, because of the spiral-like structure of sugar molecules, the relation between **D** and **E** is given by

$$\begin{aligned}\mathbf{D} &= \varepsilon_0 n^2 \mathbf{E} + ig\hat{\boldsymbol{\kappa}} \times \mathbf{E} \\ &= \varepsilon_0 n^2 [\mathbf{E} + i\alpha\hat{\boldsymbol{\kappa}} \times \mathbf{E}]\end{aligned} \tag{132}$$

where

$$\alpha = \frac{g}{\varepsilon_0 n^2}$$

and $\hat{\boldsymbol{\kappa}}$ is the unit vector along the direction of propagation of the wave. The parameter α can be positive or negative, but it is usually an extremely small number (<< 1). Without any loss of generality, we may assume propagation along the z axis so that $\kappa_x = \kappa_y = 0$ and $\kappa_z = 1$, giving

$$\hat{\boldsymbol{\kappa}} \times \mathbf{E} = \begin{pmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ 0 & 0 & 1 \\ E_x & E_y & E_z \end{pmatrix} = -\hat{\mathbf{x}} E_y + \hat{\mathbf{y}} E_x$$

Thus

$$\begin{pmatrix} D_x \\ D_y \\ D_z \end{pmatrix} = \begin{pmatrix} \varepsilon_0 n^2 & -ig & 0 \\ ig & \varepsilon_0 n^2 & 0 \\ 0 & 0 & \varepsilon_0 n^2 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} \tag{133}$$

The ε matrix is still Hermitian, but there is a "small" off-diagonal imaginary element. The presence of these off-diagonal terms give rise to optical activity. We rewrite Eq. (85)

$$\frac{n_w^2}{c^2\mu_0}[\boldsymbol{E} - (\hat{\boldsymbol{\kappa}} \cdot \boldsymbol{E})\hat{\boldsymbol{\kappa}}] = \mathbf{D}$$

We write the x and y components of the above equation, and since $\kappa_x = 0 = \kappa_y$ and $\kappa_z = 1$, we get

$$\frac{n_w^2}{c^2\mu_0} E_x = D_x = \varepsilon_0 n^2 E_x - igE_y$$

and

$$\frac{n_w^2}{c^2\mu_0} E_y = D_y = igE_x + \varepsilon_0 n^2 E_y$$

Thus

$$\left(\frac{n_w^2}{n^2} - 1\right) E_x = -i\alpha E_y$$

and

$$\left(\frac{n_w^2}{n^2} - 1\right) E_y = i\alpha E_x$$

where we used the fact that $c = 1/\sqrt{\varepsilon_0\mu_0}$. For nontrival solutions

$$\left(\frac{n_w^2}{n^2} - 1\right)^2 = \alpha^2$$

giving

$$n_w = n\sqrt{1 \pm \alpha} \tag{134}$$

and

$$E_y = \pm iE_x \tag{135}$$

We write the two solutions as $n_r \left(= n\sqrt{1+\alpha}\right)$ and $n_l \left(= n\sqrt{1-\alpha}\right)$; the corresponding propagation constants are given by

$$k = k_r = \frac{\omega}{c} n_r = \frac{\omega}{c} n\sqrt{1+\alpha} \tag{136}$$

and

$$k = k_l = \frac{\omega}{c} n_l = \frac{\omega}{c} n\sqrt{1-\alpha} \tag{137}$$

For $n_w = n_r$, if

$$E_x = E_0 e^{i(k_r z - \omega t)}$$

then

$$E_y = +iE_x = E_0 e^{i(k_r z - \omega t + \pi/2)}$$

which represents a RCP (right circularly polarized) wave and hence the subscript r. Similarly, For $n_w = n_l$, if

$$E_x = E_0 e^{i(k_l z - \omega t)}$$

then

$$E_y = -iE_x = E_0 e^{i(k_l z - \omega t - \pi/2)}$$

which represents a LCP (left circularly polarized) wave and hence the subscript l. The RCP and LCP waves are the two modes of the optically active substance, and for an arbitrary incident state of polarization, we must write it as a superposition of the two modes and study the independent propagation of the two modes (see Sec. 22.8). Now,

$$\begin{aligned} n_r - n_l &= n\left(\sqrt{1+\alpha} - \sqrt{1-\alpha}\right) \\ &\approx n\alpha \end{aligned} \tag{138}$$

If d grams of pure cane sugar is dissolved in 100 g of water solution, then for λ = 5893 Å (sodium light)

$$n_r - n_l \approx 2.2 \times 10^{-6}\, d$$

Thus, if $d = 5$ g, then $n_r - n_l \approx 1.1 \times 10^{-5}$ and $\alpha \approx 0.83 \times 10^{-5}$ where we have assumed $n \approx 4/3$. Further, the angle of rotation is given by [see Eq. (58)]

$$\Phi = \frac{\pi}{\lambda_0}(n_l - n_r)z \tag{139}$$

The specific rotation ρ is defined as the angle through which the plane of polarization rotates in traversing a distance of 1 cm; thus

$$\rho = \frac{\pi}{\lambda_0}(n_l - n_r) \tag{140}$$

where λ_0 is measured in centimeters. For the sugar solution mentioned above (5 g dissolved in 100 g of water solution)

$$\rho \approx -0.59 \text{ rad cm}^{-1}$$

the negative sign indicating that the direction of polarization rotates in the counterclockwise direction.

22.16.1 Optical Activity in Quartz

One observes optical activity for a plane polarized wave propagating along the optic axis of a quartz crystal. The general theory of propagation of electromagnetic waves in such crystals is quite difficult; however, if the propagation is not along the optic axis, the modes are very nearly linearly polarized and one may use the analysis discussed in Sec. 22.12. If the propagation is along the z axis, then we may write [cf. Eq. (133)]

$$\begin{pmatrix} D_x \\ D_y \\ D_z \end{pmatrix} = \begin{pmatrix} \varepsilon_0 n_o^2 & -ig & 0 \\ ig & \varepsilon_0 n_o^2 & 0 \\ 0 & 0 & \varepsilon_0 n_e^2 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} \tag{141}$$

where n_o and n_e are constants of the crystal. Carrying out an identical analysis, we get

$$n_r \approx n_o \left(1 + \tfrac{1}{2}\alpha\right)$$

and

$$n_l \approx n_o \left(1 - \tfrac{1}{2}\alpha\right)$$

giving

$$n_r - n_l \approx n_o \alpha$$

and

$$\rho = \frac{\pi}{\lambda_0}(n_l - n_r) \approx \frac{\pi n_o \alpha}{\lambda_0}$$

where λ_0 is measured in centimeters. For quartz,

$$\rho \approx \pm 8.54 \text{ rad cm}^{-1} \qquad \text{at } \lambda_0 = 4046.56 \text{ Å}$$
$$\approx \pm 3.79 \text{ rad cm}^{-1} \qquad \text{at } \lambda_0 = 5892.90 \text{ Å}$$
$$\approx \pm 2.43 \text{ rad cm}^{-1} \qquad \text{at } \lambda_0 = 7281.35 \text{ Å}$$

(data adapted from Ref. 7). In quartz, we can have $n_r > n_l$ or $n_r < n_l$. For $\lambda_0 = 4046.56$ Å, we readily get

$$|n_l - n_r| \approx 1.1 \times 10^{-4}$$

We may compare this with the value of $n_e - n_o \approx 0.9 \times 10^{-2}$. At higher wavelengths, the value of $|n_r - n_l|$ is much less.

22.17 THEORY OF FARADAY ROTATION

As discussed in Sec. 7.5, the equation of motion for the electron, in the presence of an external electric field **E**, is given by [see Eq. (62) of Chap. 7]:

$$\frac{d^2\mathbf{r}}{dt^2} + \omega_0^2 \mathbf{r} = -\frac{q}{m}\mathbf{E} \tag{142}$$

In the presence of a static magnetic field **B**, we have an additional $\mathbf{v} \times \mathbf{B}$ term:

$$\frac{d^2\mathbf{r}}{dt^2} + \omega_0^2 \mathbf{r} + \frac{q}{m}\dot{\mathbf{r}} \times \mathbf{B} = -\frac{q}{m}\mathbf{E} \tag{143}$$

where $\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$ represents the position vector of the electron; $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are the unit vectors, and q (= $+1.6 \times 10^{-19}$ C) is the magnitude of the electronic charge. We assume the magnetic field to be in the z direction.

$$B_x = 0 = B_y \qquad \text{and} \qquad B_z = B_0 \tag{144}$$

Thus

$$\dot{\mathbf{r}} \times \mathbf{B} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \frac{dx}{dt} & \frac{dy}{dt} & \frac{dz}{dt} \\ 0 & 0 & B_0 \end{vmatrix} = \left(\hat{\mathbf{x}}\frac{dy}{dt} - \hat{\mathbf{y}}\frac{dx}{dt}\right)B_0 \tag{145}$$

Now, for a circularly polarized light wave propagating along the z direction

$$\mathbf{E}_{\pm} = (\hat{\mathbf{x}} \pm i\hat{\mathbf{y}})E_0 e^{i(kz - \omega t)} \tag{146}$$

where the upper and lower signs correspond to RCP and LCP, respectively. If we now write the x and y components of Eq. (143), we get

$$\frac{d^2x}{dt^2} + \omega_0^2 x + \frac{qB_0}{m}\frac{dy}{dt} = -\frac{q}{m}E_0 e^{i(kz - \omega t)} \tag{147}$$

and

$$\frac{d^2y}{dt^2} + \omega_0^2 y - \frac{qB_0}{m}\frac{dx}{dt} = \mp i\frac{q}{m}E_0 e^{i(kz - \omega t)} \tag{148}$$

where the upper and lower signs correspond to RCP and LCP, respectively. Writing

$$x = x_0 e^{i(kz - \omega t)} \qquad \text{and} \qquad y = y_0 e^{i(kz - \omega t)}$$

we get

$$(\omega^2 - \omega_0^2)\, x_0 + i\omega_c\omega\, y_0 = +\frac{q}{m}E_0 \quad \Big| \times(\omega^2 - \omega_0^2) \qquad \textbf{(149)}$$

$$(\omega^2 - \omega_0^2)\, y_0 - i\omega_c\omega\, x_0 = \pm i\frac{q}{m}E_0 \quad \Big| \times -i\omega_c\omega \qquad \textbf{(150)}$$

where

$$\omega_c = \frac{qB_0}{m}$$

is the electron cyclotron frequency. If we multiply Eq. (149) by $\omega^2 - \omega_0^2$ and Eq. (150) by $-i\omega_c\omega$ and add the two equations we get

$$[(\omega^2 - \omega_0^2)^2 - \omega_c^2\omega^2]x_0 = \frac{q}{m}E_0[(\omega^2 - \omega_0^2) \pm \omega_c\omega]$$

giving

$$x_0 = \frac{qE_0}{m[(\omega^2 - \omega_0^2) \mp \omega_c\omega]}$$

Similarly,

$$y_0 = \pm i\frac{qE_0}{m[(\omega^2 - \omega_0^2) \mp \omega_c\omega]} = \pm i x_0$$

Thus the polarization is given by

$$\begin{aligned}\mathbf{P} &= -Nq\mathbf{r} \\ &= -Nq \cdot \frac{qE_0(\hat{\mathbf{x}} \pm i\hat{\mathbf{y}})}{m[(\omega^2 - \omega_0^2) \mp \omega_c\omega]}\, e^{i(kz - \omega t)} \\ &= \chi \mathbf{E}_{\pm}\end{aligned}$$

where the susceptibility χ is given by

$$\chi = \frac{Nq^2}{m} \cdot \frac{1}{(\omega_0^2 - \omega^2) \pm \omega_c\omega}$$

Thus the modes are circularly polarized, and the corresponding refractive indices are given by [cf. Eq. (86) of Chap. 7]

$$n_{\pm}^2 = 1 + \frac{Nq^2}{m\varepsilon_0} \cdot \frac{1}{(\omega_0^2 - \omega^2) \pm \omega_c\omega}$$

where the upper and lower signs correspond to RCP and LCP, respectively.

Summary

- For an electromagnetic wave propagating in the z direction, let the x and y components of the electric field be given by

  $E_x = a_1 \cos(kz - \omega t)$
  $E_y = a_2 \cos(kz - \omega t + \theta)$

  1. If $a_2 = 0$, we have an x-polarized wave. If $a_1 = 0$, we have a y-polarized wave. For $\theta = n\pi$, $n = 0, \pm1, \pm2, \ldots$, we again have a linearly polarized wave with the electric vector making an angle with the x axis—this angle is either $+\tan^{-1}(a_2/a_1)$ or $-\tan^{-1}(a_2/a_1)$.
  2. If $a_1 = a_2$ and $\theta = \pi/2, 5\pi/2, 9\pi/2, \ldots$, we have a RCP wave; for $\theta = 3\pi/2, 7\pi/2, \ldots$ we have a LCP wave.
  3. In general, we have either a LEP (left elliptically polarized) wave or a REP (right elliptically polarized) wave.

- Linearly polarized light can be produced by various methods:
  1. By allowing an unpolarized light to pass through a polaroid.
  2. By allowing an unpolarized light to fall on a dielectric surface at the Brewster angle $\theta_p \left(= \tan^{-1}\frac{n_2}{n_1}\right)$.
  3. By passing through a Nicol prism.
- If an unpolarized plane wave is incident on an uniaxial crystal, the plane wave will split into two plane waves. One is referred to as the ordinary wave (usually abbreviated as the *o*-wave), and the other is referred to as the extraordinary wave (usually abbreviated as the *e*-wave). For both waves, the space and time dependence of the vectors **E**, **D**, **B**, and **H** can be assumed to be of the form $e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}$ where **k** denotes the propagation vector and represents the direction normal to the phase fronts. In general, **k** vectors for the *o*- and *e*-waves will be different. Further,
  1. Both ordinary and extraordinary waves are linearly polarized.
  2. $\mathbf{D} \cdot \mathbf{k} = 0$ for both *o*- and *e*-waves.
  3. For the *o*-wave, the **D** vector is at right angles to the optic axis as well as to **k**.
  4. On the other hand, for the *e*-wave, **D** lies in the plane containing **k** and the optic axis, and of course, $\mathbf{D} \cdot \mathbf{k} = 0$.
- Consider a uniaxial crystal whose optic axis lies on the surface of the crystal. If the thickness of the crystal is such that a phase difference of $\pi/2$ is introduced between the ordinary and extraordinary waves, then the plate is said to be a quarter wave plate (usually abbreviated as a QWP). If a linearly polarized wave (with its **E** making 45° with the optic axis) is allowed to fall normally on a QWP, the output is a circularly polarized wave.
- When a linearly polarized light beam propagates through an optically active medium such as sugar solution, then as the beam propagates, its plane of polarization rotates.
- In a uniaxial crystal, the wave velocities associated with the ordinary and extraordinary waves are given by

$$v_w = v_{wo} = \frac{c}{n_o} \quad o\text{-wave}$$

$$v_{we}^2 = \frac{c^2}{n_{we}^2} = \frac{c^2}{n_o^2}\cos^2\psi + \frac{c^2}{n_e^2}\sin^2\psi \qquad e\text{-wave}$$

where ψ is the angle that **k** makes with the optic axis and n_0 and n_e are constants of the crystal. On the other hand, the corresponding ray velocities are given by

$$v_r = v_{ro} = \frac{c}{n_o} \qquad o\text{-ray}$$

$$\frac{1}{v_r^2} = \frac{1}{v_{re}^2} = \frac{n_{re}^2}{c^2} = \frac{\cos^2\theta}{c^2/n_0^2} + \frac{\sin^2\theta}{c^2/n_e^2} \qquad e\text{-ray}$$

where θ is the angle that the direction of energy propagation (direction of the Poynting vector) makes with the optic axis.

Problems

22.1 Discuss the state of polarization when the x and y components of the electric field are given by the following equations:

(a) $E_x = E_0 \cos(\omega t + kz)$
$E_y = \frac{1}{\sqrt{2}} E_0 \cos(\omega t + kz + \pi)$

(b) $E_x = E_0 \sin(\omega t + kz)$
$E_y = E_0 \cos(\omega t + kz)$

(c) $E_x = E_0 \sin\left(kz - \omega t + \frac{\pi}{3}\right)$
$E_y = E_0 \sin\left(kz - \omega t - \frac{\pi}{6}\right)$

(d) $E_x = E_0 \sin\left(kz - \omega t + \frac{\pi}{4}\right)$
$E_y = \frac{1}{\sqrt{2}} E_0 \sin(kz - \omega t)$

In each case, plot the rotation of the tip of the electric vector on the plane $z = 0$.

[**Ans:** (a) Linearly polarized, (b) right circularly polarized, (c) left circularly polarized, and (d) left elliptically polarized.]

22.2 The electric field components of a plane electromagnetic wave are

$E_x = 2E_0 \cos(\omega t - kz + \phi) \qquad E_y = E_0 \sin(\omega t - kz)$

Draw the diagram showing the state of polarization (i.e., circular, plane, elliptical, or unpolarized) when

(a) $\phi = 0$ (b) $\phi = \pi/2$ (c) $\phi = \pi/4$

22.3 Using the data given in Table 22.1, calculate the thickness of quartz half wave plate for $\lambda_0 = 5890$ Å.

[**Ans:** 32.34 μm]

22.4 A right circularly polarized beam is incident on a calcite half wave plate. Show that the emergent beam will be left circularly polarized.

22.5 What will be the Brewster angle for a glass slab ($n = 1.5$) immersed in water ($n = 4/3$)?

[**Ans:** 48.4°]

22.6 Consider the normal incidence of a plane wave on a quartz quarter wave plate whose optic axis is parallel to the surface (see Fig. 22.24). Thus the optic axis is along the z axis, and the propagation is along the x axis. Show that E_y propagates as an o-wave and E_z as an e-wave.

(a) Assuming

$E_y = E_0 \cos \omega t$
$E_z = E_0 \cos \omega t$ at $x = 0$

show that the emergent light is right circularly polarized.

(b) On the other hand, if one assumes

$E_y = E_0 \sin \omega t$
$E_z = E_0 \cos \omega t$ at $x = 0$

show that the emergent beam is linearly polarized.

22.7 Show that the angle between vectors **D** and **E** is the same as between the Poynting vector **S** and the propagation vector **k**.

22.8 Consider the propagation of an extraordinary wave through a KDP crystal. If the wave vector is at an angle of 45° to the optic axis, calculate the angle between **S** and **k**. Repeat the calculation for $LiNbO_3$. The values of n_o and n_e for KDP and $LiNbO_3$ are given in Table 22.1.

[**Ans:** 1.56° and 2.25°]

22.9 Prove that when the angle of incidence corresponds to the Brewster angle, the reflected and refracted rays are at right angles to each other.

22.10 (a) Consider two crossed Polaroids placed in the path of an unpolarized beam of intensity I_0 (see Fig. 22.6). If we place a third Polaroid in between the two, then, in general, some light will be transmitted through. Explain this phenomenon.

(b) Assuming the pass axis of the third Polaroid to be at 45° to the pass axis of either of the Polaroids, calculate the intensity of the transmitted beam. Assume that all the Polaroids are perfect.

[**Ans:** $\frac{1}{8} I_0$]

22.11 A quarter wave plate is rotated between two crossed Polaroids. If an unpolarized beam is incident on the first Polaroid, discuss the variation of intensity of the emergent beam as the quarter wave plate is rotated. What will happen if we have a half wave instead of a quarter wave plate?

22.12 In Prob. 22.11, if the optic axis of the quarter wave plate makes an angle of 45° with the pass axis of either Polaroid, show that only one-quarter of the incident intensity will be transmitted. If the quarter wave plate is replaced by a half wave plate, show that one-half of the incident intensity will be transmitted through.

22.13 For calcite the values of n_o and n_e for $\lambda_0 = 4046$ Å are 1.68134 and 1.49694, respectively; corresponding to $\lambda_0 = 7065$ Å the values are 1.65207 and 1.48359, respectively. We have a calcite quarter wave plate corresponding to $\lambda_0 = 4046$ Å. A left circularly polarized beam of $\lambda_0 = 7065$ Å is incident on this plate. Obtain the state of polarization of the emergent beam.

22.14 A HWP is introduced between two crossed Polaroids P_1 and P_2. The optic axis makes an angle of 15° with the pass axis of P_1 as shown in Fig. 22.39(a) and (b). If an unpolarized beam of intensity I_0 is normally incident on P_1 and if I_1, I_2, and I_3 are the intensities after P_1, after HWP, and after P_2, respectively, then calculate I_1/I_0, I_2/I_0, and I_3/I_0.

[**Ans.:** $\frac{1}{2}$, $\frac{1}{2}$, $\frac{1}{8}$]

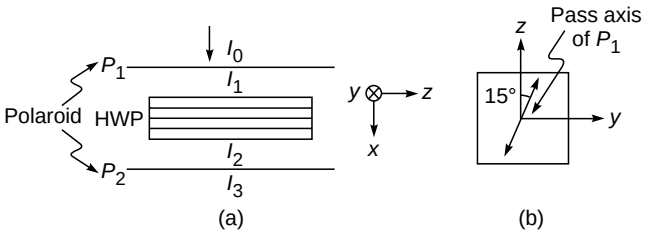


Fig. 22.39

22.15 Two prisms of calcite ($n_o > n_e$) are cemented together as shown in Fig. 22.40, so as to form a cube. Lines and dots show the direction of the optic axis. A beam of unpolarized light is incident normally from region I. Assume the angle of the prism to be 12°. Determine the path of rays in regions II, III, and IV indicating the direction of vibrations (i.e., the direction of **D**).

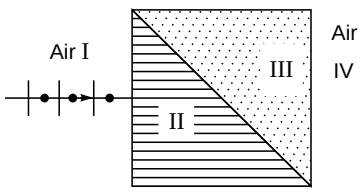


Fig. 22.40

22.16 A λ/6 plate is introduced in between the two crossed polarizers in such a way that the optic axis of the λ/6 plate makes an angle of 45° with the pass axis of the first polarizer (see Fig. 22.41). Consider an unpolarized beam of intensity I_0 to be incident normally on the polarizer. Assume the optic axis to be along the z axis and the propagation along the x axis. Write the y and z components of the electric fields (and the corresponding total intensities) after passing through (a) P_1, (b) λ/6 plate, and (c) P_2.

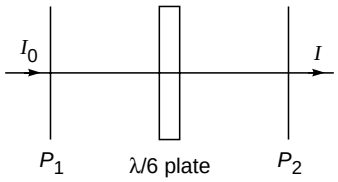


Fig. 22.41

22.17 A beam of light is passed through a polarizer. If the polarizer is rotated with the beam as an axis, the intensity I of the emergent beam does not vary. What are the possible states of polarization of the incident beam? How can you ascertain its state of polarization with the help of the given polarizer and a QWP?

22.18 Consider a Wollaston prism consisting of two similar prisms of calcite ($n_o = 1.66$ and $n_e = 1.49$) as shown in Fig. 22.29, with the angle of the prism now equal to 25°. Calculate the angular divergence of the two emerging beams.

22.19 (a) Consider a plane wave incident normally on a calcite crystal with its optic axis making an angle of 20° with the normal [see Fig. 22.21(a)]. Thus $\psi = 20°$. Calculate the angle that the Poynting vector will make with the normal to the surface. Assume $n_o \approx 1.66$ and $n_e \approx 1.49$.

(b) In (a) assume the crystal to be quartz with $n_o \approx 1.544$ and $n_e \approx 1.553$.

[**Ans:** (a) 4.31°]

22.20 Consider the incidence of the following REP beam on a sugar solution at $z = 0$:

$$E_x = 5 \cos \omega t \qquad E_y = 4 \sin \omega t$$

with $\lambda = 6328$ Å. Assume

$$n_l - n_r = 10^{-5} \qquad \text{and} \qquad n_l = 4/3$$

Study the evolution of the SOP of the beam.

22.21 Consider the incidence of the above REP beam on an elliptic core fiber with

$$\frac{\beta_x}{k_0} \approx 1.506845 \qquad \text{and} \qquad \frac{\beta_y}{k_0} \approx 1.507716$$

Calculate the SOP at $z = 0.25L_b$, $0.5L_b$, $0.75L_b$, and L_b.

22.22 When the optic axis lies on the surface of the crystal and in the plane of incidence, show (by geometrical considerations) that the angles of refraction of the ordinary and the extraordinary rays (which we denote by r_o and r_e, respectively) are related through the following equation:

$$\frac{\tan r_o}{\tan r_e} = \frac{n_o}{n_e}$$

REFERENCES AND SUGGESTED READINGS

1. W. A. Shurcliff and S. S. Ballard, *Polarized Light*, Van Nostrand, Princeton, N.J., 1964.
2. G. R. Bird and M. P. Parrish, "The Wire Grid as a near infrared polarizer," *Journal of the Optical Society of America*, Vol. 50, p. 886, 1960.
3. M. Alonso and E. J. Finn, *Physics*, Addison-Wesley, Reading, Mass., 1970.
4. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley, Reading, Mass., 1963.
5. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, England, 1970.
6. A. K. Ghatak and K. Thyagarajan, *Optical Electronics*, Cambridge University Press, Cambridge, United Kingdom, 1989. Reprinted by Foundation Books, New Delhi.
7. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, McGraw-Hill, New York, 1976.
8. *Polarized Light: Selected Reprints*, American Institute of Physics, New York, 1963.
9. S. Chandrashekhar, "Simple Model for Optical Activity," *Amer. J. Phys*, Vol. 24, p. 503, 1956.
10. P. Gay, *An Introduction to Crystal Optics*, Longmans Green and Co, London, 1967.
11. T. H. Waterman, "Polarized Light and Animal Navigation," *Scien. Amer*, July, 1955.
12. E. A. Wood, *Crystals and Light*, Van Nostrand Momentum Book No. 5, Van Nostrand, Princeton, N. J., 1964.
13. L. B. Jeunhomme, *Single Mode Fiber Optics*, Marcel Dekker, New York, 1983.

Chapter Twenty-Three

ELECTROMAGNETIC WAVES

Maxwell could say, when he was finished with his discovery, "Let there be electricity and magnetism, and there is light."

—Richard Feynman

23.1 MAXWELL'S EQUATIONS

All electromagnetic phenomena can be said to follow from Maxwell's equations. These equations are based on experimental observations and are given by

$$\nabla \cdot \mathbf{D} = \rho \quad \textbf{(1)}$$

$$\nabla \cdot \mathbf{B} = 0 \quad \textbf{(2)}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad \textbf{(3)}$$

and

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J} \quad \textbf{(4)}$$

where ρ represents the charge density and **J** the current density; and **E**, **D**, **B**, and **H** represent the electric field, electric displacement, magnetic induction, and magnetic field, respectively. Further,

$$\nabla \cdot \mathbf{D} \equiv \text{div}\, \mathbf{D}$$

and

$$\nabla \times \mathbf{E} \equiv \text{curl}\, \mathbf{E}$$

Equations (1) through (4) can be solved only if the "constitutive relations" are known which relate **D** to **E**, **B** to **H**, and **J** to **E**; the constitutive relations depend on the properties of the medium, field strengths, etc. For example, for an anisotropic medium ε is a tensor of the second tank (see Sec. 22.12); for high field strengths ε may itself depend on **E**. For a linear, isotropic, and homogeneous medium, the constitutive relations are given by

$$\mathbf{D} = \varepsilon \mathbf{E} \quad \textbf{(5)}$$

$$\mathbf{B} = \mu \mathbf{H} \quad \textbf{(6)}$$

and

$$\mathbf{J} = \sigma \mathbf{E} \quad \textbf{(7)}$$

where ε, μ, and σ denote, respectively, the dielectric permittivity, magnetic permeability, and conductivity of the medium. For a charge-free dielectric, we may write

$$\rho = 0 \quad \textbf{(8)}$$

$$\mu = \mu_0 \quad \textbf{(9)}$$

and

$$\mathbf{J} = 0 \quad \textbf{(10)}$$

where μ_0 ($= 4\pi \times 10^{-7}$ N s^2 C^{-2}) represents the magnetic permeability of vacuum. In many problems of interest, the propagation is in a dielectric medium, and the above constitutive relations are valid. If we use the above relations, Maxwell's equations simplify to

$$\nabla \cdot \mathbf{E} = 0 \quad \textbf{(11)}$$

$$\nabla \cdot \mathbf{H} = 0 \quad \textbf{(12)}$$

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \quad \textbf{(13)}$$

and

$$\nabla \times \mathbf{H} = \varepsilon \frac{\partial \mathbf{E}}{\partial t} \quad \textbf{(14)}$$

23.2 PLANE WAVES IN A DIELECTRIC

In Sec. 23.3, using the above equations, we will derive the wave equation; however, in this section we will show that plane wave solutions satisfy Maxwell's equations and will study the properties of plane waves. For plane waves propagating in the direction of **k**, the electric and magnetic fields can be written in the form

$$\mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \quad \textbf{(15)}$$

and

$$\mathbf{H} = \mathbf{H}_0 e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} \quad \textbf{(16)}$$

where $\mathbf{E}_0$ and $\mathbf{H}_0$ are space- and time-independent vectors, but may, in general, be complex. Now

$$\nabla \cdot \mathbf{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}$$

and since

$$E_x = E_{0x} e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} = E_{0x}\, e^{i(k_x x + k_y y + k_z z - \omega t)}$$

we get

$$\frac{\partial E_x}{\partial x} = ik_x E_{0x}\, e^{i(k_x x + k_y y + k_z z - \omega t)}$$

Thus the equation $\nabla \cdot \mathbf{E} = 0$ gives

$$i(k_x E_{0x} + k_y E_{0y} + k_z E_{0z})\, e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} = 0 \qquad \textbf{(17)}$$

implying

$$\mathbf{k} \cdot \mathbf{E} = 0 \qquad \textbf{(18)}$$

Similarly, the equation $\nabla \cdot \mathbf{H} = 0$ gives

$$\mathbf{k} \cdot \mathbf{H} = 0 \qquad \textbf{(19)}$$

The above two equations tell us that **E** and **H** are at right angles to **k**; thus the waves are transverse. Now, using Eq. (15) gives

$$(\nabla \times \mathbf{E})_x = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = i(k_y E_{0z} - k_z E_{0y})\, e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}$$

$$= i(\mathbf{k} \times \mathbf{E})_x$$

Thus Eq. (13) gives

$$i\,(\mathbf{k} \times \mathbf{E})_x = i\omega\mu_0 H_x \qquad \Rightarrow \qquad H_x = \frac{(\mathbf{k} \times \mathbf{E})_x}{\omega\mu_0} \qquad \textbf{(20)}$$

Similarly, we can write for the *y* and *z* components of Eq. (13). Thus we obtain the vector equation

$$\mathbf{H} = \frac{\mathbf{k} \times \mathbf{E}}{\omega\mu_0} \qquad \textbf{(21)}$$

Similarly, Eq. (14) would give us

$$\mathbf{E} = \frac{\mathbf{H} \times \mathbf{k}}{\omega\varepsilon} \qquad \textbf{(22)}$$

showing that **k**, **E**, and **H** are at right angles to one another (see Fig. 23.1). From Eq. (21) we readily get

$$H_0 = \frac{k}{\omega\mu_0} E_0 \qquad \textbf{(23)}$$

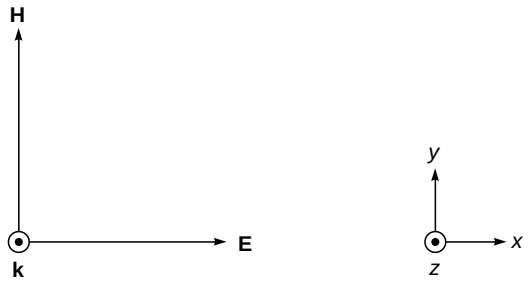


Fig. 23.1 If a plane wave is propagating in the *z* direction (which is coming out of the paper) and if at any instant of time the electric vector is along the *x* axis then the magnetic vector will be along the *y* axis.

Substituting for **H** from Eq. (21) into Eq. (22), we get

$$\mathbf{E} = \frac{1}{\omega^2\varepsilon\mu_0}\,[(\mathbf{k} \times \mathbf{E}) \times \mathbf{k}]$$

$$= \frac{1}{\omega^2\varepsilon\mu_0}\,[(\mathbf{k} \cdot \mathbf{k})\mathbf{E} - (\mathbf{k} \cdot \mathbf{E}) \times \mathbf{k}] \qquad \textbf{(24)}$$

where we have used the vector identity

$$(\mathbf{A} \times \mathbf{B}) \times \mathbf{C} = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{B} \cdot \mathbf{C})\mathbf{A} \qquad \textbf{(25)}$$

Since $\mathbf{k} \cdot \mathbf{E} = 0$, we get

$$\mathbf{E} = \frac{k^2}{\omega^2\varepsilon\mu_0}\mathbf{E}$$

Thus

$$k = \omega\sqrt{\varepsilon\mu_0} \qquad \textbf{(26)}$$

and the speed of propagation of the electromagnetic wave is given by

$$v = \frac{\omega}{k} = \frac{1}{\sqrt{\varepsilon\mu_0}} \qquad \textbf{(27)}$$

In free space

$$\varepsilon = \varepsilon_0 = 8.8542 \times 10^{-12}\ \ \mathrm{C^2\ N^{-1}\ m^{-2}} \qquad \textbf{(28)}$$

$$\mu = \mu_0 = 4\pi \times 10^{-7}\ \ \mathrm{N\ s^2\ C^{-2}} \qquad \textbf{(29)}$$

and

$$v = c = \frac{1}{\sqrt{8.8542 \times 10^{-12} \times 4\pi \times 10^{-7}}}$$

$$= 2.99794 \times 10^8\ \mathrm{m\ s^{-1}} \qquad \textbf{(30)}$$

The above equations show that the plane wave solutions given by Eqs. (15) and (16) do indeed satisfy Maxwell's equations where **k**, **E**, and **H** are at right angles to one another and related through Eqs. (21) to (23). Further, the speed of propagation of the electromagnetic wave is given by Eq. (27). If we assume the electric vector to be along the *x* axis, then the magnetic vector will be along the *y* axis so that we may write

$$\mathbf{E} = \hat{\mathbf{x}} E_0\, e^{i(kz - \omega t)} \qquad \textbf{(31)}$$

$$\mathbf{H} = \hat{\mathbf{y}} H_0\, e^{i(kz - \omega t)} \qquad \textbf{(32)}$$

with

$$H_0 = \frac{k}{\omega\mu_0} E_0 \qquad \textbf{(33)}$$

The actual electric fields are the real part of the exponentials appearing on the RHS of Eqs. (31) and (32):

$$\mathbf{E} = \hat{\mathbf{x}} E_0 \cos(kz - \omega t) \qquad \textbf{(34)}$$

$$\mathbf{H} = \hat{\mathbf{y}} H_0 \cos(kz - \omega t) \qquad \textbf{(35)}$$

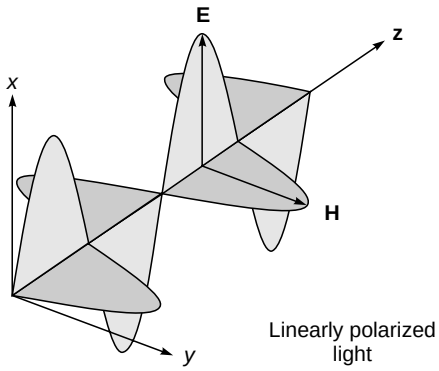


Fig. 23.2 The arrows represent the direction and magnitude of the **E** and **H** vectors (at a particular instant of time) for a plane polarized wave. The electric vectors always lie in the *xz* plane, and the magnetic vectors lie in the *yz* plane.

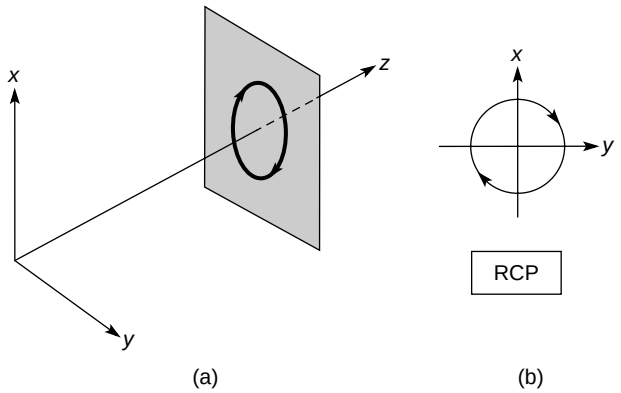


Fig. 23.3 For a right circularly polarized wave, if we look in the direction of the propagation of the wave, the electric vector rotates in a clockwise direction on the circumference of a circle.

where we have assumed E_0 and H_0 to be real. The plane wave as represented by Eq. (31) [or by Eq. (34)] is said to be linearly polarized (or *x*-polarized) because the electric vector is always along the *x* axis, and similarly, the magnetic vector is always along the *y* axis (see Fig. 23.2). Similarly, for a *y*-polarized wave, the electric vector is always along the *y* axis as shown in Fig. 22.1(b). We may also have superposition of two independent plane waves [we are considering the real part of the exponentials appearing on the RHS of Eqs. (31) and (32)]:

$$\mathbf{E}_1 = \hat{\mathbf{x}} E_0 \cos(kz - \omega t) \quad \textbf{(36)}$$

$$\mathbf{H}_1 = \hat{\mathbf{y}} H_0 \cos(kz - \omega t) \quad \textbf{(37)}$$

and

$$\mathbf{E}_2 = \hat{\mathbf{y}} E_0 \cos\left(kz - \omega t + \frac{\pi}{2}\right) = -\hat{\mathbf{y}} E_0 \sin(kz - \omega t) \quad \textbf{(38)}$$

$$\mathbf{H}_2 = -\hat{\mathbf{x}} H_0 \cos\left(kz - \omega t + \frac{\pi}{2}\right) = +\hat{\mathbf{x}} H_0 \sin(kz - \omega t) \quad \textbf{(39)}$$

The first wave is *x*-polarized, the second wave is *y*-polarized, and there is a phase difference of $\pi/2$. The superposition of these two waves gives the resultant

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 = E_0[\hat{\mathbf{x}} \cos(kz - \omega t) - \hat{\mathbf{y}} \sin(kz - \omega t)] \quad \textbf{(40)}$$

and $$\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_2 = H_0[\hat{\mathbf{y}} \cos(kz - \omega t) + \hat{\mathbf{x}} \sin(kz - \omega t)] \quad \textbf{(41)}$$

Now, at $z = 0$

$$E_x = E_0 \cos \omega t \quad \text{and} \quad E_y = E_0 \sin \omega t \quad \textbf{(42)}$$

and the tip of the electric vector will rotate (on the circumference of a circle) in the clockwise direction as shown in Fig. 23.3; this will represent a right circularly polarized (usually abbreviated as RCP) wave. Also, at $z = 0$

$$H_x = -H_0 \sin \omega t \quad \text{and} \quad H_y = H_0 \cos \omega t$$

and the tip of the **H** vector will also rotate (on the circumference of a circle) in the clockwise direction.

From the above equations we may draw the following inferences for plane waves:

1. Both **E** and **H** are at right angles to the direction of propagation. Thus the waves are transverse (see Fig. 23.1).
2. The vectors **E** and **H** are at right angles to each other; thus if the direction of propagation is along the *z* axis and if **E** is assumed to point in the x direction, then **H** will point in the *y* direction (see Fig. 23.1).
3. Since $k/\omega\mu$ is a real number, the electric and magnetic vectors are in phase; thus if at any instant **E** is zero, then **H** is also zero; similarly, when **E** attains its maximum value, **H** also attains its maximum value; etc.
4. The refractive index *n* of a dielectric (characterized by dielectric permittivity ε and magnetic permeability μ_0) is given by

$$n = \frac{c}{v} = \sqrt{\frac{\varepsilon\mu_0}{\varepsilon_0\mu_0}} = \sqrt{\frac{\varepsilon}{\varepsilon_0}} = \sqrt{\kappa} \quad \textbf{(43)}$$

   where $\kappa(= \varepsilon/\varepsilon_0)$ is known as the dielectric constant of the medium.
5. The electric and magnetic waves are interdependent; neither can exist without the other. Physically, an electric field varying in time produces a magnetic field varying in space and time; this changing magnetic field produces an electric field varying in space and time, and so on. This mutual generation of electric and magnetic fields results in the propagation of the electromagnetic wave.
6. Maxwell's equations are linear in **E** and **H**. So if $(\mathbf{E}_1, \mathbf{H}_1)$ and $(\mathbf{E}_2, \mathbf{H}_2)$ are two independent solutions of Maxwell's

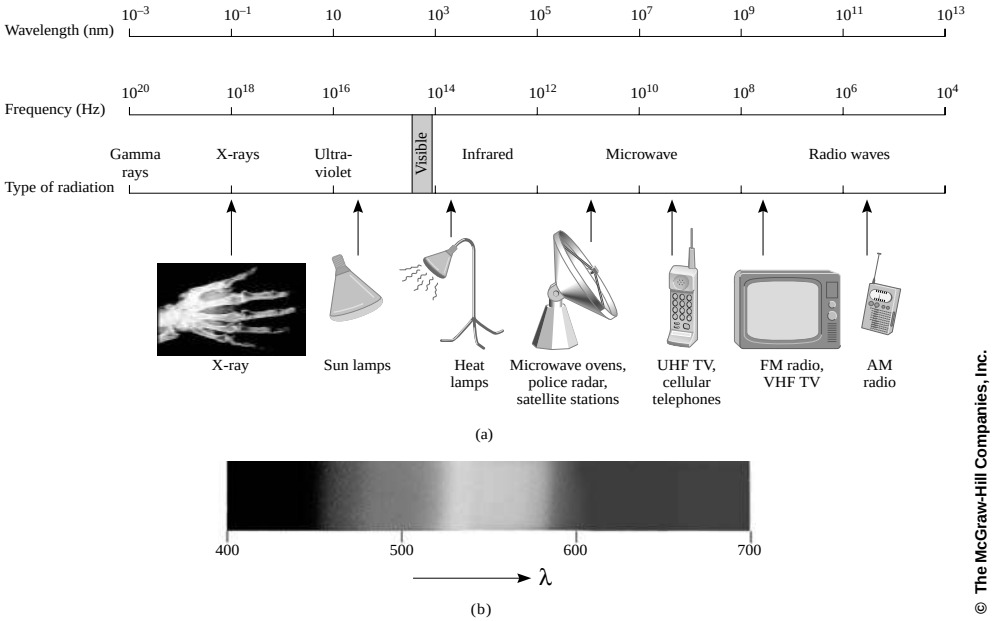


Fig. 23.4 The electromagnetic spectrum; visible light occupies a very small portion of the spectrum. A color photograph appears in the insert at the back of the book.

equations, then $(\mathbf{E}_1 + \mathbf{E}_2,\ \mathbf{H}_1 + \mathbf{H}_2)$ will also be a solution of Maxwell's equations. This is the superposition principle according to which the resultant displacement produced by two independent disturbances is the vector sum of the displacements produced by the disturbances independently.[1]

7. The plane wave as represented by Eq. (31) is said to be linearly polarized because the electric vector is always along the *x* axis and, similarly, the magnetic vector is always along the *y* axis (see Fig. 23.1).

8. There exists a wide and continuous variation of frequency (and wavelength) of electromagnetic waves as shown in Fig. 23.4. The radio waves correspond to wavelengths in the range of 10 to 1000 m whereas the wavelengths of X-rays are in the region of angstroms (1 Å = 10^{-10} m). The range of wavelengths of various kinds of electromagnetic waves is shown in Fig. 23.4, and as can be seen, the visible region (4×10^{-7} m < λ < 7×10^{-9} m) occupies a very small portion of the electromagnetic spectrum. The methods for production of different kinds of electromagnetic waves are different. For example, gamma rays are produced in nuclear decay processes, X-rays are usually produced by the sudden stopping and deflection of electrons, and radio waves are produced by varying the charge on an antenna. However, all wavelengths propagate with an identical speed in vacuum (which is denoted by *c*) and are always produced by accelerated charges.

23.3 THE THREE-DIMENSIONAL WAVE EQUATION IN A DIELECTRIC

In Sec. 23.2 we showed that plane wave solutions indeed satisfy Maxwell's equations. In this section we will show that the wave equation can be derived from Maxwell's equations. If we take the curl of Eq. (13), we obtain

$$\text{curl curl } \mathbf{E} = -\mu_0 \frac{\partial}{\partial t} \text{curl } \mathbf{H} = -\varepsilon\mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \qquad \textbf{(44)}$$

where we have used Eq. (14). Now, the operator $\nabla^2\mathbf{E}$ is *defined* by the following equation:

$$\nabla^2\mathbf{E} \equiv \text{grad div } \mathbf{E} - \text{curl curl } \mathbf{E} \qquad \textbf{(45)}$$

Using Cartesian coordinates, we can easily show that

$$(\nabla^2 \mathbf{E})_x = \frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} = \text{div grad } E_x$$

[1] Thus the superposition principle is a consequence of the linearity of Maxwell's equations. If, for example, the fields associated with the electromagnetic wave are so high that the dielectric permittivity ε depends on **E** itself, then Maxwell's equations will become nonlinear and the superposition principle will not remain valid. Indeed, when we discuss any nonlinear phenomenon, the superposition principle does not hold.

i.e., a Cartesian component of $\nabla^2\mathbf{E}$ is the div grad of the Cartesian component.[2] Thus, using

curl curl $\mathbf{E}$ = grad div $\mathbf{E} - \nabla^2\mathbf{E}$

we obtain

$$\text{grad div } \mathbf{E} - \nabla^2\mathbf{E} = -\varepsilon\mu_0 \frac{\partial^2\mathbf{E}}{\partial t^2} \tag{46}$$

or

$$\nabla^2\mathbf{E} = \varepsilon\mu_0 \frac{\partial^2\mathbf{E}}{\partial t^2} \tag{47}$$

where we have used the equation div $\mathbf{E} = 0$ [see Eq. (11)]. Equation (47) is known as the three-dimensional wave equation, and each Cartesian component of $\mathbf{E}$ satisfies the scalar wave equation (see Sec. 11.9)

$$\nabla^2\psi = \varepsilon\mu_0 \frac{\partial^2\psi}{\partial t^2} \tag{48}$$

The velocity of propagation v of the wave is simply given by

$$v = \frac{1}{\sqrt{\varepsilon\mu_0}} \tag{49}$$

In a similar manner, we can derive the wave equation satisfied by $\mathbf{H}$

$$\nabla^2\mathbf{H} = \varepsilon\mu_0 \frac{\partial^2\mathbf{H}}{\partial t^2} \tag{50}$$

It can be easily seen that the solutions expressed by Eqs. (31) and (32) [or Eqs. (34) and (35)] indeed satisfy Eqs. (47) and (50) provided

$$\frac{\omega}{k} = \frac{1}{\sqrt{\varepsilon\mu_0}} \tag{51}$$

which is the speed of propagation of the electromagnetic wave. Around 1860, Maxwell derived the wave equation, *predicted* the existence of electromagnetic waves, and calculated the speed of theses waves to be about 3.1074×10^8 m s^{-1}; this he found to be very close to the velocity of light which at that time was known to be 3.14858×10^8 m s^{-1} (as measured by Fizeau in1849). Just based on the closeness of these two numbers and with "*faith in the rationality of nature*," he propounded the electromagnetic theory of light and predicted that light must be an electromagnetic wave.[3] In the words of Maxwell himself, the speed of electromagnetic waves

> . . . calculated from the electromagnetic measurements of Kohlrausch and Weber, agrees so exactly with the velocity of light calculated from the experiments of M. Fizeau, that we can scarcely avoid the inference that *light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena.*

It was only in 1888 that Heinrich Hertz carried out experiments that could produce and detect electromagnetic waves of frequencies smaller than those of light. Hertz showed that the velocity of electromagnetic waves that he generated was the same as that of light.

In 1931 (during the centennial celebration of Maxwell's birth), Max Planck said, "[Maxwell's equations] . . . remain for all time one of the greatest triumphs of human intellectual endeavor." Albert Einstein said, "[The work of Maxwell was] . . . the most profound and the most fruitful that physics has experienced since the time of Newton."

23.4 THE POYNTING VECTOR

We rewrite Eqs. (3) and (4) as

$$\text{curl } \mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \tag{52}$$

and

$$\text{curl } \mathbf{H} = \frac{\partial\mathbf{D}}{\partial t} + \mathbf{J} \tag{53}$$

Now

$$\text{div } (\mathbf{E} \times \mathbf{H}) = \mathbf{H} \cdot \text{curl } \mathbf{E} - \mathbf{E} \cdot \text{curl } \mathbf{H} \tag{54}$$

Thus

$$\text{div } (\mathbf{E} \times \mathbf{H}) = -\mathbf{H} \cdot \frac{\partial\mathbf{B}}{\partial t} - \mathbf{J} \cdot \mathbf{E} - \mathbf{E} \cdot \frac{\partial\mathbf{D}}{\partial t} \tag{55}$$

For a linear material,

$$\begin{aligned} \mathbf{H} \cdot \frac{\partial\mathbf{B}}{\partial t} + \mathbf{E} \cdot \frac{\partial\mathbf{D}}{\partial t} &= \mu\mathbf{H} \cdot \frac{\partial\mathbf{H}}{\partial t} + \varepsilon\mathbf{E} \cdot \frac{\partial\mathbf{E}}{\partial t} \\ &= \frac{1}{2}\mu\frac{\partial}{\partial t}(\mathbf{H} \cdot \mathbf{H}) + \frac{1}{2}\varepsilon\frac{\partial}{\partial t}(\mathbf{E} \cdot \mathbf{E}) \\ &= \frac{1}{2}\frac{\partial}{\partial t}(\mathbf{B} \cdot \mathbf{H} + \mathbf{D} \cdot \mathbf{E}) \end{aligned}$$

Thus Eq. (55) can be rewritten in the form

$$\text{div } \mathbf{S} + \frac{\partial u}{\partial t} = -\mathbf{J} \cdot \mathbf{E} \tag{56}$$

where

$$\mathbf{S} \equiv \mathbf{E} \times \mathbf{H} \tag{57}$$

[2] However, $(\nabla^2\mathbf{E})_r \neq \text{div grad } E_r$.

[3] We also note that the physical laws described by Eqs. (1), (2), and (3) were known before Maxwell; he introduced only the term $\partial\mathbf{D}/\partial t$ (which is the concept of displacement current) in Eq. (4), and it is the presence of this term which leads to the prediction of electromagnetic waves.

is known as the *Poynting vector* and[4]

$$u = \tfrac{1}{2}\mathbf{B}\cdot\mathbf{H} + \tfrac{1}{2}\mathbf{D}\cdot\mathbf{E} \tag{58}$$

Equation (56) resembles the equation of continuity, and for a physical interpretation we note that if a charge q (moving with velocity v) is acted on by an electromagnetic field, then the work done by the field in moving it through a distance $d\mathbf{s}$ is $\mathbf{F}\cdot d\mathbf{s}$; thus the work done per unit time is

$$\begin{aligned}\mathbf{F}\cdot\frac{d\mathbf{s}}{dt} &= \mathbf{F}\cdot\mathbf{v}\\ &= (q\mathbf{E} + q\mathbf{v}\times\mathbf{B})\cdot\mathbf{v}\\ &= q\mathbf{E}\cdot\mathbf{v}\end{aligned} \tag{59}$$

If there are N charged particles per unit volume, each carrying a charge q, then the work done per unit volume is

$$Nq\mathbf{v}\cdot\mathbf{E} = \mathbf{J}\cdot\mathbf{E} \tag{60}$$

where $\mathbf{J}$ represents the current density. The energy appears in the form of kinetic (or heat) energy of the charged particles. Thus the term $\mathbf{J}\cdot\mathbf{E}$ represents the familiar Joule loss, and therefore, the quantity $\mathbf{J}\cdot\mathbf{E}$ on the RHS of Eq. (56) represents the rate at which energy is produced per unit volume per unit time. Consequently, we may interpret Eq. (56) as an equation of continuity[5] for energy with u representing the energy per unit volume. The quantities $\frac{1}{2}\mathbf{D}\cdot\mathbf{E}$ and $\frac{1}{2}\mathbf{B}\cdot\mathbf{H}$ represent the electrical and magnetic energies per unit volume, respectively. Further, we may interpret $\mathbf{S}\cdot d\mathbf{a}$ as the electromagnetic energy crossing the area $d\mathbf{a}$ per unit time. For plane waves in a dielectric, we may write

$$\begin{aligned}\mathbf{E} &= \hat{\mathbf{x}}\mathbf{E}_0\cos(kz-\omega t)\\ \mathbf{H} &= \hat{\mathbf{y}}\mathbf{H}_0\cos(kz-\omega t) = \hat{\mathbf{y}}\frac{k}{\omega\mu}\mathbf{E}_0\cos(kz-\omega t)\end{aligned} \tag{61}$$

Thus

$$\begin{aligned}\mathbf{S} &= \mathbf{E}\times\mathbf{H}\\ &= \hat{\mathbf{z}}\frac{k}{\omega\mu}E_0^2\cos^2(kz-\omega t)\end{aligned} \tag{62}$$

which implies that the energy flow is in the z direction (which represents the direction of propagation of the wave) and that an amount of energy

$$\frac{k}{\omega\mu}E_0^2\cos^2(kz-\omega t)$$

crosses a unit area (perpendicular to the z axis) per unit time. For optical beams $\omega \approx 10^{15}\ \mathrm{s}^{-1}$ and the $\cos^2$ term fluctuates with extreme rapidity[6] and any detector would record only an average value. Since

$$\begin{aligned}\langle\cos^2(kz-\omega t)\rangle &= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{+T}\cos^2(kz-\omega t)\,dt\\ &= \tfrac{1}{2}\end{aligned}$$

we obtain

$$\langle\mathbf{S}\rangle = \hat{\mathbf{z}}\frac{k}{2\omega\mu}E_0^2 \tag{63}$$

where $\langle\cdots\rangle$ denotes the time average of the quantity inside the angular brackets (see Sec. 17.5).
In general, for a plane wave

$$\mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{64}$$

$$\mathbf{H} = \mathbf{H}_0 e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{65}$$

and as shown in Sec. 23.2

$$\mathbf{H} = \frac{\mathbf{k}\times\mathbf{E}}{\omega\mu} \tag{66}$$

---

[4] Equation (58) is valid even for anisotropic media because in the principal axis system [see Eq. (66) of Chap. 22]

$$\begin{aligned}\mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t} &= \frac{1}{2}\varepsilon_x\frac{\partial E_x^2}{\partial t} + \frac{1}{2}\varepsilon_y\frac{\partial E_y^2}{\partial t} + \frac{1}{2}\varepsilon_z\frac{\partial E_z^2}{\partial t}\\ &= \frac{1}{2}\frac{\partial}{\partial t}(D_xE_x + D_yE_y + D_zE_z)\end{aligned}$$

[5] The equation of continuity is always written in the form

$$\operatorname{div}\mathbf{J} + \frac{\partial\rho}{\partial t} = 0$$

where ρ represents the charge density and $\mathbf{J}$ the current density; i.e., $\mathbf{J}\cdot d\mathbf{a}$ represents the amount of charge crossing the area $d\mathbf{a}$ per unit time.

[6] See also Secs. 14.3 and 14.6.

$$\mathbf{E} = \frac{\mathbf{H} \times \mathbf{k}}{\omega \varepsilon} \tag{67}$$

showing that **k**, **E** and **H** are at right angles to one another. The Poynting vector **S** is obviously in the direction of **k**.

We must hasten to point out that $\mathbf{S} \cdot d\mathbf{a}$ does not *always* represent the rate of energy flow through area $d\mathbf{a}$; for example, we may have *static* electric and magnetic field where $\mathbf{E} \times \mathbf{H}$ is finite, but we know that there is no energy flow. However, the integral

$$\oint \mathbf{S} \cdot d\mathbf{a}$$

over a closed surface rigorously represents the net energy flowing out of the surface. This follows immediately if we carry out a volume integral of Eq. (56) to give

$$\int \text{div}\mathbf{S}\, dV + \frac{\partial}{\partial t}\int u\, dV = -\int \mathbf{J} \cdot \mathbf{E}\, dV$$

or

$$-\frac{\partial}{\partial t}\int u\, dV = \oint \mathbf{S} \cdot d\mathbf{a} + \int \mathbf{J} \cdot \mathbf{E}\, dV$$

where we have used the divergence theorem. The quantity on the LHS represents the rate of decrease of the total energy; this must be equal to the Joule loss plus the net flow out of the surface enclosing the volume.

23.4.1 The Oscillating Dipole

Consider an oscillating dipole in the *z* direction:

$$\mathbf{p} = p_0 e^{-i\omega t}\hat{\mathbf{z}}$$

At large distances from such a dipole the fields are of the form (see, e.g., Ref. 8, p. 258)

$$\mathbf{E} = -\left(\frac{k^2 p_0}{4\pi\varepsilon_0}\right)(\sin\theta)\frac{e^{i(kr-\omega t)}}{r}\hat{\boldsymbol{\theta}} \tag{68}$$

$$\mathbf{H} = -\left(\frac{\omega k p_0}{4\pi}\right)(\sin\theta)\frac{e^{i(kr-\omega t)}}{r}\hat{\boldsymbol{\phi}} \tag{69}$$

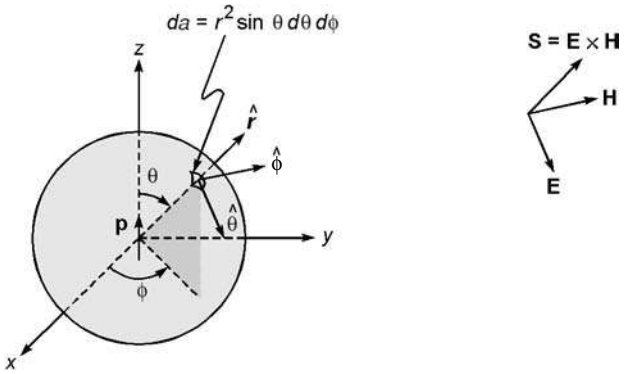


Fig. 23.5 The direction of the electric and magnetic fields and of the Poynting vector from an oscillating dipole. To calculate the total energy radiated per unit time, we must integrate the Poynting vector over the surface of a sphere.

where $k = \omega\sqrt{\varepsilon_0\mu_0}$ and the other symbols have their usual meaning; the unit vectors $\hat{\mathbf{r}}, \hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$ are shown in Fig. 23.5. Notice that the fields fall off as $1/r$ and that they are in phase. Further, the ratio of the amplitudes of the magnetic and electric fields is

$$\frac{\omega k p_0/4\pi}{k^2 p_0/4\pi\varepsilon_0} = \frac{\omega\varepsilon_0}{k} = \frac{k}{\omega\mu_0} \tag{70}$$

which is consistent with Eq. (23). Because of the $\sin\theta$ factor in Eqs. (68) and (69), the dipole does not produce any field along the direction of oscillation (see also Secs. 22.2.4 and 22.8). Thus[7]

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}$$

$$= \frac{\omega k^3 p_0^2}{16\pi^2\varepsilon_0}(\sin^2\theta)\frac{\cos^2(kr-\omega t)}{r^2}\hat{\mathbf{r}} \tag{71}$$

Equation (71) shows that **S** falls off as $1/r^2$, as it indeed should for a spherical wave (this is the inverse square law). If we integrate over a sphere of radius *r*, we obtain

[7] To calculate the Poynting vector, we must take the products of the real parts of **E** and **H**. Note that in the complex representation if $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$, then

$$\text{Re}\,(\mathbf{E}) = \text{Re}\,(\mathbf{E}_1) + \text{Re}\,(\mathbf{E}_2)$$

However,

$$(\text{Re}\,\mathbf{E}_1) \times (\text{Re}\,\mathbf{E}_2) \neq \text{Re}\,(\mathbf{E}_1 \times \mathbf{E}_2)$$

Here Re (**E**) denotes the real part of **E**.

$$P = \oint \mathbf{S}\cdot d\mathbf{a} = r^2 \int\int \mathbf{S}\cdot\hat{\mathbf{r}}\sin\theta\, d\theta d\phi$$

$$= \frac{\omega k^3 p_0^2}{16\pi^2\varepsilon_0}\cos^2(kr-\omega t)\int_0^{\pi}\sin^2\theta\sin\theta\, d\theta\int_0^{2\pi} d\phi$$

$$= \frac{\omega k^3 p_0^2}{6\pi\varepsilon_0}\cos^2(kr-\omega t) \qquad (72)$$

where P represents the instantaneous radiated power. Since the $\cos^2$ term fluctuates very rapidly, the average radiated power is given by

$$\overline{P} = \frac{\omega k^3 p_0^2}{12\pi\varepsilon_0} \qquad (73)$$

23.5 ENERGY DENSITY AND INTENSITY OF AN ELECTROMAGNETIC WAVE

In Sec. 23.4 we showed that the energy per unit volume associated with a plane wave is given by

$$u = \frac{1}{2}\mathbf{D}\cdot\mathbf{E} + \frac{1}{2}\mathbf{B}\cdot\mathbf{H} = \frac{1}{2}\varepsilon E^2 + \frac{1}{2\mu}B^2 \qquad (74)$$

For a linearly polarized plane wave, we may write

$$E_x = E_0\cos(kz-\omega t) \qquad E_y = 0,\ E_z = 0 \qquad (75)$$

$$B_x = 0 \qquad B_y = B_0\cos(kz-\omega t) \qquad B_z = 0 \qquad (76)$$

Thus

$$u = \frac{1}{2}\varepsilon E_0^2\,(kz-\omega t) + \frac{1}{2\mu}B_0^2\cos^2(kz-\omega t)$$

Since

$$B_0 = \sqrt{\varepsilon\mu}\,E_0$$

[see Eq. (33)], we get

$$\frac{B_0^2}{2\mu} = \frac{1}{2}\varepsilon E_0^2$$

Thus the energy associated with the electric field is equal to the energy associated with the magnetic field. If we take the time average of the $\cos^2$ terms, we get

$$\langle u\rangle = \frac{1}{2}\varepsilon E_0^2 \qquad (77)$$

Further, to obtain the intensity of the beam, we must multiply $\langle u\rangle$ by the speed of propagation, which will give us the energy crossing a unit area in unit time. Thus, the intensity is given by

$$I = \frac{1}{2}\varepsilon v E_0^2 = \frac{1}{2}\sqrt{\frac{\varepsilon}{\mu}}\,E_0^2 \qquad (78)$$

This should be consistent with Eq. (63). Indeed, if we substitute $k = \omega\sqrt{\varepsilon\mu}$ in Eq. (63), we obtain Eq. (78). In free space

$$I = \frac{1}{2}\varepsilon_0 c E_0^2$$

$$= \frac{1}{2}(8.854\times10^{-12}\ \mathrm{C^2\ N^{-1}\ m^{-2}})\times(3\times10^8\ \mathrm{m\ s^{-1}})E_0^2$$

$$= (1.33\times10^{-3}\ \mathrm{W\ V^{-2}})E_0^2$$

For example, for a 100 W lamp, the intensity at a distance of 10 m is

$$I = \frac{100}{4\pi(10)^2} \simeq 7.96\times10^{-2}\ \mathrm{W\ m^{-2}}$$

where we have assumed light to spread out uniformly in all directions. Thus

$$E_0 = \left(\frac{7.96\times10^{-2}}{1.33\times10^{-3}}\right)^{1/2} \approx 7.74\ \mathrm{V\ m^{-1}}$$

It is of interest to mention here that since a laser beam is almost perfectly parallel, it can be focused by a lens to a cross-sectional area of less than 10^{-6} cm (see Sec. 18.4.1). Thus for a 10^5 W laser beam, the intensity at the focal plane is

$$I = \frac{10^5\ \mathrm{W}}{10^{-10}\ \mathrm{m^2}} = 10^{15}\ \mathrm{W\ m^{-2}}$$

Thus

$$E_0 = \left(\frac{10^{15}}{1.33\times10^{-3}}\right)^{1/2} \approx 0.87\times10^9\ \mathrm{V\ m^{-1}}$$

Such high electric fields can cause extreme high temperatures which may result in the burning of a target (see Figs. 18.18 and 18.19).

23.6 RADIATION PRESSURE[8]

Let us consider a linearly polarized electromagnetic wave propagating in the $+z$ direction; we assume the electric field to be along the x direction and the magnetic field along the y direction (see Fig. 23.1). The electromagnetic wave is assumed to interact with a charge q; the electric field makes the charge move up and down along the x axis. Thus the charge

[8] See Ref. 4, Sec. 34.9. A rigorous analysis is given in Ref. 8, Chap. 12.

acquires a certain velocity in the x direction, and since the magnetic field is along the y axis, a force

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B} \tag{79}$$

acts on the charge q. This force acts along the z axis[9] (i.e., along the direction of propagation of the wave) and constitutes what is known as *radiation pressure*. Thus,

$$\mathbf{F} = qvB\hat{\mathbf{z}} \tag{80}$$

since

$$B = \mu_0 H = \frac{k}{\omega} E = \frac{E}{c} \tag{81}$$

[see Eq. (21)], we get

$$\mathbf{F} = \frac{qEv}{c}\hat{\mathbf{z}} \tag{82}$$

Now qEv represents the work done by the field on the charge per unit time; thus, if we consider a unit volume, then

$$\mathbf{F} = \frac{1}{c}\frac{du}{dt}\hat{\mathbf{z}} \tag{83}$$

But the force is equal to the change in momentum per unit time; consequently, the momentum per unit volume associated with the plane wave is given by

$$\mathbf{p} = \frac{u}{c}\hat{\mathbf{z}} \tag{84}$$

In Chap. 25, we will show that light essentially consists of corpuscles called photons. Each photon carries an energy equal to $h\nu$; the photon momentum, therefore, is given by

$$p = \frac{h\nu}{c} \tag{85}$$

Let us consider a plane wave incident normally on a perfect absorber. If we consider an area dS on the absorbing surface, then the momentum transferred to area dS in time dt is

$$p\, dS\, c dt$$

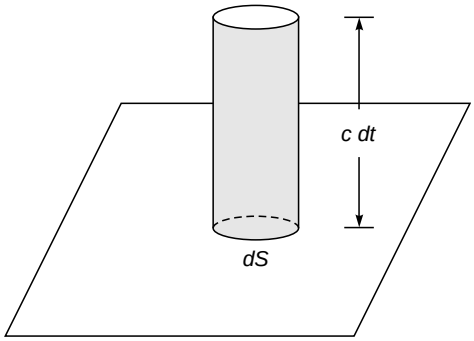


Fig. 23.6 A cylindrical volume to calculate radiation pressure.

which represents the momentum contained in a cylindrical volume $dS\, c\, dt$ (see Fig. 23.6). Thus the force acting on the area dS is

$$pc\, dS$$

Hence

$$P_{\text{rad}} = cp = u \tag{86}$$

where P_{rad} represents the radiation pressure due to a plane wave incident on a perfect absorber. On the other hand, for a perfect reflector, the momentum of the reflected wave is equal and opposite to the momentum associated with the incident wave. Thus the momentum transferred is twice the above value and hence

$$P_{\text{rad}} = 2cp = 2u \tag{87}$$

To have a numerical appreciation, let us consider a light beam of intensity $I = 3000\ \text{W m}^{-2}$ falling on a perfectly reflecting mirror. Since $I = cu$, we have

$$u = \frac{3000\ \text{W m}^{-2}}{3\times 10^8\ \text{m s}^{-1}} = 10^{-5}\ \text{J m}^{-3}$$

---

[9] Using the analysis of Sec. 7.5, we can show that in the presence of a field $\mathbf{E} = \hat{\mathbf{x}} E_0 \cos(kz - \omega t)$, the displacement is given by

$$\mathbf{x} = \hat{\mathbf{x}} qE_0 A \cos(kz - \omega t + \phi)$$

where we have explicitly shown that the amplitude is proportional to q and E_0. Thus

$$\mathbf{v} = \frac{d\mathbf{x}}{dt} = \hat{\mathbf{x}} qE_0 A\omega \sin(kz - \omega t + \phi)$$

Now

$$\mathbf{B} = \hat{\mathbf{y}} B_0 \cos(kz - \omega t) = \hat{\mathbf{y}}\frac{E_0}{c}\cos(kz - \omega t)$$

Thus

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B} = +\ \hat{\mathbf{z}} q^2 \frac{E_0^2 \omega}{c}\mathbf{A}[\cos(kz - \omega t)]\,[\sin(kz - \omega t)\cos\phi + \cos(kz - \omega t)\sin\phi]$$

If we carry out a time averaging, then

$$\langle \mathbf{F} \rangle = \hat{\mathbf{z}}\frac{q^2 E_0^2 \omega}{2c} A \sin\phi = \hat{\mathbf{z}}\frac{1}{c}\langle q\mathbf{E} \cdot \mathbf{v} \rangle$$

Since $\sin\phi$ is always positive [see Eqs. (22) and (51) of Chap. 8], the force is always in the z direction.

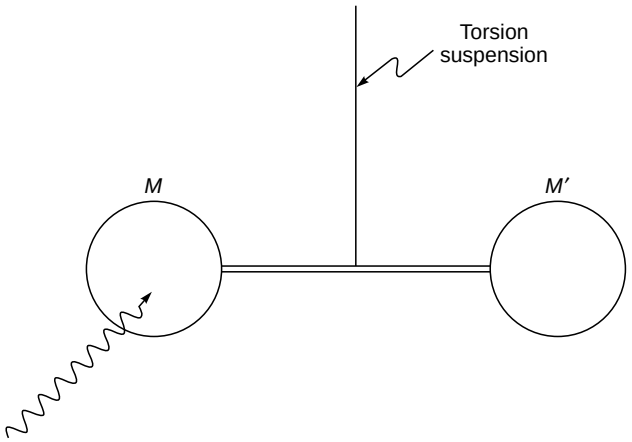


Fig. 23.7 An experimental arrangement to measure radiation pressure.

The radiation pressure is

$$10^{-5}\,\mathrm{N\,m^{-2}}$$

which may be compared with the atmospheric pressure ($\approx 10^5\,\mathrm{N\,m^{-2}}$).

It has been possible to measure the radiation pressure by allowing a light beam to fall on a highly polished mirror *M* (see Fig. 23.7).[10] The radiation pressure caused a twist in the suspension which was measured. The intensity of the beam can be determined by allowing it to fall on an absorber (such as a blackened disk) and measuring the temperature rise. In a particular run, the radiation pressure was found to be about $7.01 \times 10^{-6}\,\mathrm{N\,m^{-2}}$ which was in good agreement with the predicted value of $7.05 \times 10^{-6}\,\mathrm{N\,m^{-2}}$.

For oblique incidence on a perfect reflector, the change in momentum per unit volume is $2p\cos\theta$, and the radiation pressure is

$$P_{\text{rad}} = 2cp\cos^2\theta = 2u\cos^2\theta \qquad \textbf{(88)}$$

where θ represents the angle of incidence.

23.7 THE WAVE EQUATION IN A CONDUCTING MEDIUM

In Sec. 23.3 we assumed $\mathbf{J} = 0$. For a conducting medium

$$\mathbf{J} = \sigma\mathbf{E} \qquad \textbf{(89)}$$

where σ represents the conductivity of the medium. Thus, Maxwell's equations become

$$\operatorname{div}\mathbf{E} = 0 \qquad \textbf{(90)}$$

$$\operatorname{div}\mathbf{H} = 0 \qquad \textbf{(91)}$$

$$\operatorname{curl}\mathbf{E} = -\mu\frac{\partial\mathbf{H}}{\partial t} \qquad \textbf{(92)}$$

$$\operatorname{curl}\mathbf{H} = \sigma\mathbf{E} + \varepsilon\frac{\partial\mathbf{E}}{\partial t} \qquad \textbf{(93)}$$

Taking the curl of Eq. (92), we get

$$\operatorname{curl}\operatorname{curl}\mathbf{E} = -\mu\frac{\partial}{\partial t}\operatorname{curl}\mathbf{H}$$

or

$$\operatorname{grad}\operatorname{div}\mathbf{E} - \nabla^2\mathbf{E} = -\mu\sigma\frac{\partial\mathbf{E}}{\partial t} - \mu\varepsilon\frac{\partial^2\mathbf{E}}{\partial t^2}$$

Using Eq. (90), we get

$$\nabla^2\mathbf{E} - \mu\sigma\frac{\partial\mathbf{E}}{\partial t} - \mu\varepsilon\frac{\partial^2\mathbf{E}}{\partial t^2} = 0 \qquad \textbf{(94)}$$

which is the wave equation for a conducting medium. For a plane wave of the type

$$\mathbf{E} = \mathbf{E}_0\, e^{i(kz - \omega t)} \qquad \textbf{(95)}$$

we obtain

$$-k^2 + i\omega\mu\sigma + \omega^2\varepsilon\mu = 0 \qquad \textbf{(96)}$$

which shows that k must be a complex number. If we write

$$k = \alpha + i\beta \qquad \textbf{(97)}$$

then

$$-(\alpha^2 + 2i\alpha\beta - \beta^2) + i\omega\mu\sigma + \omega^2\varepsilon\mu = 0$$

Equating real and imaginary parts, we get

$$\alpha^2 - \beta^2 = \omega^2\varepsilon\mu \qquad \textbf{(98)}$$

and

$$\beta = \frac{\omega\mu\sigma}{2\alpha} \qquad \textbf{(99)}$$

Substituting for β in Eq. (98) and solving for α, we get

$$\alpha = \omega\sqrt{\varepsilon\mu}\left[\frac{1}{2} \pm \frac{1}{2}\left(1 + \frac{\sigma^2}{\omega^2\varepsilon^2}\right)^{1/2}\right]^{1/2} \qquad \textbf{(100)}$$

We must choose the positive sign; the negative sign would make α complex. Thus

$$\alpha = \omega\sqrt{\varepsilon\mu}\left[\frac{1}{2} + \frac{1}{2}\left(1 + \frac{\sigma^2}{\omega^2\varepsilon^2}\right)^{1/2}\right]^{1/2}$$
$$\beta = \frac{\omega\mu\sigma}{2\alpha} \qquad \textbf{(101)}$$

[10] The experiment was first carried out by Lebedev in Russia in 1899; the experimental arrangement shown in Fig. 23.7 is similar to that of Nichols and Hull who performed the experiment in 1901 and confirmed the prediction of radiation pressure.

Now, when k is complex, Eq. (95) becomes

$$\mathbf{E} = \mathbf{E}_0 \exp(-\beta z) \exp[i(\alpha z - \omega t)] \quad \textbf{(102)}$$

which represents an attenuated wave. The attenuation is due to the Joule loss. For a good conductor[11]

$$\frac{\sigma}{\omega\varepsilon} >> 1 \quad \textbf{(103)}$$

and we obtain

$$\alpha \approx \beta \approx \left(\frac{\omega\mu\sigma}{2}\right)^{1/2} \quad \textbf{(104)}$$

Indeed if $\sigma/\omega\varepsilon << 1$ (say, $\lesssim 0.01$), the medium can be classified as a dielectric; and if $\sigma/\omega\varepsilon >> 1$ (say, $\gtrsim 100$), the medium can be classified as a conductor. For

$$0.01 \lesssim \frac{\sigma}{\omega\varepsilon} \lesssim 100$$

the medium is said to be a quasi-conductor. Thus, depending on the frequency, a particular material can behave as a dielectric or as a conductor. For example, for freshwater $\varepsilon/\varepsilon_0 \approx 80$ and $\sigma \approx 10^{-3}$ mho m^{-1}. (Both ε and σ can be assumed to be constants at low frequencies.) Thus

$$\frac{\sigma}{\varepsilon} \approx \frac{10^{-3}}{80 \times 8.85 \times 10^{-12}} \approx 1.4 \times 10^6 \text{ s}^{-1}$$

For $\omega = 2\pi \times 10$ s^{-1}

$$\frac{\sigma}{\omega\varepsilon} \approx 2 \times 10^4$$

and for $\omega = 2\pi \times 10^{10}$ s^{-1}

$$\frac{\sigma}{\omega\varepsilon} \approx 2 \times 10^{-5}$$

Thus, freshwater behaves as a good conductor for $\nu \leq 10^3$ s^{-1} and as a dielectric for $\nu \geq 10^7$ s^{-1}. On the other hand, for copper we may assume $\varepsilon \approx \varepsilon_0$ and $\sigma \approx 5.8 \times 10^7$ mho m^{-1}, and for $\omega \approx 2\pi \times 10^{10}$ s^{-1}

$$\frac{\sigma}{\omega\varepsilon} \approx \frac{5.8 \times 10^7}{2\pi \times 10^{10} \times 8.9 \times 10^{-12}} \approx 10^8$$

Thus even for such frequencies it behaves as an excellent conductor.

From Eq. (102), it can be easily seen that the field decreases by a factor e in traversing a distance

$$\delta = \frac{1}{\beta}$$

which is known as the penetration depth. For copper,

$$\mu = \mu_0 = 4\pi \times 10^{-7} \text{ N s}^2/\text{C}^2$$

and

$$\delta \approx \left(\frac{2}{\omega\mu\sigma}\right)^{1/2} \approx \left(\frac{2}{2\pi\nu \times 4\pi \times 10^{-7} \times 5.8 \times 10^7}\right)^{1/2}$$

$$\approx \frac{0.065}{\sqrt{\nu}} \quad \text{m}$$

Thus for $\nu \approx 100$ s^{-1}, $\delta \approx 0.0065$ m = 0.65 cm whereas for $\nu \approx 10^8$ s^{-1}, $\delta \approx 6.5 \times 10^{-6}$ m, showing that the penetration decreases with an increase in frequency.

23.8 THE CONTINUITY CONDITIONS

In this section we will derive the continuity conditions for electric and magnetic fields at the interface of two media. Let us first consider the equation

$$\text{div } \mathbf{B} = 0 \quad \textbf{(105)}$$

At the interface of two media, we consider a pillbox which encloses an area ΔS of the interface (see Fig. 23.8). Let the height of the pillbox be l.

Now if we integrate div $\mathbf{B}$ over the cylindrical volume, then, using Gauss' theorem, we obtain

$$0 = \int \text{div } \mathbf{B}\, dV = \oint_{S_1} \mathbf{B} \cdot d\mathbf{a} + \oint_{S_2} \mathbf{B} \cdot d\mathbf{a} + \oint_{S_3} \mathbf{B} \cdot d\mathbf{a}$$

where S_1 and S_2 represent the flat faces of the cylinder and S_3 represents the curved surface of the cylinder. If we let

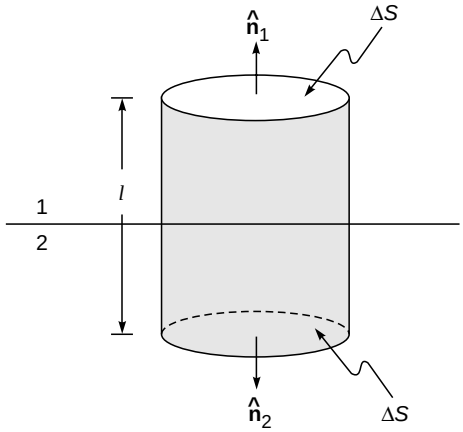


Fig. 23.8 A cylindrical pillbox at the interface of two dielectrics.

[11] The corresponding expressions for an insulator are given in Prob. 23.8.

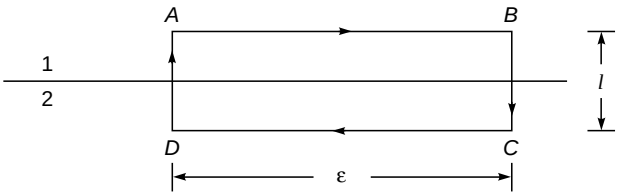


Fig. 23.9 A rectangular loop at the interface of two dielectrics.

$l \to 0$, then the third integral vanishes and we obtain

$$\oint_{S_1} \mathbf{B} \cdot d\mathbf{a} = -\oint_{S_2} \mathbf{B} \cdot d\mathbf{a}$$

or $$\mathbf{B_1} \cdot \hat{\mathbf{n}}_1 \Delta S = -\mathbf{B_2} \cdot \hat{\mathbf{n}}_2 \Delta S \quad \textbf{(106)}$$

or $$B_{1n} = B_{2n} \quad \textbf{(107)}$$

where the directions of $\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}}_2$ are shown in Fig. 23.8. Thus, the normal component of **B** is continuous across the interface.

Similarly, in the absence of free charges

$$\text{div } \mathbf{D} = 0$$

and we obtain[12]

$$D_{1n} = D_{2n} \quad \textbf{(108)}$$

showing that the normal component of **D** is also continuous across the interface.

We next consider the equation

$$\text{curl } \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0$$

We consider a rectangular loop *ABCD* as shown in Fig. 23.9. Now

$$0 = \oint_S \text{curl } \mathbf{E} \cdot d\mathbf{a} + \oint_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a} \quad \textbf{(109)}$$

where the surface integral is over any surface bounding the loop *ABCD*. Using Stokes' theorem, we get

$$\oint \mathbf{E} \cdot d\mathbf{l} = -\int \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a} \quad \textbf{(110)}$$

or

$$\left(\int_{AB} + \int_{BC} + \int_{CD} + \int_{DA} \right) \mathbf{E} \cdot d\mathbf{l} = -\int \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a}$$

If we let $l \to 0$, then the integrals along *BC* and *DA* tend to zero, and since the area of the loop also tends to zero, the RHS vanishes. Thus we obtain

$$\int_{AB} \mathbf{E} \cdot d\mathbf{l} + \int_{CD} \mathbf{E} \cdot d\mathbf{l} = 0$$

or $$(\mathbf{E}_1 \cdot \hat{\mathbf{t}})\varepsilon + [\mathbf{E}_2 \cdot (-\hat{\mathbf{t}})]\varepsilon = 0$$

or $$\mathbf{E}_{1t} = \mathbf{E}_{2t}$$

where $\mathbf{E}_{1t}$ and $\mathbf{E}_{2t}$ represent the tangential components of **E** which are continuous across the interface.

Similarly, Eq. (8) gives us[13]

$$\mathbf{H}_{1t} = \mathbf{H}_{2t}$$

In summary, in the absence of any surface current and surface charges, the normal components of **B** and **D** and the tangential components of **H** and **E** are continuous across an interface.

23.9 PHYSICAL SIGNIFICANCE OF MAXWELL'S EQUATIONS

Let us first consider the equation

$$\text{div } \mathbf{D} = \rho \quad \textbf{(111)}$$

In free space

$$\mathbf{D} = \varepsilon_0 \mathbf{E} \quad \textbf{(112)}$$

and Eq. (111) becomes

$$\text{div } \mathbf{E} = \frac{\rho}{\varepsilon_0} \quad \textbf{(113)}$$

If we integrate the above equation over a volume *V*, we obtain

$$\int \text{div } E \, dV = \frac{1}{\varepsilon_0} \int \rho \, dV$$

Applying the divergence theorem, we get

$$\oint \mathbf{E} \cdot d\mathbf{a} = \frac{1}{\varepsilon_0} Q \quad \textbf{(114)}$$

[12] Rigorously

$$D_{1n} - D_{2n} = \sigma$$

where σ represents the surface charge density.

[13] More rigorously, $\mathbf{H}_{1t} - \mathbf{H}_{2t}$ is equal to the normal component of the surface current density. However, if there are no surface currents, which is indeed true for most cases, then $H_{1t} = H_{2t}$.

which is simply Gauss' law.[14] That is, the electric flux through a closed surface is the total charge inside the volume divided by ε_0. In a similar manner, the equation

$$\text{div } \mathbf{B} = 0 \tag{115}$$

gives

$$\oint \mathbf{B} \cdot d\mathbf{a} = 0 \tag{116}$$

i.e., the magnetic flux through a closed surface is always zero; this implies the absence of magnetic monopoles.

We next consider the equation

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{117}$$

which associates a space- and time-dependent electric field with a changing magnetic field. Now, Stokes' theorem tells us that

$$\oint_\Gamma \mathbf{E} \cdot d\mathbf{l} = \int_S \text{curl } \mathbf{E} \cdot d\mathbf{a} \tag{118}$$

where the LHS represents a line integral over a closed path Γ and the RHS represents a surface integral over any surface bounding path Γ. Thus

$$\oint_\Gamma \mathbf{E} \cdot d\mathbf{l} = \int_S \text{curl } \mathbf{E} \cdot d\mathbf{a} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a} \tag{119}$$

or

$$\oint_\Gamma \mathbf{E} \cdot d\mathbf{l} = -\frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{a} \tag{120}$$

where in the last step we have used the fact that the surface S is fixed.[15] The LHS of the above equation represents the induced emf in a closed circuit which is equal to the negative of the rate of change of the magnetic flux through the circuit. This is the famous Faraday law of induction; although this law was discovered by Faraday, it was put into differential form [see Eq. (117)] by Maxwell.

We now come to the last of the Maxwell's equations,[16]

$$\text{curl } \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{121}$$

Ampere's law (which was known before Maxwell), when expressed as a differential equation, was of the form[17]

$$\text{curl } \mathbf{H} = \mathbf{J} \tag{122}$$

which implies that a magnetic field is produced only by currents. For example, if we have a long wire carrying a current, we know that it produces a magnetic field. Since the divergence of the curl of any vector is zero, we obtain

$$\text{div } \mathbf{J} = 0 \tag{123}$$

which may be compared with the equation of continuity

$$\text{div } \mathbf{J} + \frac{\partial \rho}{\partial t} = 0 \tag{124}$$

Thus Eq. (122) is valid only when $\partial\rho/\partial t = 0$. Thus, for Ampere's law to be consistent with the equation of continuity,

---

[14] For a dielectric we get

$$\oint \mathbf{D} \cdot d\mathbf{a} = Q$$

where

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}$$

with $\mathbf{P}$ being the dipole moment per unit volume. For a linear homogeneous medium,

$$\mathbf{P} = \chi \mathbf{E}$$

where χ is known as the susceptibility. Thus

$$\mathbf{D} = \varepsilon \mathbf{E}$$

where

$$\varepsilon \equiv \varepsilon_0 + \chi$$

is known as the dielectric permittivity of the medium.

[15] Equation (120) is not valid for a moving system (see, for example, Ref. 3, p. 526).

[16] $\mathbf{H} = \dfrac{\mathbf{B}}{\mu_0} - \mathbf{M}$, where $\mathbf{M}$ is the magnetic moment per unit volume. For a linear material $\mathbf{M} = \chi_m \mathbf{H}$ and therefore $\mathbf{B} = \mu \mathbf{H}$, where $\mu = \mu_0(1 + \chi_m)$.

[17] Once again, it was Maxwell who expressed Ampere's law as a differential equation.

Maxwell argued that there must be an additional term on the RHS of Eq. (122).[18]

The introduction of the term $\partial \mathbf{D}/\partial t$ (which is known as the displacement current) revolutionized physics. Physically it implies that not only does a current produce a magnetic field, but also a changing electric field produces a magnetic field (as indeed happens during the charging and discharging of a condenser).[19] It is the presence of the term $\partial \mathbf{D}/\partial t$ which leads to the wave equation (see Sec. 23.3) and, therefore, the prediction of electromagnetic waves. We can thus argue on physical grounds that a changing electric field produces a magnetic field which varies in space and time, and this changing magnetic field produces an electric field varying in space and time, and so on. This mutual generation of electric and magnetic fields results in the propagation of electromagnetic waves.

Summary

- In a homogeneous dielectric (with dielectric constant ε) Maxwell's equations take the form

$$\text{div } \mathbf{E} = 0$$
$$\text{div } \mathbf{H} = 0$$
$$\text{curl } \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t}$$
$$\text{curl } \mathbf{H} = \varepsilon \frac{\partial \mathbf{E}}{\partial t}$$

  where we have assumed the medium to be nonmagnetic so that $\mu \approx \mu_0 = 4\pi \times 10^{-7}$ N s^2 C^{-2}. The third equation is Faraday's law. The RHS of the fourth equation is known as the displacement current which was introduced by Maxwell; the inclusion of the displacement current term enabled Maxwell to derive the wave equation

$$\nabla^2 \mathbf{E} = \varepsilon\mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2}$$

- In free space $\varepsilon = \varepsilon_0 = 8.854 \times 10^{-12}$ C^2 N^{-1} m^{-2}, and therefore the velocity of the electromagnetic waves in free space is given by

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} \approx 3 \times 10^8 \text{ m s}^{-1}$$

  (The exact value is 2.99792458×10^8 m s^{-1}). Maxwell found that the velocity of the electromagnetic waves was very close to the measured velocity of light and with "faith in rationality of nature" he said that *light is an electromagnetic wave*.

- For an x-polarized electromagnetic wave propagating in the $+z$ direction, we may write

$$\mathbf{E} = \hat{\mathbf{x}} E_0 \cos(kz - \omega t)$$
$$\mathbf{H} = \hat{\mathbf{y}} H_0 \cos(kz - \omega t)$$

  with

$$H_0 = \frac{k}{\omega\mu_0} E_0 \qquad \frac{\omega}{k} = \frac{1}{\sqrt{\varepsilon\mu_0}} = v = \frac{c}{n} \qquad n = \sqrt{\frac{\varepsilon}{\varepsilon_0}}$$

  which represents the refractive index of the dielectric. The corresponding average energy density is given by

$$< u > = \tfrac{1}{2} \varepsilon E_0^2 \qquad \text{J m}^{-3}$$

  and the intensity is given by

$$I = \tfrac{1}{2} \varepsilon v E_0^2 = \tfrac{1}{2} \varepsilon_0 c n E_0^2 \qquad \text{W m}^{-3}$$

  For $I = 10^{15}$ W m^{-2}, $E_0 \approx 0.9 \times 10^9$ V m^{-1}; such high electric field can cause a spark in air.

- The momentum associated with a plane wave is given by

$$\mathbf{p} = \frac{u}{c} \hat{\mathbf{z}}$$

Problems

23.1 On the surface of the Earth we receive about 1.37 kW of energy per square meter from the Sun. Calculate the electric field associated with the sunlight (on the surface of the Earth), assuming that it is essentially monochromatic with $\lambda = 6000$ Å.

[**Ans:** ~1000 V m^{-1}]

23.2 (a) On the surface of the Earth, we receive about 1370 W m^{-2} of energy. Show that the corresponding radiation pressure is about 4.6 μPa (1 Pa $\approx 10^{-5}$ N m^{-2}).

(b) A 100 W sodium lamp ($\lambda \approx 5890$ Å) is assumed to emit waves uniformly in all directions. What is the radiation pressure on a plane mirror at a distance of 10 m from the bulb?

23.3 A 1 kW transmitter is emitting electromagnetic waves (of wavelength 40 m) uniformly in all directions. Calculate the electric field at a distance of 1 km from the transmitter.

[**Ans:** $E_0 \approx 0.25$ V m^{-1}]

23.4 Ocean water can be assumed to be a nonmagnetic dielectric with $\kappa (= \varepsilon/\varepsilon_0) = 80$ and $\sigma = 4.3$ mho m^{-1}. (a) Calculate the frequency at which the penetration depth will be 10 cm.

---

[18] Consequently

$$\text{div curl } \mathbf{H} = 0 = \text{div } \mathbf{J} + \frac{\partial}{\partial t} \text{div } \mathbf{D}$$

or

$$0 = \text{div } \mathbf{J} + \frac{\partial \rho}{\partial t}$$

which is the equation of continuity [we have used Eq. (111)].

[19] For static fields, $\partial \mathbf{D}/\partial t = 0$ and we obtain Ampere's law.

(b) Show that for frequencies less than 10^8 s^{-1}, it can be considered as a good conductor.

[**Ans:** (a) ~6×10^6 s^{-1}]

23.5 For silver one may assume $\mu \approx \mu_0$ and $\sigma \approx 3 \times 10^7$ mho m^{-1}. Calculate the skin depth at 10^8 s^{-1}.

[**Ans:** $\approx 9 \times 10^{-4}$ cm]

23.6 Show that for frequencies $\lesssim 10^8$ s^{-1}, a sample of silicon will act as a good conductor. For silicon one may assume $\varepsilon/\varepsilon_0 \approx 12$ and $\sigma \approx 2$ mho m^{-1}. Also calculate the penetration depth for this sample at $\nu = 10^6$ s^{-1}.

[**Ans:** $\approx 9 \times 10^{-4}$ cm]

23.7 In a conducting medium show that **H** also satisfies an equation similar to Eq. (94).

23.8 Using the analysis given in Sec. 23.7 and assuming $\sigma/\omega\varepsilon << 1$ (which is valid for an insulator), show that

$$\alpha \approx \omega\sqrt{\varepsilon\mu}\left[1+\frac{1}{8}\left(\frac{\sigma}{\omega\varepsilon}\right)^2\right] = \frac{2\pi}{\lambda_0} n\left[1+\frac{1}{8}\left(\frac{\sigma}{\omega\varepsilon}\right)^2\right]$$

and

$$\beta \approx \omega\sqrt{\varepsilon\mu}\left[\frac{1}{2}\left(\frac{\sigma}{\omega\varepsilon}\right)\right] = \frac{2\pi}{\lambda_0} n\left[\frac{1}{2}\left(\frac{\sigma}{\omega\varepsilon}\right)\right]$$

where

$$n = \sqrt{\varepsilon/\varepsilon_0}$$

23.9 For the glass used in a typical optical fiber at $\lambda_0 \approx 8500$ Å, $n = (\varepsilon/\varepsilon_0)^{1/2} = 1.46$, $\sigma \approx 3.4 \times 10^{-6}$ mho m^{-1}. Calculate $\sigma/\omega\varepsilon$ and show that we can use the formulas given in Prob. 23.8. Calculate β and loss in dB km^{-1}. [*Hint:* The power would decrease as $\exp(-2\beta z)$; loss in dB km^{-1} is defined in Sec. 27.8.]

[**Ans:** $\sigma/\omega\varepsilon \approx 8 \times 10^{-11}$;
$\beta \approx 4.3 \times 10^{-4}$ m^{-1}; loss ≈ 3.7 dB km^{-1}]

REFERENCES AND SUGGESTED READINGS

See at the end of Chap. 24.

Chapter Twenty-Four

REFLECTION AND REFRACTION OF ELECTROMAGNETIC WAVES

All of electromagnetism is contained in Maxwell's equations. . . . Untold number of experiments have confirmed Maxwell's equations. If we take away the scaffolding he used to build it, we find that Maxwell's beautiful edifice stands on its own.

—Richard Feynman

24.1 INTRODUCTION

In Chap. 23 we discussed Maxwell's equations and showed the existence of electromagnetic waves. We also showed that at an interface, the tangential components of **E** and **H** and the normal components of **D** and **B** must be continuous. Using these continuity conditions, we will, in this chapter, study the reflection and refraction of plane waves at an interface of two dielectrics (Sec. 24.2) and at an interface of a dielectric and a metal (Sec. 24.3). In Sec. 24.4 we will consider reflectivity (and transmittivity) of a dielectric film.

24.2 REFLECTION AND REFRACTION AT AN INTERFACE OF TWO DIELECTRICS

Let us consider the incidence of a plane polarized electromagnetic wave on an interface of two media; we assume the plane $x = 0$ to represent the interface. Let (ε_1, μ_1) and (ε_2, μ_2) represent the dielectric permittivity and magnetic permeability, respectively, of the media below and above the plane $x = 0$; we will assume both media to be lossless dielectrics, and the case of reflection by a conducting surface will be discussed in Sec. 24.3. Let $\mathbf{E}_1$, $\mathbf{E}_2$, and $\mathbf{E}_3$ denote the electric fields associated with the incident wave, refracted wave, and reflected wave, respectively. For an incident plane wave, these fields will be of the form

$$\begin{aligned} \mathbf{E}_1 &= \mathbf{E}_{10}\, e^{i(\mathbf{k}_1 \cdot \mathbf{r} - \omega t)} \\ \mathbf{E}_2 &= \mathbf{E}_{20}\, e^{i(\mathbf{k}_2 \cdot \mathbf{r} - \omega t)} \\ \mathbf{E}_3 &= \mathbf{E}_{30}\, e^{i(\mathbf{k}_3 \cdot \mathbf{r} - \omega t)} \end{aligned} \tag{1}$$

where $\mathbf{E}_{10}$, $\mathbf{E}_{20}$, and $\mathbf{E}_{30}$ are independent of space and time but may, in general, be complex. The vectors $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$ represent the propagation vectors associated with the incident, refracted, and reflected waves, respectively. Since the fields must satisfy Maxwell's equations, we must have (see Sec. 23.2)

$$\begin{aligned} \mathbf{k}_1^2 &= \omega^2 \varepsilon_1 \mu_1 \\ \mathbf{k}_2^2 &= \omega_2^2 \varepsilon_2 \mu_2 \\ \mathbf{k}_3^2 &= \omega_3^2 \varepsilon_1 \mu_1 \end{aligned} \tag{2}$$

As discussed in Sec. 23.9, the fields have to satisfy certain boundary conditions at the interface (which corresponds to $x = 0$) where Eqs. (1) take the form

$$\begin{aligned} \mathbf{E}_1 &= \mathbf{E}_{10}\, e^{i(k_{1y}\, y + k_{1z}\, z - \omega t)} \\ \mathbf{E}_2 &= \mathbf{E}_{20}\, e^{i(k_{2y}\, y + k_{2z}\, z - \omega_2 t)} \\ \mathbf{E}_3 &= \mathbf{E}_{30}\, e^{i(k_{3y}\, y + k_{3z}\, z - \omega_3 t)} \end{aligned}$$

where k_{1x}, k_{1y}, and k_{1z} represent the x, y, and z components, respectively, of $\mathbf{k}_1$; similarly for $\mathbf{k}_2$ and $\mathbf{k}_3$. Now, for example, the z component of the electric field (which is a tangential component) must be continuous at $x = 0$ for *all* values of y, z, and t. Consequently, the coefficients of y, z, and t in the exponents appearing the above equation must be equal. Thus

$$\omega = \omega_2 = \omega_3 \tag{3}$$

showing that all the waves have the same frequency. Hence Eqs. (2) simplify to

$$k_1^2 = \omega^2 \varepsilon_1 \mu_1 = k_3^2 \tag{4}$$

$$k_2^2 = \omega^2 \varepsilon_2 \mu_2 \tag{5}$$

Further, we must have

$$k_{1y} = k_{2y} = k_{3y} \tag{6}$$

and

$$k_{1z} = k_{2y} = k_{3z} \tag{7}$$

Without any loss of generality we may choose the y axis such that

$$k_{1y} = 0$$

(i.e., $\mathbf{k}_1$ is assumed to lie in the xz plane—see Fig. 24.1). Consequently,

$$k_{2y} = k_{3y} = 0 \tag{8}$$

Equation (8) implies that vectors $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$ will lie in the same plane. Further, from Eq. (7) we get

$$k_1 \sin\theta_1 = k_2 \sin\theta_2 = k_3 \sin\theta_3 \tag{9}$$

Since $k_1 = k_3$ [see Eq. (4)], we must have $\theta_1 = \theta_3$; i.e., the angle of incidence is equal to angle of reflection. Further,

$$\frac{\sin\theta_1}{\sin\theta_2} = \left(\frac{\varepsilon_2\mu_2}{\varepsilon_1\mu_1}\right)^{1/2} \tag{10}$$

If $v_1 \left(=1/\sqrt{\varepsilon_1\mu_1}\right)$ and $v_2 \left(=1/\sqrt{\varepsilon_2\mu_2}\right)$ represent the speeds of propagation of the waves in media 1 and 2, then[1]

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2} = \frac{n_2}{n_1} \tag{11}$$

where $$n_1 \left(= \frac{c}{v_1} = c\sqrt{\varepsilon_1\mu_1}\right)$$

and $$n_2 \left(= \frac{c}{v_2} = c\sqrt{\varepsilon_2\mu_2}\right) \tag{12}$$

represent the refractive indices of media 1 and 2, respectively. Equation (10) is the well known Snell's law.

We will now derive expressions for the reflection and transmission coefficients when a plane polarized wave is incident on an interface of two dielectrics. We will first consider the case when the electric vector lies in the plane of incidence, which will be followed by the case when the electric vector is at right angles to the plane of incidence.

Case 1. E parallel to the plane of incidence: We will assume the electric vector to lie in the plane of incidence as shown in Fig. 24.1. We will show later that if the electric vector associated with the incident wave lies in the plane of incidence, then the electric vectors associated with the reflected and transmitted waves also lie in the plane of incidence. Similarly, if the electric vector associated with the incident wave is normal to the plane of incidence, then the electric vectors associated with the reflected and transmitted waves also lie normal to the plane of incidence—see the discussion just before Example 24.5. The magnetic vectors are along the y axis. Clearly, the z component of the electric field represents a tangential component which should be continuous across the surface. Thus

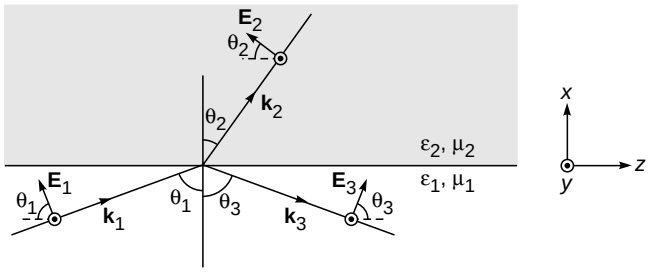


Fig. 24.1 The reflection of a plane wave with its electric vector parallel to the plane of incidence.

$$E_{1z} + E_{3z} = E_{2z}$$

or $$-E_1 \cos\theta_1 + E_3 \cos\theta_1 = -E_2 \cos\theta_2 \tag{13}$$

Thus $\{-E_{10} \exp[i(\mathbf{k}_1 \cdot \mathbf{r} - \omega t)]$

$$+ E_{30} \exp[i(\mathbf{k}_3 \cdot \mathbf{r} - \omega t)]\}_{x=0} \cos\theta_1$$

$$= \{-E_{20} \exp[i(\mathbf{k}_2 \cdot \mathbf{r} - \omega t)]\}_{x=0} \cos\theta_2 \tag{14}$$

Once again, since this condition has to be satisfied at all space points in the plane $x = 0$ and at all times, the exponents must be identically equal which leads to Eqs. (3), (6), and (7). Thus

$$(E_{10} - E_{30}) \cos\theta_1 = E_{20} \cos\theta_2 \tag{15}$$

Further, the normal component of $\mathbf{D}$ must also be continuous, and since $\mathbf{D} = \varepsilon\mathbf{E}$, we must have

$$\varepsilon_1 E_{1x} + \varepsilon_1 E_{3x} = \varepsilon_2 E_{2x}$$

[1] Equation (11) remains valid even when the second medium is anisotropic. As a simple example, if we assume the second medium to be uniaxial with its optic axis along the normal, then for the extraordinary wave we have

$$n_1 \sin\theta_1 = n_{we}(\theta_2) \sin\theta_2$$

where n_{we} is given by Eq. (98) of Chap. 22 with ψ replaced by θ_2; θ_2 represents the direction of $\mathbf{k}_2$, *not* of the ray. The above equation would determine θ_2 (see, e.g., Chap. 3 of Ref. 5).

or

$$\varepsilon_1(E_{10} + E_{30}) \sin\theta_1 = \varepsilon_2 E_{20} \sin\theta_2 \qquad \textbf{(16)}$$

Substituting for E_{20} from Eq. (15), we get

$$\varepsilon_1(E_{10} + E_{30}) \sin\theta_1 = \varepsilon_2 \sin\theta_2 \frac{E_{10} - E_{30}}{\cos\theta_2} \cos\theta_1$$

or

$$(\varepsilon_2 \sin\theta_2 \cos\theta_1 + \varepsilon_1 \sin\theta_1 \cos\theta_2) E_{30} = (\varepsilon_2 \sin\theta_2 \cos\theta_1 - \varepsilon_1 \sin\theta_1 \cos\theta_2) E_{10}$$

Thus
$$r_{\|} = \frac{E_{30}}{E_{10}} = \frac{\varepsilon_2 \sin\theta_2 \cos\theta_1 - \varepsilon_1 \sin\theta_1 \cos\theta_2}{\varepsilon_2 \sin\theta_2 \cos\theta_1 + \varepsilon_1 \sin\theta_1 \cos\theta_2} \qquad \textbf{(17)}$$

where $r_{\|}$ denotes the amplitude reflection coefficient, the subscript $\|$ refers to parallel polarization. If we now divide Eq. (15) by E_{10} and substitute the expression for E_{30}/E_{10} from Eq. (17), we get

$$\left(1 - \frac{\varepsilon_2 \sin\theta_2 \cos\theta_1 - \varepsilon_1 \sin\theta_1 \cos\theta_2}{\varepsilon_2 \sin\theta_2 \cos\theta_1 + \varepsilon_1 \sin\theta_1 \cos\theta_2}\right) \cos\theta_1 = \frac{E_{20}}{E_{10}} \cos\theta_2$$

or

$$t_{\|} = \frac{E_{20}}{E_{10}} = \frac{2\varepsilon_1 \sin\theta_1 \cos\theta_1}{\varepsilon_2 \sin\theta_2 \cos\theta_1 + \varepsilon_1 \sin\theta_1 \cos\theta_2} \qquad \textbf{(18)}$$

where $t_{\|}$ denotes the amplitude transmission coefficient.

To calculate the reflection coefficient, we must determine the ratio of the x components of the Poynting vectors (see Sec. 23.4) associated with the reflected and transmitted waves. The reason why we should take the ratio of the x component can be easily understood by referring to Fig. 24.2. If S_1 denotes the magnitude of the Poynting vector associated with the incident wave, then the energy incident on area dA (on the surface $x = 0$) per unit time is $S_{1x}\, dA = S_1\, dA \cos\theta_1$. Similarly, the energy transmitted through area dA is

$$S_{2x}\, dA = S_2 \cos\theta_2\, dA$$

and the energy reflected from area dA is

$$S_{3x}\, dA = S_3 \cos\theta_1\, dA$$

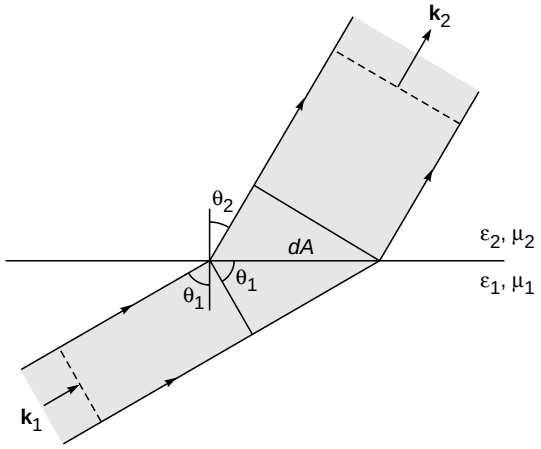


Fig. 24.2 If the cross-sectional area of the incident beam is $dA \cos\theta_1$, then the cross-sectional area of the transmitted beam is $dA \cos\theta_2$, where θ_1 and θ_2 represent the angles of incidence and refraction, respectively.

If $R_{\|}$ and $T_{\|}$ denote the reflection and transmission coefficients, then[2]

$$R_{\|} = \frac{S_{3x}}{S_{1x}} = \frac{S_3 \cos\theta_1}{S_1 \cos\theta_1} \qquad \textbf{(19)}$$

$$= \frac{\langle \mathbf{E}_3 \times \mathbf{H}_3 \rangle}{\langle \mathbf{E}_1 \times \mathbf{H}_1 \rangle} = \frac{\sqrt{\varepsilon_1/\mu_1}\, |E_{30}|^2}{\sqrt{\varepsilon_1/\mu_1}\, |E_{10}|^2} \qquad \text{(see Sec. 20.4)}$$

$$= \left|\frac{E_{30}}{E_{10}}\right|^2$$

or

$$R_{\|} = \left(\frac{\varepsilon_2 \sin\theta_2 \cos\theta_1 - \varepsilon_1 \sin\theta_1 \cos\theta_2}{\varepsilon_2 \sin\theta_2 \cos\theta_1 + \varepsilon_1 \sin\theta_1 \cos\theta_2}\right)^2 \qquad \textbf{(20)}$$

and

$$T_{\|} = \frac{S_{2x}}{S_{1x}} = \frac{S_2 \cos\theta_2}{S_1 \cos\theta_1}$$

$$= \frac{\langle \mathbf{E}_2 \times \mathbf{H}_2 \rangle \cos\theta_2}{\langle \mathbf{E}_1 \times \mathbf{H}_1 \rangle \cos\theta_1}$$

$$= \frac{\sqrt{\varepsilon_2/\mu_2}\, |E_{20}|^2 \cos\theta_2}{\sqrt{\varepsilon_1/\mu_1}\, |E_{10}|^2 \cos\theta_1}$$

$$= \sqrt{\frac{\varepsilon_2}{\varepsilon_1}} \sqrt{\frac{\varepsilon_2}{\varepsilon_1}} \frac{\sin\theta_2}{\sin\theta_1}$$

[2] To calculate the Poynting vector, we must use the real parts of **E** and **H**; see Sec. 23.4.

$$\times \left(\frac{2\varepsilon_1 \sin\theta_1 \cos\theta_1}{\varepsilon_2 \sin\theta_2 \cos\theta_1 + \varepsilon_1 \sin\theta_1 \cos\theta_2} \right)^2 \times \frac{\cos\theta_2}{\cos\theta_1}$$

where we have substituted for $\sqrt{\mu_1/\mu_2}$ from Eq. (10). Thus

$$T_{\parallel} = \frac{4\varepsilon_1\varepsilon_2 \sin\theta_1 \sin\theta_2 \cos\theta_1 \cos\theta_2}{(\varepsilon_2 \sin\theta_2 \cos\theta_1 + \varepsilon_1 \sin\theta_1 \cos\theta_2)^2} \quad \textbf{(21)}$$

It can easily be seen that

$$R_{\parallel} + T_{\parallel} = 1 \quad \textbf{(22)}$$

For nonmagnetic media, $\mu_1 \approx \mu_2 \approx \mu_0 = 4\pi \times 10^{-7}$ N A^{-2}, and the expression for the amplitude reflection coefficient [Eq. (17)] simplifies to[3]

$$r_{\parallel} = \frac{n_2^2 \sin\theta_2 \cos\theta_1 - n_1^2 \sin\theta_1 \cos\theta_2}{n_2^2 \sin\theta_2 \cos\theta_1 + n_1^2 \sin\theta_1 \cos\theta_2}$$

Since

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \quad \textbf{(23)}$$

we get

$$r_{\parallel} = \frac{n_2 \cos\theta_1 - n_1 \cos\theta_2}{n_2 \cos\theta_1 + n_1 \cos\theta_2} \quad \textbf{(24a)}$$

$$= \frac{\sin\theta_1 \cos\theta_1 - \sin\theta_2 \cos\theta_2}{\sin\theta_1 \cos\theta_1 + \sin\theta_2 \cos\theta_2} \quad \textbf{(24b)}$$

or

$$r_{\parallel} = \frac{\sin 2\theta_1 - \sin 2\theta_2}{\sin 2\theta_1 + \sin 2\theta_2}$$

$$= \frac{2\cos(\theta_1 + \theta_2)\sin(\theta_1 - \theta_2)}{2\sin(\theta_1 + \theta_2)\cos(\theta_1 - \theta_2)}$$

$$= \frac{\tan(\theta_1 - \theta_2)}{\tan(\theta_1 + \theta_2)} \quad \textbf{(24c)}$$

Similarly, starting from Eq. (18), we easily obtain

$$t_{\parallel} = \frac{2\cos\theta_1 \sin\theta_2}{\sin(\theta_1 + \theta_2)\cos(\theta_1 - \theta_2)} \quad \textbf{(25)}$$

From Eqs. (24) and (25) we may deduce the following:

(a) *No reflection when $n_2 = n_1$*
When $n_2 = n_1$, $\theta_2 = \theta_1$ and we get

$$r_{\parallel} = 0 \text{ and } t_{\parallel} = 1$$

Thus there is no reflection when the second medium has the same refractive index as the first medium (obviously!). Thus if we have a transparent solid immersed in a liquid of the same refractive index, the solid will not be seen!

(b) *Polarization by reflection: Brewster's law:* If the angle of incidence is such that

$$\theta_1 + \theta_2 = \frac{\pi}{2} \qquad \text{then} \qquad r_{\parallel} = 0$$

i.e., there is no reflected beam; consequently the entire energy will appear in the transmitted beam. But $t_{\parallel} = 2\cos^2\theta_1$. Why? (See Fig. 24.2.) Thus, if an unpolarized beam is incident at an angle such that $\theta_1 + \theta_2 = \pi/2$, then the parallel component of the **E** vector will not be reflected and the reflected light will be polarized with its **E** vector perpendicular to the plane of incidence (see Fig. 24.3). This is the famous *Brewster law*. The corresponding angle of incidence is known as the Brewster angle (or the polarizing angle) and is usually denoted by θ_p.

Notice that the angle of refraction will be $\pi/2 - \theta_p$, and therefore Snell's law takes the form

$$\frac{n_2}{n_1} = \frac{\sin\theta_1}{\sin\theta_2} = \frac{\sin\theta_p}{\sin(\pi/2 - \theta_p)} = \tan\theta_p \quad \textbf{(26)}$$

or

$$\theta_p = \tan^{-1}\left(\frac{n_2}{n_1}\right) \quad \textbf{(27)}$$

Thus, when the angle of incidence is equal to $\tan^{-1}(n_2/n_1)$, then the reflected beam is plane polarized. Further, the transmitted beam is partially polarized. It is easily seen that at the polarizing angle, the reflected ray is at right angles to the refracted ray. In Ref. 4, a beautiful physical argument has been given as to why the reflected light should be linearly

[3] We are using here the fact that for nonmagnetic media

$$n = \frac{c}{v} = \sqrt{\frac{\varepsilon\mu}{\varepsilon_0\mu_0}} \approx \sqrt{\frac{\varepsilon}{\varepsilon_0}}$$

Thus

$$n^2 = \frac{\varepsilon}{\varepsilon_0}$$

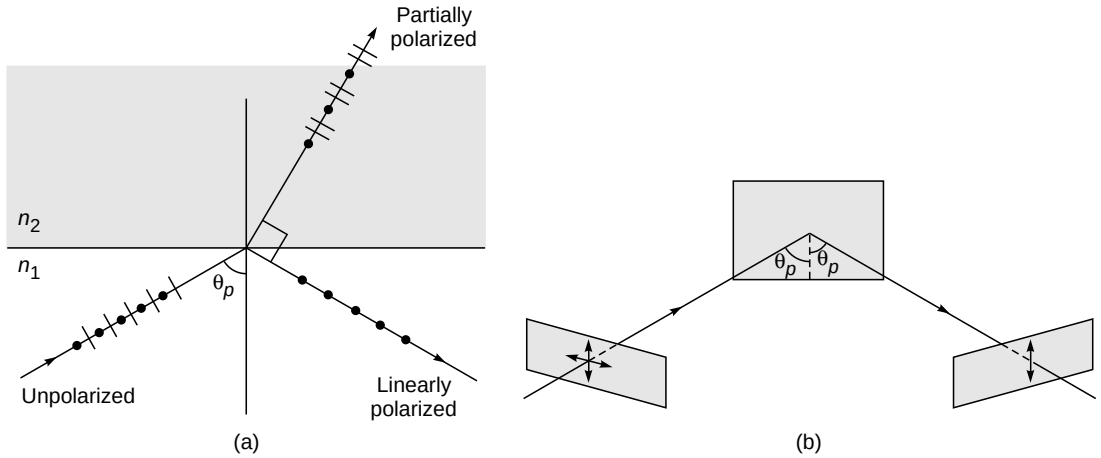


Fig. 24.3 When an unpolarized beam of light is incident on a dielectric at the polarizing angle [i.e., the angle of incidence is equal to $\tan^{-1}(n_2/n_1)$], then the reflected beam is plane polarized with its **E** vector perpendicular to the plane of incidence. The transmitted beam is partially polarized. The dashed line in (b) is normal to the reflecting surface.

polarized when the angle of incidence corresponds to the Brewster angle.

(c) *Phase change on reflection and Stokes' relations:* When light is incident on a denser medium, $\theta_2 < \theta_1$ and for $\theta_1 + \theta_2 > \pi/2$ (i.e., $\theta_1 > \theta_p$), $r_{||}$ is negative, implying a phase change of π. However, no such phase change occurs when $\theta_1 < \theta_p$. We will discuss this point in detail later.

The amplitude reflection and transmission coefficients satisfy Stokes' relations (see Example 24.1).

(d) *Reflection at grazing incidence:* For grazing incidence $(\theta_1 \approx \pi/2)$, Eq. (23) can be written in the form[4]

$$r_{||} = \frac{(\sin\theta_1/\sin\theta_2)\ \sin\alpha_1 - \sin\alpha_2}{(\sin\theta_1/\sin\theta_2)\ \sin\alpha_1 + \sin\alpha_2}$$

$$= \frac{n\sin\alpha_1 - \sin\alpha_2}{n\sin\alpha_1 + \sin\alpha_2} \qquad \textbf{(28)}$$

where $n = n_2/n_1$, $\alpha_1 = \pi/2 - \theta_1$, and $\alpha_2 = \pi/2 - \theta_2$ and at grazing incidence both these angles will be small. Now

$$n = \frac{\sin\theta_1}{\sin\theta_2} = \frac{\cos\alpha_1}{\cos\alpha_2}$$

or

$$\sin\alpha_2 = (1 - \cos^2\alpha_2)^{1/2} = \left(1 - \frac{\cos^2\alpha_1}{n^2}\right)^{1/2}$$

Thus

$$r_{||} = \frac{n\sin\alpha_1 - \left(1 - \dfrac{\cos^2\alpha_1}{n^2}\right)^{1/2}}{n\sin\alpha_1 + \left(1 - \dfrac{\cos^2\alpha_1}{n^2}\right)^{1/2}}$$

$$\approx \frac{n\alpha_1 - \left(1 - \dfrac{1}{n^2}\right)^{1/2}}{n\alpha_1 + \left(1 - \dfrac{1}{n^2}\right)^{1/2}} \qquad \textbf{(29)}$$

where we have replaced $\sin\alpha_1$ by α_1 and $\cos\alpha_1$ by 1 (thus we have retained terms proportional to α_1 but neglected terms of higher order—this will be justified when α_1 is small). Thus

$$r_{||} \approx -\left[1 - \frac{n\alpha_1}{\sqrt{(n^2-1)/n^2}}\right]\left[1 + \frac{n\alpha_1}{\sqrt{(n^2-1)/n^2}}\right]^{-1}$$

$$\approx -\left[1 - \frac{2n^2\alpha_1}{\sqrt{n^2-1}}\right] \to -1 \qquad \text{as } \alpha_1 \to 0 \qquad \textbf{(30)}$$

which shows that the reflection is complete at grazing incidence. The transmission coefficient tends to zero as is indeed obvious from Eq. (25). Thus, if we hold a glass plate horizontally at the level of the eye (see Fig. 24.4), the angle of incidence will be close to $\pi/2$ and the plate will act as a mirror.

[4] The second medium must be a denser medium (i.e., $n_2 > n_1$); otherwise, the beam will undergo total internal reflection [see part (e)].

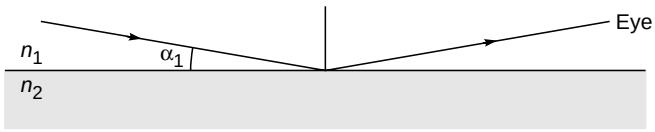


Fig. 24.4 When light is incident at grazing angle (i.e., $\alpha_1 \approx 0$), the reflection is almost complete.

(e) *Total internal reflection:* When an electromagnetic wave is incident on a rarer medium (i.e., $n_2 < n_1$), then $\theta_2 > \theta_1$ and Snell's law [Eq. (12)] can be written in the form

$$\sin\theta_2 = \frac{n_1}{n_2}\sin\theta_1 = \sqrt{\frac{\varepsilon_1}{\varepsilon_2}}\sin\theta_1 \qquad \textbf{(31)}$$

where the media have been assumed to be nonmagnetic, i.e., we have assumed

$$n_1 = \sqrt{\frac{\varepsilon_1}{\varepsilon_0}} \qquad \text{and} \qquad n_2 = \sqrt{\frac{\varepsilon_2}{\varepsilon_0}} \qquad \textbf{(32)}$$

Clearly when

$$\theta_1 > \theta_c \qquad \sin\theta_2 > 1$$

where

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) = \sin^{-1}\sqrt{\frac{\varepsilon_2}{\varepsilon_1}} \qquad \textbf{(33)}$$

The angle θ_c is known as the critical angle. Now, for nonmagnetic media the amplitude reflection coefficient will be given by [see Eq. (17)]

$$r_{\parallel} = \frac{\varepsilon_2\cos\theta_1\sin\theta_2 - \varepsilon_1\sin\theta_1\cos\theta_2}{\varepsilon_2\cos\theta_1\sin\theta_2 + \varepsilon_1\sin\theta_1\cos\theta_2}$$

$$= \frac{\cos\theta_1 - \sqrt{\varepsilon_1/\varepsilon_2}\sqrt{1-\sin^2\theta_2}}{\cos\theta_1 + \sqrt{\varepsilon_1/\varepsilon_2}\sqrt{1-\sin^2\theta_2}}$$

$$= \frac{\cos\theta_1 - (\varepsilon_1/\varepsilon_2)\sqrt{\varepsilon_2/\varepsilon_1 - \sin^2\theta_1}}{\cos\theta_1 + (\varepsilon_1/\varepsilon_2)\sqrt{\varepsilon_2/\varepsilon_1 - \sin^2\theta_1}}$$

or

$$r_{\parallel} = \frac{\cos\theta_1 - (\varepsilon_1/\varepsilon_2)\sqrt{\sin^2\theta_c - \sin^2\theta_1}}{\cos\theta_1 + (\varepsilon_1/\varepsilon_2)\sqrt{\sin^2\theta_c - \sin^2\theta_1}} \qquad \textbf{(34)}$$

where we have used Eqs. (31) and (33). Clearly, for $\theta_1 > \theta_c$ the quantity under the square root becomes negative and we may write

$$\frac{\varepsilon_1}{\varepsilon_2}\sqrt{\sin^2\theta_c - \sin^2\theta_1} = \frac{\varepsilon_1}{\varepsilon_2}\sqrt{\frac{\varepsilon_2}{\varepsilon_1} - \sin^2\theta_1} = i\gamma \qquad \textbf{(35)}$$

where

$$\gamma = \frac{\varepsilon_1}{\varepsilon_2}\sqrt{\sin^2\theta_1 - \frac{\varepsilon_2}{\varepsilon_1}} \qquad \textbf{(36)}$$

is a real number.

Substituting this into Eq. (34), we get

$$r_{\parallel} = \frac{\cos\theta_1 - i\gamma}{\cos\theta_1 + i\gamma} \qquad \textbf{(37)}$$

and the reflection coefficient will be given by

$$R = |r_{\parallel}|^2 = 1 \qquad \textbf{(38)}$$

showing that the entire energy is reflected into the first medium. This is the well-known phenomenon of *total internal reflection.* We may, however, note two points:

1. Since $r_{\parallel}$ is a complex number, there is a phase change on reflection (see Examples 24.3 and 24.6).
2. The amplitude transmission coefficient is given by

$$t_{\parallel} = \frac{2\varepsilon_1\sin\theta_1\cos\theta_1}{\varepsilon_1\sin\theta_1\cos\theta_2 + \varepsilon_2\cos\theta_1\sin\theta_2}$$

which is *not zero*. Thus the field in the rarer medium is not zero (see Example 24.4).

Example 24.1 Figure 24.5 shows that if the media are interchanged, the angles of incidence and refraction are reversed. If $r_{\parallel}'$ and $t_{\parallel}'$ denote the amplitude reflection and transmission coefficients corresponding to Fig. 24.5(b), then show that

$$1 + r_{\parallel}\, r_{\parallel}' = t_{\parallel}\, t_{\parallel}' \qquad \textbf{(39)}$$

(This is one of Stokes' relations—see. Sec. 14.12.)

Solution: Coefficient $r_{\parallel}$ is given by Eq. (17). To calculate $r_{\parallel}'$ in Eq. (17) we replace ε_1 by ε_2, θ_1 by θ_2, and θ_2 by θ_1 (see Fig. 24.5), and we readily obtain

$$r_{\parallel}' = -r_{\parallel} \qquad \textbf{(40)}$$

Thus

$$1 + r_{\parallel}r_{\parallel}' = 1 - r_{\parallel}^2$$

$$= 1 - \frac{(\varepsilon_2\sin\theta_2\cos\theta_1 - \varepsilon_1\sin\theta_1\cos\theta_2)^2}{(\varepsilon_2\sin\theta_2\cos\theta_1 + \varepsilon_1\sin\theta_1\cos\theta_2)^2}$$

$$= \frac{4\varepsilon_1\varepsilon_2\sin\theta_1\cos\theta_1\sin\theta_2\cos\theta_2}{(\varepsilon_2\sin\theta_2\cos\theta_1 + \varepsilon_1\sin\theta_1\cos\theta_2)^2}$$

Now $t_{\parallel}$ is given by Eq. (18); if we make the above-mentioned replacements, we get

$$t_{\parallel}' = \frac{2\varepsilon_2\sin\theta_2\cos\theta_2}{\varepsilon_1\sin\theta_1\cos\theta_2 + \varepsilon_2\sin\theta_2\cos\theta_1} \qquad \textbf{(41)}$$

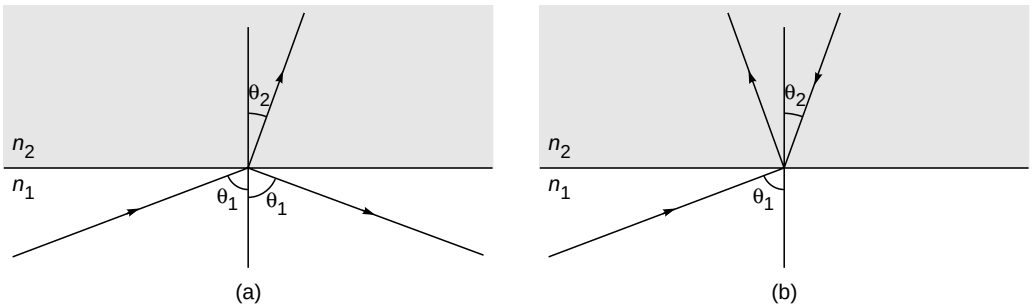


Fig. 24.5 The angles of incidence and refraction are reversed if the media are interchanged.

If we multiply the above expression for $t_{||}'$ by Eq. (18), we readily get Eq. (39).

Example 24.2 In deriving the reflection and transmission coefficients, instead of assuming the continuity of the normal component of **D**, if we assume the continuity of the tangential component of **H**, show that the same results for the reflection and transmission coefficients are obtained.

Solution: It is obvious from Fig. 24.1 that the magnetic field will be in the y direction[5] which represents a tangential component. Thus if $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3$ represent the magnetic fields associated with the incident, transmitted, and reflected waves, respectively, then we may write

$$\begin{aligned}\mathbf{H}_1 &= \hat{\mathbf{y}} H_{10}\, e^{i(\mathbf{k}_1\cdot\mathbf{r}-\omega t)}\\ \mathbf{H}_2 &= \hat{\mathbf{y}} H_{20}\, e^{i(\mathbf{k}_2\cdot\mathbf{r}-\omega t)}\\ \mathbf{H}_3 &= \hat{\mathbf{y}} H_{30}\, e^{i(\mathbf{k}_3\cdot\mathbf{r}-\omega t)}\end{aligned} \tag{42}$$

Continuity of the y component of the field gives

$$H_{10} + H_{30} = H_{20} \tag{43}$$

But

$$\mathbf{H} = \frac{\mathbf{k}\times\mathbf{E}}{\omega\mu} \tag{44}$$

Thus

$$\frac{k_1}{\omega\mu_1}(E_{10} + E_{30}) = \frac{k_2}{\omega\mu_2}E_{20} \tag{45}$$

Continuity of the tangential component of **E** gives [see Eq. (15)]

$$\begin{aligned}(E_{10} - E_{30})\cos\theta_1 &= E_{20}\cos\theta_2\\ &= \frac{k_1}{k_2}\frac{\mu_2}{\mu_1}(E_{10} + E_{30})\cos\theta_2\end{aligned}$$

Thus

$$r_{||} = \frac{E_{30}}{E_{10}}$$

$$= \frac{k_2/\mu_2\cos\theta_1 - k_1/\mu_1\cos\theta_2}{k_2/\mu_2\cos\theta_1 + k_1/\mu_1\cos\theta_2} \tag{46}$$

$$= \frac{\sqrt{\varepsilon_2/\mu_2}\cos\theta_1 - \sqrt{\varepsilon_1/\mu_1}\cos\theta_2}{\sqrt{\varepsilon_2/\mu_2}\cos\theta_1 + \sqrt{\varepsilon_1/\mu_1}\cos\theta_2} \tag{47}$$

If we now use Snell's law, i.e.,

$$\frac{\sin\theta_1}{\sin\theta_2} = \left(\frac{\varepsilon_2\mu_2}{\varepsilon_1\mu_1}\right)^{1/2}$$

we get Eq. (17). However, from Eq. (47) we get the reflection coefficient at normal incidence

$$r_{||} = \frac{\sqrt{\varepsilon_2/\mu_2} - \sqrt{\varepsilon_1/\mu_1}}{\sqrt{\varepsilon_2/\mu_2} + \sqrt{\varepsilon_1/\mu_1}} \approx \frac{n_2 - n_1}{n_2 + n_1} \tag{48}$$

the last relation holds only for nonmagnetic media ($\mu_2 \approx \mu_1 \approx \mu_0$). Thus

$$R = \left|\frac{E_{30}}{E_{10}}\right|^2 = \left(\frac{n_2 - n_1}{n_2 + n_1}\right)^2 \tag{49}$$

For a beam incident from air onto glass $n_1 = 1.0$, and $n_2 = 1.5$, and therefore,

$$R = 0.04 \tag{50}$$

Thus about 4% of the light is reflected and 96% is transmitted into glass.

Example 24.3 Calculate the phase change in the beam which undergoes total internal reflection.

Solution:

$$r_{||} = \frac{\cos\theta_1 - i\gamma}{\cos\theta_1 + i\gamma} \tag{51}$$

$$= \frac{Ae^{-i\phi}}{Ae^{i\phi}} = e^{-2i\phi}$$

where

$$A = (\cos^2\theta_1 + \gamma^2)^{1/2}$$

[5] The vector $\mathbf{E}\times\mathbf{H}$ is along the direction of propagation.

$$\cos\phi = \frac{\cos\theta_1}{(\cos^2\theta_1 + \gamma^2)^{1/2}} \qquad \sin\phi = \frac{\gamma}{(\cos^2\theta_1 + \gamma^2)^{1/2}}$$

Thus

$$E_{30} = E_{10}\, e^{-2i\phi} \tag{52}$$

and the phase change Δ is given by

$$\Delta = 2\phi = 2\tan^{-1}\frac{\gamma}{\cos\theta_1}$$

$$= 2\tan^{-1}\left(\frac{\varepsilon_1}{\varepsilon_2}\frac{\sqrt{\sin^2\theta_1 - \sin^2\theta_c}}{\cos\theta_1}\right) \tag{53}$$

Example 24.4 Determine the nature of the transmitted wave when the beam undergoes total internal reflection.

Solution: The electric field associated with the transmitted wave is given by [see Eq. (1)]

$$\mathbf{E}_2 = \mathbf{E}_{20}\exp[i(\mathbf{k}_2\cdot\mathbf{r} - \omega t)]$$

$$= \mathbf{E}_{20}\exp[i(k_{2x}x + k_{2z}z - \omega t)]$$

$$= \mathbf{E}_{20}\exp[i(k_2\, x\cos\theta_2 + k_2\, z\sin\theta_2 - \omega t)] \tag{54}$$

(see Fig. 24.1). Now

$$\frac{\sin\theta_1}{\sin\theta_2} = \sqrt{\frac{\varepsilon_2}{\varepsilon_1}}$$

therefore $$\sin\theta_2 = \sqrt{\frac{\varepsilon_1}{\varepsilon_2}}\sin\theta_1$$

and

$$\cos\theta_2 = \sqrt{1 - \frac{\varepsilon_1}{\varepsilon_2}\sin^2\theta_1}$$

$$= \sqrt{\frac{\varepsilon_1}{\varepsilon_2}}\sqrt{\frac{\varepsilon_2}{\varepsilon_1} - \sin^2\theta_1}$$

$$= \sqrt{\frac{\varepsilon_2}{\varepsilon_1}}\, i\gamma$$

Thus

$$\mathbf{E}_2 = \mathbf{E}_{20}\, e^{-\beta x}\exp\left\{i\left[\left(k_2\sqrt{\frac{\varepsilon_1}{\varepsilon_2}}\sin\theta_1\right)z - \omega t\right]\right\} \tag{55}$$

where

$$\beta = k_2\sqrt{\frac{\varepsilon_2}{\varepsilon_1}}\,\gamma = \frac{\omega}{c}\sqrt{n_1^2\sin^2\theta_1 - n_2^2} \tag{56}$$

The field given by Eq. (55) represents a wave propagating in the $+z$ direction with an amplitude decreasing exponentially in the x direction. Such a wave is known as a *surface wave* or an *evanescent wave* (see Fig. 24.6). Such waves have many interesting applications.[6]

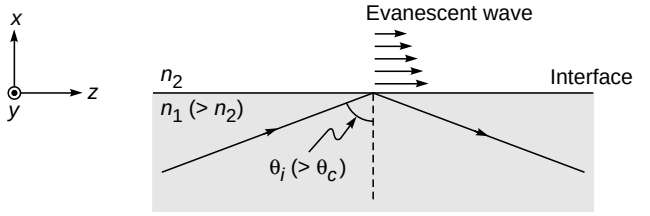


Fig. 24.6 An evanescent wave is generated in the rarer medium when a beam undergoes total internal reflection. The evanescent wave propagates along the z axis, and the amplitude decreases along the x axis.

Case 2. E perpendicular to the plane of incidence: Let us next consider the reflection and refraction of a linearly polarized plane wave with its electric vector perpendicular to the plane of incidence; the reflection is assumed to occur at the interface of two dielectrics. Thus the electric vectors will be along the y axis (see Fig. 24.7) and we may write

$$\mathbf{E}_1 = \hat{\mathbf{y}}E_{10}\exp[i(\mathbf{k}_1\cdot\mathbf{r} - \omega t)]$$
$$\mathbf{E}_2 = \hat{\mathbf{y}}E_{20}\exp[i(\mathbf{k}_2\cdot\mathbf{r} - \omega t)] \tag{57}$$
$$\mathbf{E}_3 = \hat{\mathbf{y}}E_{30}\exp[i(\mathbf{k}_3\cdot\mathbf{r} - \omega t)]$$

where $\mathbf{E}_1$, $\mathbf{E}_2$, and $\mathbf{E}_3$ denote the electric vectors associated with the incident, transmitted, and reflected waves, respectively. Since the y axis is tangential to the interface, the y component of $\mathbf{E}$ must be continuous across the interface; consequently

$$E_{10} + E_{30} = E_{20} \tag{58}$$

The directions of the magnetic fields[7] are also shown in Fig. 24.7; they lie in the plane of incidence and are given by

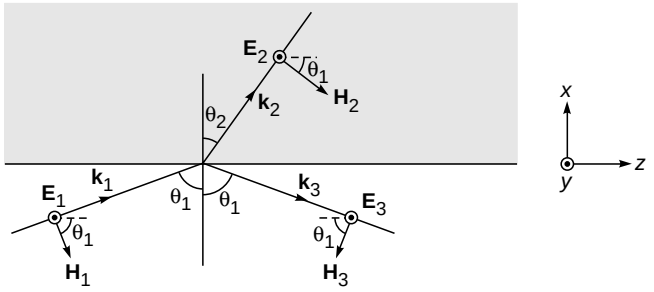


Fig. 24.7 The reflection and refraction of a plane wave with the electric vector lying perpendicular to the plane of incidence.

[6] See, for example, Ref. 2.

[7] Note that since the displacement vector **D** has no component normal to the interface, the continuity of the normal component of **D** will not give us any equation.

$$\begin{aligned}
\mathbf{H}_1 &= \mathbf{H}_{10}\exp[i(\mathbf{k}_1\cdot\mathbf{r}-\omega t)] \\
&= \frac{\mathbf{k}_1\times\mathbf{E}_{10}}{\omega\mu_1}\exp[i(\mathbf{k}_1\cdot\mathbf{r}-\omega t)] \\
\mathbf{H}_2 &= \mathbf{H}_{20}\exp[i(\mathbf{k}_2\cdot\mathbf{r}-\omega t)] \\
&= \frac{\mathbf{k}_2\times\mathbf{E}_{20}}{\omega\mu_2}\exp[i(\mathbf{k}_2\cdot\mathbf{r}-\omega t)] \\
\mathbf{H}_3 &= \mathbf{H}_{30}\exp[i(\mathbf{k}_3\cdot\mathbf{r}-\omega t)] \\
&= \frac{\mathbf{k}_3\times\mathbf{E}_{30}}{\omega\mu_1}\exp[i(\mathbf{k}_3\cdot\mathbf{r}-\omega t)]
\end{aligned} \tag{59}$$

(Notice that **H** lies in the plane of incidence.) Since $\mathbf{k}_1$ is at right angles to $\mathbf{E}_{10}$, the magnitude of $\mathbf{H}_{10}$ is simply $k_1E_{10}/\omega\mu_1$; similarly for $\mathbf{H}_{20}$ and $\mathbf{H}_{30}$. It is obvious from Fig. 24.7 that for the z component of the magnetic field to be continuous, we must have

$$H_{10}\cos\theta_1 - H_{30}\cos\theta_1 = H_{20}\cos\theta_2 \tag{60}$$

or

$$\frac{k_1}{\omega\mu_1}(E_{10}-E_{30})\cos\theta_1 = \frac{k_2}{\omega\mu_2}E_{20}\cos\theta_2 \tag{61}$$

Substituting the expression for E_{20} from Eq. (58), we get

$$\frac{k_1}{\omega\mu_1}(E_{10}-E_{30})\cos\theta_1 = \frac{k_2}{\omega\mu_2}(E_{10}+E_{30})\cos\theta_2$$

Rearranging, we get

$$r_\perp = \frac{E_{30}}{E_{10}}$$

$$= \frac{k_1/\omega\mu_1\cos\theta_1 - k_2/\omega\mu_2\cos\theta_2}{k_1/\omega\mu_1\cos\theta_1 + k_2/\omega\mu_2\cos\theta_2} \tag{62}$$

$$= \frac{\sqrt{\varepsilon_1/\mu_1}\cos\theta_1 - \sqrt{\varepsilon_2/\mu_2}\cos\theta_2}{\sqrt{\varepsilon_1/\mu_1}\cos\theta_1 + \sqrt{\varepsilon_2/\mu_2}\cos\theta_2} \tag{63}$$

$$\approx \frac{\sin\theta_2\cos\theta_1 - \sin\theta_1\cos\theta_2}{\sin\theta_2\cos\theta_1 + \sin\theta_1\cos\theta_2}$$

$$= -\frac{\sin(\theta_1-\theta_2)}{\sin(\theta_1+\theta_2)} \tag{64}$$

Further

$$t_\perp = \frac{E_{20}}{E_{10}} = 1+\frac{E_{30}}{E_{10}}$$

$$= \frac{2\sqrt{\varepsilon_1/\mu_1}\cos\theta_1}{\sqrt{\varepsilon_1/\mu_1}\cos\theta_1 + \sqrt{\varepsilon_2/\mu_2}\cos\theta_2} \tag{65}$$

$$\approx \frac{2\sin\theta_2\cos\theta_1}{\sin(\theta_1+\theta_2)} \tag{66}$$

where the subscript $\perp$ on r and t refers to the state of polarization in which the **E** vector is perpendicular to the plane of incidence. Equations (62), (63), and (65) are exact, whereas Eqs. (64) and (66) are valid for nonmagnetic media. Once again, we can show that when $\theta_1 > \theta_c$, total internal reflection will occur and for grazing incidence the reflection is complete.

As mentioned earlier, the subscript $||$ in r and t refers to polarization parallel to the plane of incidence. Often, instead of the subscript $||$, the subscript p is used; the letter p stands for the word *parallel*. Similarly, the subscript $\perp$ in r and t refers to perpendicular polarization; the subscript is often represented by the subscript s, the letter s stands for the German word *senkrecht* which means perpendicular.[8]

We summarize now the amplitude reflection and transmission coefficients for the two cases; the results are valid for nonmagnetic media:

$$r_p = r_{||} = \frac{n_2\cos\theta_1 - n_1\cos\theta_2}{n_2\cos\theta_1 + n_1\cos\theta_2}$$

$$= \frac{(n_2/n_1)^2\cos\theta_1 - \sqrt{(n_2/n_1)^2-\sin^2\theta_1}}{(n_2/n_1)^2\cos\theta_1 + \sqrt{(n_2/n_1)^2-\sin^2\theta_1}} \tag{67}$$

$$= \frac{\tan(\theta_1-\theta_2)}{\tan(\theta_1+\theta_2)} \tag{68}$$

$$r_s = r_\perp = \frac{n_1\cos\theta_1 - n_2\cos\theta_2}{n_1\cos\theta_1 + n_2\cos\theta_2}$$

$$= \frac{\cos\theta_1 - \sqrt{(n_2/n_1)^2-\sin^2\theta_1}}{\cos\theta_1 + \sqrt{(n_2/n_1)^2-\sin^2\theta_1}} \tag{69}$$

$$= -\frac{\sin(\theta_1-\theta_2)}{\sin(\theta_1+\theta_2)} \tag{70}$$

$$t_p = t_{||} = \frac{2n_1\cos\theta_1}{n_2\cos\theta_1 + n_1\cos\theta_2}$$

$$= \frac{2\cos\theta_1\sin\theta_2}{\sin\theta_1\cos\theta_1 + \sin\theta_2\cos\theta_2} \tag{71}$$

[8] The parallel polarization (or the p polarization) is also called the *transverse magnetic* (or the TM) polarization as the magnetic field is perpendicular to the plane of incidence. On the other hand, the perpendicular polarization (or the s polarization) is also called the *transverse electric* (or the TE) polarization as the electric field is perpendicular to the plane of incidence.

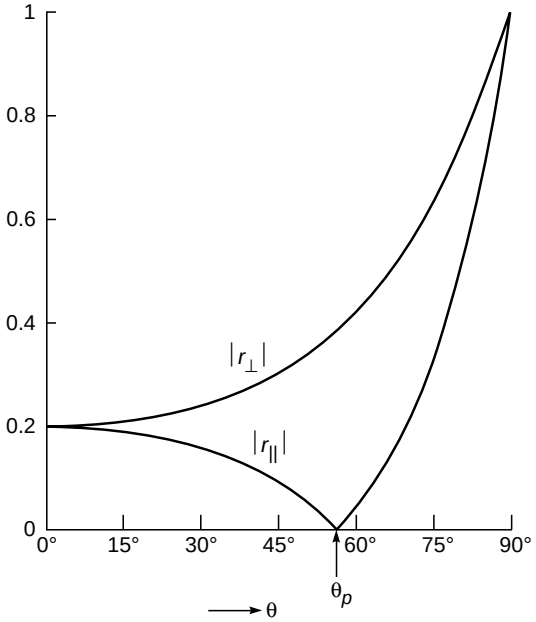


Fig. 24.8 Variation of $|r_{||}|$ and $|r_\perp|$ with the angle of incidence when $n_2 = 1.5$ and $n_1 = 1.0$.

$$t_s = t_\perp = \frac{2n_1 \cos\theta_1}{n_1 \cos\theta_1 + n_2 \cos\theta_2} = \frac{2\cos\theta_1 \sin\theta_2}{\sin(\theta_1 + \theta_2)} \qquad \textbf{(72)}$$

Equation, (67) to (72) are known as the *Fresnel equations*.[9] We write

$$r = |r| e^{i\phi} \qquad \textbf{(73)}$$

The variations of $|r_{||}|$, $|r_\perp|$, $\phi_{||}$, and $\phi_\perp$ are plotted in Figs. 24.8, and 24.9 for $n_2/n_1 = 1.5$. The directions of the **E** vector in the reflected components are shown in Fig. 24.10.

Referring to Fig. 24.8, we note that when

$$\theta_1 = \theta_p = \tan^{-1}\frac{n_2}{n_1} \approx 56° \qquad |r_{||}| = 0$$

This is Brewster's angle. At grazing incidence (i.e., as $\theta_1 \to 90°$), both $|r_{||}|$ and $|r_\perp|$ tend to 1 implying complete reflection. At normal incidence (i.e., $\theta_1 = 0$) any state of polarization can be thought of as parallel polarization or perpendicular polarization,[10] and we should expect $r_{||}$ and $r_\perp$ to give the same result. Figure 24.8 shows that both $|r_{||}|$ and $|r_\perp|$ have the same value; however, Fig. 24.9 shows that whereas the perpendicular component predicts a phase change of π, there is no phase change associated with the parallel component. There is, however, no inconsistency, if we study the direction of the electric vector associated with the reflected component [see Fig. 24.10 (b) and (d)].

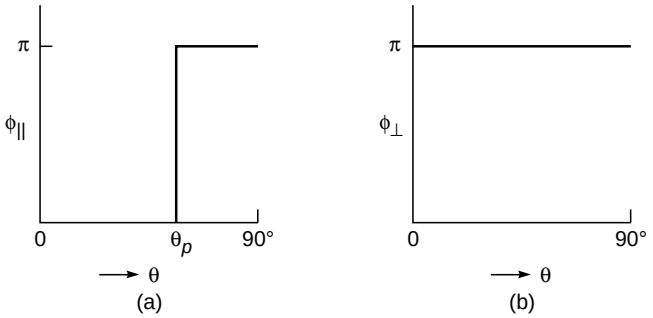


Fig. 24.9 The phase change on reflection (a) for the parallel component and (b) for the perpendicular component for $n_2 = 1.5$ and $n_1 = 1.0$; $\phi_\perp = \pi$ for all values of θ.

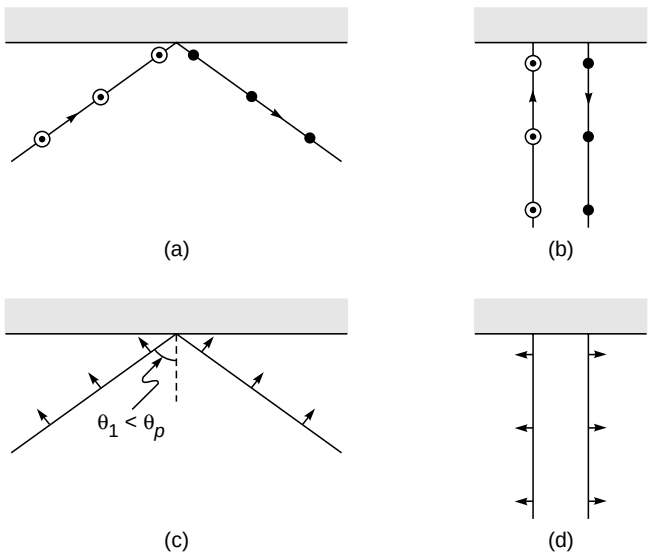


Fig. 24.10 For the perpendicular component, there is a phase change of π at all angles [(a) and (b)]. For the parallel component there is no phase change for $\theta_1 < \theta_p$ [see (c) and (d)]. Notice that at normal incidence, the electric field changes direction in both cases.

We must now recapitulate. We first considered the case when the electric field associated with the incident wave was in the plane of incidence, and we assumed that the electric fields associated with the reflected and transmitted waves were also in the plane of incidence. Had we assumed that the reflection at the interface resulted in electric fields (E_{2y} and E_{3y}) along the y direction associated with the transmitted and reflected waves, then the continuity of E_y and H_z at $x = 0$ would have given

$$E_{3y} = E_{2y}$$

[9] An alternative derivation of Fresnel equations is given in Ref. 4, §33.6.

[10] This is so because at normal incidence the direction of propagation is coincident with the normal to the reflecting surface, and *any* plane containing the normal could be thought of as the plane of incidence.

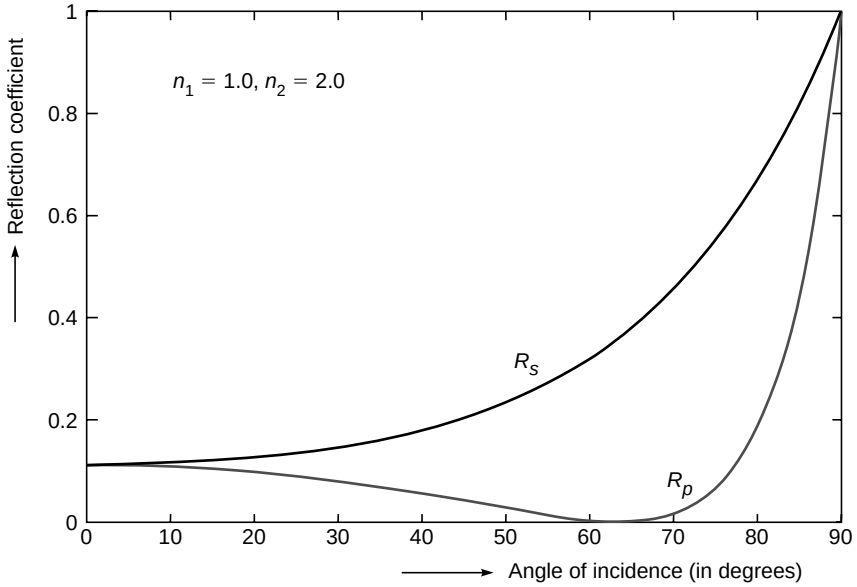


Fig. 24.11 The reflection coefficients for the *p* (parallel) and *s* (perpendicular) components when a light beam is incident from a rarer Medium (of refractive index 1.0) on a denser medium of refractive index 2.0. The Brewster angle is 63.43° where R_p is zero and the reflected wave is *s*-polarized.

and

$$\frac{k_{3x}E_{3y}}{\omega\mu_0} = \frac{k_{2x}E_{2y}}{\omega\mu_0}$$

$$\Rightarrow \qquad -n_1 \cos\theta_1 E_{3y} = n_2 \cos\theta_2 E_{2y}$$

The above two equations would immediately result in the solution $E_{2y} = E_{3y} = 0$. Thus we may conclude that if the incident electric field lies in the plane of incidence, then the electric fields associated with the reflected and transmitted waves must also lie in that plane. Similarly, if the incident electric field is perpendicular to the plane of incidence, then the electric fields associated with the reflected and transmitted waves will also lie perpendicular to the same plane. In general, for an arbitrary state of polarization of the incident wave, we must resolve the incident electric field in components which are parallel and perpendicular to the plane of incidence, consider the reflection (and transmission) of each of the components, and then superpose to find the resultant state of polarization (see Example 24.9). Indeed by studying the polarization characteristics of the reflected wave, we can determine the (complex) refractive index of the material. This is known as the field of *ellipsometry*—a subject of profound importance (see, e.g., Ref. 1). We will consider the reflection by a material of complex refractive index in Sec. 24.3.

In Fig. 24.11 we have plotted the reflection coefficients for the parallel (*p*) and the perpendicular (*s*) components when light is incident from air on a denser medium of refractive index 2.0; notice that $R_p = 0$ at Brewster's angle (or the polarizing angle), showing that, at this angle of incidence, the reflected light will always be *s*-polarized. On the other hand, in Fig. 24.12 we have plotted the reflection coefficients for the parallel (*p*) and the perpendicular (*s*) components when light is incident from a denser medium of refractive index 2.0 on air; notice that both R_s and $R_p = 1$ at all angles of incidence greater than the critical angle. Further, at Brewster's angle $R_p = 0$, showing that, at this angle of incidence, the reflected light will again be *s*-polarized.

Example 24.5 Let us consider the incidence of a plane electromagnetic wave on an air-glass interface (see Fig. 24.1). Thus $n_1 = 1.0$ and $n_2 = 1.5$, giving

$$\theta_p = \tan^{-1}(1.5) \approx 56.31°$$

For $\theta_1 = 30°$; $\theta_2 \approx 19.47°$ we get

$$r_{||} \approx 0.1589 \qquad t_{||} \approx 0.7725$$
$$r_{\perp} \approx -0.2404 \qquad r_{\perp} \approx 0.7596$$

On the other hand, for $\theta_1 = 89°$ (grazing incidence), $\theta_2 = 41.80°$ and

$$r_{||} \approx -0.9321 \quad (\sim 87\%\ \text{reflection}) \quad t_{||} \approx 0.0452$$
$$r_{\perp} \approx -0.9693 \quad \text{and} \quad t_{\perp} \approx 0.0307$$

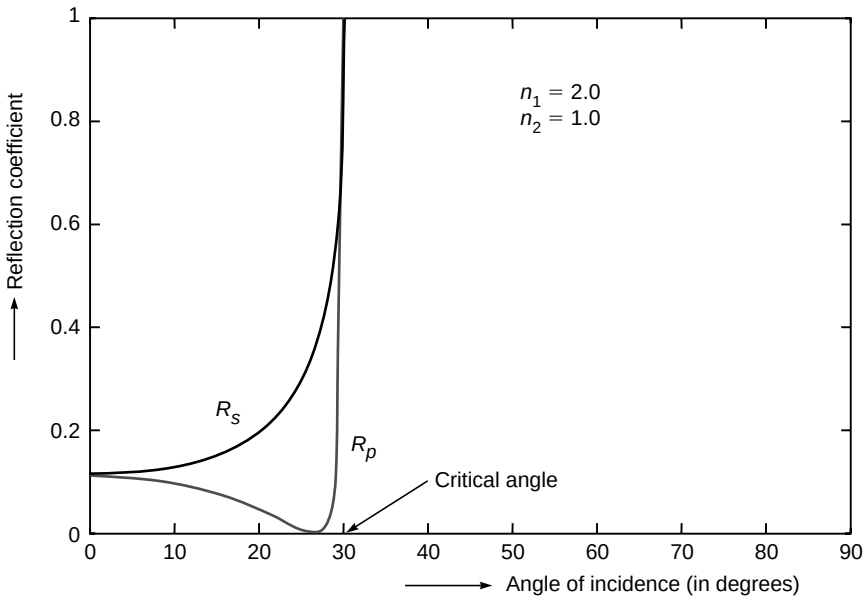


Fig. 24.12 The reflection coefficients for the *p* (parallel) and *s* (perpendicular) components when a light beam is incident from a denser medium (of refractive index 2.0) on a rarer medium of refractive index 1.0. The Brewster angle is 26.56° where R_p is zero and the reflected wave is *s*-polarized. The critical angle is 30° beyond which the reflection coefficient is unity.

Example 24.6 We next consider the incidence of a plane electromagnetic wave on a rarer medium such as a glass-air interface. Thus $n_1 = 1.5$ and $n_2 = 1.0$, giving

$$\theta_p = \tan^{-1}\frac{1}{1.5} \approx 33.69° \qquad \text{and} \qquad \theta_c = \sin^{-1}\frac{1}{1.5} \approx 41.81°$$

1. For $\theta_1 = 30°$, $\theta_2 = 48.59°$ and

$$r_{\|} \approx -0.06788 \qquad t_{\|} \approx +1.3982$$
$$r_{\perp} \approx +0.3252 \qquad t_{\perp} \approx +1.3252$$

2. For $\theta_1 = 60°$, $\cos\theta_2 = i\alpha$ with $\alpha \approx 0.82916$. Thus

$$r_{\|} = \frac{n_2\cos\theta_1 - n_1\cos\theta_2}{n_2\cos\theta_1 + n_1\cos\theta_2}$$
$$= \frac{0.5 - i1.5\alpha}{0.5 + i1.5\alpha}$$
$$\approx -0.7217 - i0.6922$$
$$\approx e^{-0.7567\pi i}$$

[Use of Eq. (53) would give the same result.]

$$t_{\|} = \frac{2n_1\cos\theta_1}{n_2\cos\theta_1 + n_1\cos\theta_2}$$
$$\approx \frac{1.5}{0.5 + i1.5\alpha} \approx 0.41739 - i1.0382$$
$$\approx 1.1190e^{-0.3783\pi i}$$

$$r_{\perp} = \frac{n_1\cos\theta_1 - n_2\cos\theta_2}{n_1\cos\theta_1 + n_2\cos\theta_2}$$
$$\approx \frac{0.75 - i\alpha}{0.75 + i\alpha}$$
$$\approx -0.1 - i0.9950 \approx e^{-0.532\pi i}$$

(Notice that $|r_{\|}| = |r_{\perp}| = 1$.)
and

$$t_{\perp} = \frac{2n_1\cos\theta_1}{n_1\cos\theta_1 + n_2\cos\theta_2}$$
$$\approx 0.9 - i0.995$$
$$\approx 1.3416e^{-0.266\pi i}$$

Example 24.7 Consider a linearly polarized electromagnetic wave (with its electric vector along the *y* direction of magnitude 5 V m^{-1}) propagating in vacuum. It is incident on a dielectric interface at $x = 0$ at an angle of incidence of 30°. The frequency associated with the wave is 6×10^{14} Hz. The refractive index of the dielectric is 1.5. Write the complete expressions for the electric and magnetic fields associated with the incident, reflected, and transmitted waves.

Solution: The wave vector associated with the incident wave is given by

$$\mathbf{k}_1 = (k_0 \cos 30)\hat{\mathbf{x}} + (k_0 \sin 30)\hat{\mathbf{z}}$$
$$= \frac{\sqrt{3}}{2}k_0\hat{\mathbf{x}} + \frac{1}{2}k_0\hat{\mathbf{z}}$$

Thus

$$\mathbf{E}_1 = \hat{\mathbf{y}}\ 5\ \exp\left[i\left(\frac{\sqrt{3}}{2}k_0 x + \frac{1}{2}k_0 z - \omega t\right)\right] \qquad \text{V m}^{-1}$$

where

$$k_0 = \frac{2\pi}{\lambda_0} = 4\pi \times 10^6\ \text{m}^{-1} \qquad \omega = 12\pi \times 10^{14}\ \text{Hz}$$

Now

$$\sin\theta_2 = \frac{n_1 \sin\theta_1}{n_2} = \frac{1}{3} \qquad \Rightarrow \qquad \cos\theta_2 = \frac{\sqrt{8}}{3}$$

Thus

$$r_\perp = \frac{n_1\cos\theta_1 - n_2\cos\theta_2}{n_1\cos\theta_1 + n_2\cos\theta_2} = -0.2404$$

$$\Rightarrow \qquad R_s = R_\perp = 0.057796$$

and

$$t_\perp = \frac{2\cos\theta_1 \sin\theta_2}{\sin(\theta_1+\theta_2)} = 0.7596$$

implying

$$T_s = T_\perp = \frac{n_2\cos\theta_2}{n_1\cos\theta_1}\ |t_\perp|^2 = 0.942204$$

showing that $R_\perp + T_\perp = 1$. Now

$$\mathbf{k}_2 = \hat{\mathbf{x}}(n_2 k_0 \cos\theta_2) + \hat{\mathbf{z}}(n_2 k_0 \sin\theta_2)$$
$$= \hat{\mathbf{x}}(\sqrt{2}\,k_0) + \hat{\mathbf{z}}\left(\tfrac{1}{2}k_0\right)$$

and

$$\mathbf{k}_3 = -\hat{\mathbf{x}}k_0 \cos\theta_1 + \hat{\mathbf{z}}(k_0 \sin\theta_1)$$
$$= -\hat{\mathbf{x}}\left(\frac{\sqrt{3}}{2}k_0\right) + \hat{\mathbf{z}}\left(\frac{1}{2}k_0\right)$$

Thus the electric fields associated with the transmitted and reflected waves are given by

$$\mathbf{E}_2 = 3.8\hat{\mathbf{y}}\ \exp\left[i\left(\sqrt{2}\,k_0 x + \frac{1}{2}k_0 z - \omega t\right)\right] \qquad \text{V/m}^{-1}$$

and

$$\mathbf{E}_3 = -1.2\hat{\mathbf{y}}\ \exp\left[i\left(-\frac{\sqrt{3}}{2}k_0 x + \frac{1}{2}k_0 z - \omega t\right)\right] \qquad \text{V/m}^{-1}$$

respectively. Notice that the values of k_z in $\mathbf{E}_1$, $\mathbf{E}_2$, and $\mathbf{E}_3$ are the same [see Eq. (7)]. The corresponding magnetic fields can be calculated by using Eq. (59) to obtain

$$\mathbf{H}_1 = 5\ \frac{k_1}{\omega\mu_0}\ (-\hat{\mathbf{x}}\ \sin\theta_1 + \hat{\mathbf{z}}\ \cos\theta_1)$$
$$\times \exp\left[i\left(\frac{\sqrt{3}}{2}k_0 x + \frac{1}{2}k_0 z - \omega t\right)\right]$$

$$\mathbf{H}_2 = 3.8\ \frac{k_2}{\omega\mu_0}\ (-\hat{\mathbf{x}}\ \sin\theta_2 + \hat{\mathbf{z}}\ \cos\theta_2)$$
$$\times \exp\left[i\left(\sqrt{2}\,k_0 x + \frac{1}{2}k_0 z - \omega t\right)\right]$$

and

$$\mathbf{H}_3 = -1.2\ \frac{k_1}{\omega\mu_0}\ (-\hat{\mathbf{x}}\ \sin\theta_1 - \hat{\mathbf{z}}\ \cos\theta_1)$$
$$\times \exp\left[i\left(-\frac{\sqrt{3}}{2}k_0 x + \frac{1}{2}k_0 z - \omega t\right)\right]$$

where

$$\frac{k_1}{\omega\mu_0} = \frac{k_0}{\omega\mu_0} = \frac{1}{c\mu_0} = \frac{1}{120\pi} \text{ mks units}$$

and

$$\frac{k_2}{\omega\mu_0} = \frac{k_0 n_2}{\omega\mu_0} = \frac{n_2}{c\mu_0} = \frac{1}{80\pi} \text{ mks units}$$

Example 24.8 Consider once again the situation described in the above example except that the magnetic vector is now along the *y* direction. Formulate the complete expressions for the electric fields associated with the incident, reflected, and transmitted waves.

Solution: Referring to Fig. 24.1, we have

$$\mathbf{E}_1 = 5\left(\frac{1}{2}\hat{\mathbf{x}} - \frac{\sqrt{3}}{2}\hat{\mathbf{z}}\right)$$
$$\times \exp\left[i\left(\frac{\sqrt{3}}{2}k_0 x + \frac{1}{2}k_0 z - \omega t\right)\right] \qquad \text{V m}^{-1}$$

Now

$$r_{||} = \frac{n_2\cos\theta_1 - n_1\cos\theta_2}{n_2\cos\theta_1 + n_1\cos\theta_2} = 0.1589$$

$$\Rightarrow \qquad R_{||} = 0.02525$$

and

$$t_{||} = \frac{2n_1\cos\theta_1}{n_2\cos\theta_1 + n_1\cos\theta_2} = 0.7726$$

implying

$$T_{||} = \frac{n_2\cos\theta_2}{n_1\cos\theta_1}\,|\,t_{||}|^2 = 0.97475$$

showing that $R_{||} + T_{||} = 1$. Furthermore,

$$\mathbf{E}_2 = 3.863\left(\frac{1}{3}\hat{\mathbf{x}} - \frac{\sqrt{8}}{3}\hat{\mathbf{z}}\right)$$
$$\times \exp\left[i\left(\sqrt{2}\,k_0 x + \frac{1}{2}k_0 z - \omega t\right)\right] \qquad \text{V m}^{-1}$$

$$\mathbf{E}_3 = 0.7945\left(\frac{1}{2}\hat{\mathbf{x}} + \frac{\sqrt{3}}{2}\hat{\mathbf{z}}\right)$$
$$\times \exp\left[i\left(-\frac{\sqrt{3}}{2}k_0 x + \frac{1}{2}k_0 z - \omega t\right)\right] \qquad \text{V m}^{-1}$$

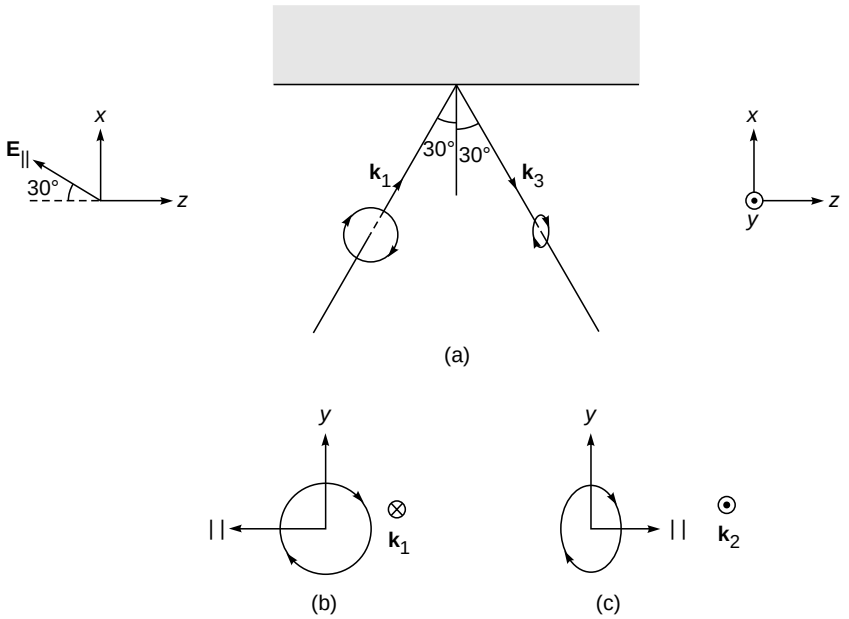


Fig. 24.13 (a) A right circularly polarized beam is incident on an air-glass interface at 30°. The reflected beam is left elliptically polarized. (b) The direction of rotation of the **E** vector for the incident wave. The direction of propagation (shown as ⊗) is into the page. (c) The direction of rotation of the **E** vector for the reflected wave. The direction of propagation (shown as ⊙) is coming out of the page.

Example 24.9 For the situation described in Example 24.7, consider a right circularly polarized wave incident at the air-glass interface at $\theta_1 = 30°$. Determine the state of polarization of the reflected and transmitted fields.

Solution: We refer to Fig. 24.13. We must resolve the electric field in components parallel and perpendicular to the plane of incidence. Neglecting the space-dependent parts, if we write for the y component of the incident field

$$E_\perp = E_y = E_0 \cos \omega t$$

then for the beam to be right circularly polarized, the parallel component must be given by

$$E_{||} = E_0 \cos\left(\omega t + \frac{\pi}{2}\right) = -E_0 \sin \omega t$$

The direction of the "parallel axis" is as shown in Fig. 24.13(b) consistent with Fig. 24.1. Thus

$$E_x = E_{||} \sin \theta_1 = -E_0 \sin \theta_1 \sin \omega t$$

and

$$E_z = -E_{||} \cos \theta_1 = +E_0 \cos \theta_1 \sin \omega t$$

In the reflected field, the "parallel component" will be along the direction shown in Fig. 24.13(c), consistent with Fig. 24.1. Now, associated with the reflected wave

$$E_y = E_\perp = r_\perp E_0 \cos \omega t \simeq -0.24 E_0 \cos \omega t$$

and

$$E_{||} = -r_{||} E_0 \sin \omega t \simeq -0.16 E_0 \sin \omega t$$

If we now refer to Fig. 24.13(c), the electric vector will rotate in the clockwise direction, and since the propagation is out of the page, the reflected wave is left elliptically polarized. We can carry out a similar analysis for the transmitted wave to show that it is right elliptically polarized.

24.3 REFLECTION BY A CONDUCTING MEDIUM

If we consider a plane wave incident from a dielectric onto a conducting medium (with conductivity σ), the expressions for E_{20}/E_{10} and E_{30}/E_{10} will remain the same, except for the fact that k_2 will now be complex (sec Sec. 23.7). Snell's law

$$k_1 \sin \theta_1 = k_2 \sin \theta_2 \qquad \textbf{(74)}$$

will also remain valid, but since k_2 is complex, $\sin \theta_2$ will also be complex.

If we consider the case when the electric fields are perpendicular to the plane of incidence (see Fig. 24.7), we have [see Eq. (62)]

$$r_\perp = \frac{E_{30}}{E_{10}}$$

$$= \frac{(k_1/\omega\mu_1)\cos\theta_1 - (k_2/\omega\mu_2)\cos\theta_2}{(k_1/\omega\mu_1)\cos\theta_1 + (k_2/\omega\mu_2)\cos\theta_2} \quad (75)$$

where

$$k_2 = \alpha + i\beta \quad (76)$$

$$\alpha = \omega\sqrt{\varepsilon_2\mu_2}\left[\frac{1}{2} + \frac{1}{2}\sqrt{1+\left(\frac{\sigma}{\omega\varepsilon_2}\right)^2}\right]^{1/2}$$

$$\beta = \frac{\omega\sigma\mu_2}{2\alpha} \quad (77)$$

(see Sec. 23.7). For normal incidence

$$r_\perp = \frac{E_{30}}{E_{10}} = \frac{1 - \dfrac{\alpha+i\beta}{\omega\sqrt{\varepsilon_1\mu_1}}\dfrac{\mu_1}{\mu_2}}{1 + \dfrac{\alpha+i\beta}{\omega\sqrt{\varepsilon_1\mu_1}}\dfrac{\mu_1}{\mu_2}} \quad (78)$$

and

$$t_\perp = \frac{E_{20}}{E_{10}} = 1 + r_\perp = \frac{2}{1 + \dfrac{\alpha+i\beta}{\omega\sqrt{\varepsilon_1\mu_1}}\dfrac{\mu_1}{\mu_2}} \quad (79)$$

For a good conductor, $\sigma/\varepsilon\omega >> 1$ and

$$\alpha \approx \beta \approx \left(\frac{\omega\sigma\mu_2}{2}\right)^{1/2} \quad (80)$$

Thus

$$r_\perp = \frac{E_{30}}{E_{10}} \simeq \frac{1-(1+i)\Delta}{1+(1+i)\Delta} \quad (81)$$

$$t_\perp = \frac{E_{20}}{E_{10}} \simeq \frac{2}{1+(1+i)\Delta} \quad (82)$$

where

$$\Delta = \left(\frac{\sigma\mu_1}{2\mu_2\varepsilon_1\omega}\right)^{1/2} \quad (83)$$

For infinite conductivity, $\Delta \to \infty$ and

$$E_{30} = -E_{10} \qquad E_{20} = 0 \quad (84)$$

showing that there is a phase change of π on reflection. Further, the energy is completely reflected, and the field inside the conductor is identically zero.[11] For a finite (but large) value of σ, an approximate expression for the reflection coefficient can be obtained in the following manner:

$$R = \left|\frac{E_{30}}{E_{10}}\right|^2 = \left|-\frac{1-1/(1+i)\Delta}{1+1/(1+i)\Delta}\right|^2$$

$$\approx \left|\left(1-\frac{1}{(1+i)\Delta}\right)\left(1-\frac{1}{(1+i)\Delta}\right)\right|^2$$

$$\approx \left|1-\frac{2}{(1+i)\Delta}\right|^2 \approx 1-\frac{2}{\Delta}$$

$$\approx 1 - 2\left(\frac{2\mu_2\varepsilon_1\omega}{\sigma\mu_1}\right)^{1/2} \quad (85)$$

For nonmagnetic media, with

$$\varepsilon_1 \approx \varepsilon_0 = 8.854\times10^{-12}\ \mathrm{C^2\ N^{-1}\ m^{-2}}$$

$$\omega \approx 2\pi\times10^{10}\ \mathrm{s^{-1}} \qquad \sigma \approx 3\times10^7\ \mathrm{mho\ m^{-1}}\ \text{(silver)}$$

$$R \approx 0.9996$$

Thus about 99.96% of light is reflected. This is the reason why metals are such good reflectors. Notice that the reflection coefficient increases with decrease in frequency.

When the incidence is not normal, one must substitute the following expression for $\cos\theta_2$.

$$\cos\theta_2 = \sqrt{1-\sin^2\theta_2} = \sqrt{1-(k_1/k_2)^2\sin^2\theta_1}$$

$$= \left[1 - \frac{\omega^2\varepsilon_1\mu_1}{(\alpha+i\beta)^2}\sin^2\theta_1\right]^{1/2}$$

$$\approx 1 - \frac{1}{2}\frac{\omega^2\varepsilon_1\mu_1\sin^2\theta_1}{(\omega^2\sigma^2\mu_2^2/4)(1+i)^2}$$

$$\approx 1$$

The last expression is valid for good conductors. Thus for the transmitted wave we can write

$$\begin{aligned}\mathbf{E}_2 &= \mathbf{E}_{20}\exp[i(\mathbf{k}_2\cdot\mathbf{r}-\omega t)]\\ &= \mathbf{E}_{20}\exp[i(k_2\cos\theta_2\,x + k_2\sin\theta_2\,z - \omega t)]\\ &= \mathbf{E}_{20}\exp\{i[\alpha(1+i)x + k_1\sin\theta_1\,z - \omega t]\}\\ &\approx \mathbf{E}_{20}\exp(-\alpha x)\exp[i(\alpha x + k_1\sin\theta_1\,z - \omega t)]\end{aligned} \quad (86)$$

For a good conductor, $\alpha >> k_1$, and the wave (having an amplitude exponentially decreasing in the x direction) propagates along the x axis.

[11] Note that if $\sigma \to \infty$ (i.e., for a perfect conductor), then $r_{||} \to +1$ and $r_\perp \to -1$ [see Eqs. (46) and (75)] even for nonnormal incidence. Thus, if a right circularly polarized wave is incident on a perfect conductor, then the reflected light will be left circularly polarized.

24.4 REFLECTIVITY OF A DIELECTRIC FILM

In this section we will calculate the reflectivity of a dielectric film for a plane wave incident normally on it. We will determine the thickness of the film for which the film will become antireflecting and compare our results with those obtained in Sec. 15.4. In Prob. 24.9, we will apply our results to a Fabry–Perot interferometer (cf. Sec. 16.2).

We consider a plane wave incident normally on a dielectric film of thickness d (see Fig. 24.14). Without any loss of generality, we assume the electric field to be along the y axis. Thus the electric fields in media 1, 2, and 3 are given by

$$\begin{aligned}
\mathbf{E}_1 &= \hat{\mathbf{y}} E_{10}^{+} e^{i(k_1 x - \omega t)} + \hat{\mathbf{y}}\, E_{10}^{-} e^{-i(k_1 x + \omega t)} \\
\mathbf{E}_2 &= \hat{\mathbf{y}} E_{20}^{+} e^{i(k_2 x - \omega t)} + \hat{\mathbf{y}}\, E_{20}^{-} e^{-i(k_2 x + \omega t)} \\
\mathbf{E}_3 &= \hat{\mathbf{y}} E_{30}^{+} e^{i[k_3 (x-d) - \omega t]}
\end{aligned} \tag{87}$$

where E_{10}^{+} and E_{10}^{-} represent the amplitudes of the forward and backward propagating waves, respectively, in region 1; similarly for other fields. Since the third medium extends to infinity, there is no backward propagating wave in region 3. For $\mathbf{E}_3$, for the sake of convenience we have introduced a phase factor of exp ($-ik_3 d$); this term makes the analysis more straightforward.

The corresponding magnetic field is given by [see Eq. (66) of Chap. 23]

$$\mathbf{H} = \frac{\mathbf{k} \times \mathbf{E}}{\omega \mu} \tag{88}$$

where $\mathbf{k} = \begin{cases} k\hat{\mathbf{x}} & \text{for waves propagating in } +x \text{ direction} \\ -k\hat{\mathbf{x}} & \text{for waves propagating in } +x \text{ direction} \end{cases}$

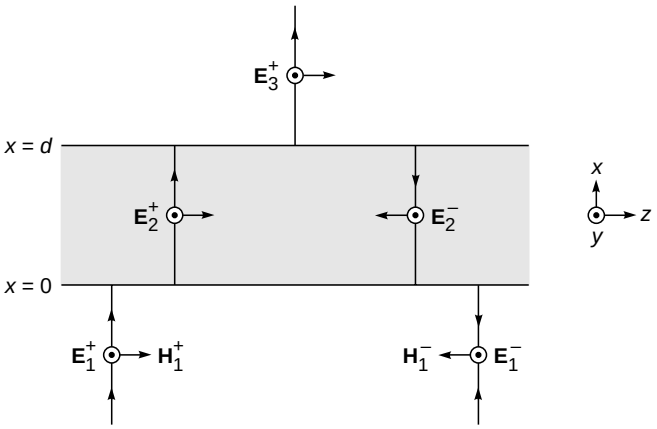


Fig. 24.14 Reflection of a plane wave incident normally on a dielectric slab of thickness d.

Thus

$$\begin{aligned}
\mathbf{H}_1 &= \hat{\mathbf{z}} \frac{k_1}{\omega \mu_0} \left(E_{10}^{+} e^{i(k_1 x - \omega t)} - E_{10}^{-} e^{-i(k_1 x + \omega t)}\right) \\
\mathbf{H}_2 &= \hat{\mathbf{z}} \frac{k_2}{\omega \mu_0} \left(E_{20}^{+} e^{i(k_2 x - \omega t)} - E_{20}^{-} e^{-i(k_2 x + \omega t)}\right) \\
\mathbf{H}_3 &= \hat{\mathbf{z}} \frac{k_3}{\omega \mu_0} E_{30}^{+} e^{i[k_3 (x-d) - \omega t]}
\end{aligned} \tag{89}$$

Both E_y and H_z represent tangential components and should therefore be continuous at interfaces $x = 0$ and $x = d$. The continuity conditions at $x = 0$ give

$$E_{10}^{+} + E_{10}^{-} = E_{20}^{+} + E_{20}^{-}$$

and

$$\frac{k_1}{\omega \mu_0} (E_{10}^{+} - E_{10}^{-}) = \frac{k_2}{\omega \mu_0} (E_{20}^{+} - E_{20}^{-})$$

or

$$E_{10}^{+} - E_{10}^{-} = \frac{n_2}{n_1} (E_{20}^{+} - E_{20}^{-})$$

where we have used the relations

$$k_2 = \frac{\omega}{c} n_2 \qquad \text{and} \qquad k_1 = \frac{\omega}{c} n_1$$

Simple manipulations give

$$\begin{pmatrix} E_{10}^{+} \\ E_{10}^{-} \end{pmatrix} = \begin{pmatrix} \dfrac{n_1 + n_2}{2n_1} & \dfrac{n_1 - n_2}{2n_1} \\ \dfrac{n_1 - n_2}{2n_1} & \dfrac{n_1 + n_2}{2n_1} \end{pmatrix} \begin{pmatrix} E_{20}^{+} \\ E_{20}^{-} \end{pmatrix} \tag{90}$$

Similarly, the continuity of E_y and H_z at $x = d$ gives

$$\begin{aligned}
E_{20}^{+} e^{i\delta} + E_{20}^{-} e^{-i\delta} &= E_{30}^{+} \\
E_{20}^{+} e^{i\delta} - E_{20}^{-} e^{-i\delta} &= \frac{n_3}{n_2} E_{30}^{+}
\end{aligned}$$

where $\delta = k_2 d$. Elementary manipulations give

$$\begin{pmatrix} E_{20}^{+} \\ E_{20}^{-} \end{pmatrix} = \begin{pmatrix} \dfrac{n_2 + n_3}{2n_2} e^{-i\delta} \\ \dfrac{n_2 - n_3}{2n_2} e^{i\delta} \end{pmatrix} E_{30}^{+} \tag{91}$$

Combining Eqs. (90) and (91), we get

$$\begin{aligned}
E_{10}^{+} = \Bigg[&\left(\frac{n_1 + n_2}{2n_1}\right)\left(\frac{n_2 + n_3}{2n_2} e^{-i\delta}\right) \\
&+ \left(\frac{n_1 - n_2}{2n_1}\right)\left(\frac{n_2 - n_3}{2n_2} e^{i\delta}\right) \Bigg] E_{30}^{+}
\end{aligned} \tag{92}$$

and

$$E_{10}^{-} = \left[\left(\frac{n_1 - n_2}{2n_1}\right)\left(\frac{n_2 + n_3}{2n_2}\, e^{-i\delta}\right) + \left(\frac{n_1 + n_2}{2n_1}\right)\left(\frac{n_2 - n_3}{2n_2}\, e^{i\delta}\right)\right] E_{30}^{+} \quad \textbf{(93)}$$

Dividing Eq. (92) by Eq. (93), we get the amplitude reflection coefficient

$$r = \frac{E_{10}^{-}}{E_{10}^{+}} = \frac{r_1 e^{-i\delta} + r_2 e^{i\delta}}{e^{-i\delta} + r_1 r_2 e^{i\delta}} \quad \textbf{(94)}$$

where

$$r_1 = \frac{n_1 - n_2}{n_1 + n_2} \quad \textbf{(95)}$$

and

$$r_2 = \frac{n_2 - n_3}{n_2 + n_3} \quad \textbf{(96)}$$

represent the Fresnel reflection coefficients at the first and second interfaces, respectively. The reflectivity is therefore given by

$$R = |r|^2 = \frac{r_1^2 + r_2^2 + 2r_1 r_2 \cos 2\delta}{1 + r_1^2 r_2^2 + 2 r_1 r_2 \cos 2\delta} \quad \textbf{(97)}$$

In Secs. 16.2 to 16.4, we discussed the above equation in detail with $r_2 = r_1$; however, the definition of δ here differs by a factor of 2 from the definition of δ in Chap. 16 (see Prob. 24.10). A more general analysis shows that the above equation remains valid even for oblique incidence with δ now equal to $k_2 d \cos\theta_2$, with θ_2 being the angle of refraction in the second medium and r_1 and r_2 representing the appropriate Fresnel reflection coefficients corresponding to the particular angle of incidence and state of polarization.

Summary

- Consider the incidence of a linearly polarized electromagnetic wave on an interface of two dielectrics (which we assume to be $x = 0$); the x z plane is assumed to be the plane of incidence. Let $n_1 \left(= \sqrt{\varepsilon_1/\varepsilon_0}\right)$ and $n_2 \left(= \sqrt{\varepsilon_2/\varepsilon_0}\right)$ be the refractive indices of the two media. The incident wave, refracted wave, and reflected waves can be written as

$$\mathbf{E}_1 = \mathbf{E}_{10} \exp[i(\mathbf{k}_1 \cdot \mathbf{r} - \omega t)] \quad \text{incident wave}$$
$$\mathbf{E}_2 = \mathbf{E}_{20} \exp[i(\mathbf{k}_2 \cdot \mathbf{r} - \omega t)] \quad \text{refracted wave}$$
$$\mathbf{E}_3 = \mathbf{E}_{30} \exp[i(\mathbf{k}_3 \cdot \mathbf{r} - \omega t)] \quad \text{reflected wave}$$

where $\mathbf{E}_{10}$, $\mathbf{E}_{20}$, and $\mathbf{E}_{30}$ are independent of space and time and

$$k_1 = \frac{\omega}{c}\, n_1 = k_3 \qquad k_2 = \frac{\omega}{c}\, n_2$$

$$k_1 \sin\theta_1 = k_2 \sin\theta_2 = k_3 \sin\theta_3$$

where θ_1, θ_2, and θ_3 are the angle of incidence, angle of refraction, and angle of reflection, respectively. The above equations readily give

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \qquad \text{Snell's law}$$

and $\theta_1 = \theta_3$.

- For $\mathbf{E}_1$ lying in the xz plane (which is the plane of incidence)

$$\mathbf{E}_{10} = E_{10}(\hat{\mathbf{x}} \sin\theta_1 - \hat{\mathbf{z}} \cos\theta_1)$$
$$\mathbf{E}_{20} = t_{||} E_{10}(\hat{\mathbf{x}} \sin\theta_2 - \hat{\mathbf{z}} \cos\theta_2)$$
$$\mathbf{E}_{30} = r_{||} E_{10}(\hat{\mathbf{x}} \sin\theta_1 + \hat{\mathbf{z}} \cos\theta_1)$$
$$r_{||} = \frac{n_2 \cos\theta_1 - n_1 \cos\theta_2}{n_2 \cos\theta_1 + n_1 \cos\theta_2} = \frac{\tan(\theta_1 - \theta_2)}{\tan(\theta_1 + \theta_2)}$$
$$t_{||} = \frac{2n_1 \cos\theta_1}{n_2 \cos\theta_1 + n_1 \cos\theta_2} = \frac{2\cos\theta_1 \sin\theta_2}{\sin(\theta_1 + \theta_2)\cos(\theta_1 - \theta_2)}$$

Notice that $r_{||} = 0$ when $\theta_1 + \theta_2 = \pi/2$, implying

$$\theta_1 = \theta_p = \tan^{-1}\left(\frac{n_2}{n_1}\right)$$

This is Brewster's angle.

- For $\mathbf{E}_1$ perpendicular to the plane of incidence (i.e., along $\hat{\mathbf{y}}$),

$$\mathbf{E}_{10} = E_{10}\hat{\mathbf{y}} \qquad \mathbf{E}_{20} = t_{\perp} E_{10}\hat{\mathbf{y}} \qquad \mathbf{E}_{30} = r_{\perp} E_{10}\hat{\mathbf{y}}$$

with

$$r_{\perp} = \frac{n_1 \cos\theta_1 - n_2 \cos\theta_2}{n_1 \cos\theta_1 + n_2 \cos\theta_2} = -\frac{\sin(\theta_1 - \theta_2)}{\sin(\theta_1 + \theta_2)}$$

and

$$t_{\perp} = \frac{2n_1 \cos\theta_1}{n_1 \cos\theta_1 + n_2 \cos\theta_2} = \frac{2\sin\theta_2 \cos\theta_1}{\sin(\theta_1 + \theta_2)}$$

- In both cases, if $n_2 < n_1$ and $\theta_1 > \theta_c = \sin^{-1}(n_2/n_1)$, we have total internal reflection. We can still use the above expressions for $r_{||}$, $t_{||}$, $r_{\perp}$, and $t_{\perp}$ but we must remember that

$$\sin\theta_2 = \frac{n_1}{n_2} \sin\theta_1 > 1$$

and $$\cos\theta_2 = \sqrt{1 - \sin^2\theta_2} = i\alpha$$

will be pure imaginary. Thus $r_{||}$, $t_{||}$, $r_{\perp}$, and $t_{\perp}$ will be complex quantities with $|r_{||}| = 1 = |r_{\perp}|$, showing that the entire energy is reflected; however, there will be an evanescent wave in the second medium whose field will decay along the x axis and propagate along the z axis.

Problems

24.1 Show that in the limit of $\theta_1 \to 0$ (i.e., at normal incidence) the reflection coefficient is the same for parallel and perpendicular polarizations.

24.2 Consider a magnetic dielectric with a permeability such that $\mu/\mu_0 = \varepsilon/\varepsilon_0$. Show that for such a material the reflection coefficient for normal incidence is identically equal to zero. This realization is equivalent to the situation where the impedance is matched at the junction of two transmission lines.
(The quantity $\sqrt{\mu/\varepsilon}$ can be considered as the intrinsic impedance of the medium.)

24.3 A right circularly polarized beam is incident on a perfect conductor at 45°. Show that the reflected beam is left circularly polarized.

24.4 Assume $n_1 = 1.5$ and $n_2 = 1.0$ (see Example 24.6).

(a) For $\theta_1 = 45°$ show that

$$r_{||} = +\,0.28 - i0.96 \qquad t_{||} = 1.92 - i1.44$$

Similarly, calculate $r_\perp$ and $t_\perp$.

(b) On the other hand, for $\theta_1 = 33.69°$ show that

$$r_{||} = 0 \qquad t_{||} = 1.5$$
$$r_\perp = +0.3846 \qquad t_\perp = 1.3846$$

[**Ans:** (a) $r_\perp = 0.8 - 0.6i$; $t_\perp = 1.8 - 0.6i$]

24.5 Consider a right circularly polarized beam incident on a medium of refractive index 1.6 at an angle of 60°. Calculate $r_{||}$ and $r_\perp$ and show that the reflected beam is right elliptically polarized with its major axis much longer than its minor axis. What will happen at 58°?

[**Ans:** $r_{||} = -0.0249$, $r_\perp = -0.4581$]

24.6 Consider a *y*-polarized wave incident on a glass-air interface ($n_1 = 1.5$, $n_2 = 1.0$) at $\theta_1 = 45°$ and at $\theta_1 = 80°$. Write the complete expressions for the transmitted field, and show that in the latter case it is an evanescent wave with depth of penetration ($= 1/\beta$) equal to about 8.8×10^{-8} m; assume $\lambda = 6000$ Å.

24.7 For gold, at $\lambda_0 = 6530$ Å the complex refractive index is given by $n_2 = 0.166 + 3.15i$. Calculate k_2 and show that the reflectivity at normal incidence is approximately 94%. [*Hint:* Use Eq. (75) directly.] On the other hand, at $\lambda_0 = 4000$ Å, $n_2 = 1.658 + 1.956i$; show that the reflectivity is only 39%.

24.8 Show that for $\delta = 0$, Eq. (97) takes the form

$$R = \left(\frac{n_1 - n_3}{n_1 + n_3}\right)^2 \qquad \textbf{(98)}$$

as it indeed should be.

24.9 Using the various equations in Sec. 24.4, calculate the transmittivity and show that

$$T = \frac{\frac{1}{2} n_3 |E_3^+|^2}{\frac{1}{2} n_1 |E_1^+|^2} = 1 - R$$

24.10 Assume the third medium in Fig. 24.12 to be identical to the first medium; i.e., $n_3 = n_1$. Thus

$$r_2 = -r_1 = -\frac{n_1 - n_2}{n_1 + n_2}$$

Using Eq. (97), show that

$$R = \frac{F \sin^2 \delta}{1 + F \sin^2 \delta} \qquad \textbf{(99)}$$

where $$F = \frac{4r_1^2}{(1 - r_1^2)^2} \qquad \textbf{(100)}$$

is called the coefficient of finesse. Equation (99) is identical to the result derived in Sec. 16.2 while discussing the theory of the Fabry–Perot interferometer. The definition of δ here differs by a factor of 2 from the definition of δ in Chap. 16.

24.11 When the angle of incidence is equal to the Brewster's angle, show that $T_{||}$ [as given by Eq. (21)] is equal to unity.

REFERENCES AND SUGGESTED READINGS

1. J. M. Bennett and H. E. Bennett, "Polarization," *Handbook of Optics*, Ed. W. J. Driscoll, McGraw-Hill, New York, 1978.
2. O. Bryngdahl, "Evanescent Waves in Optical Imaging," *Progress in Optics*, Ed. E. Wolf, Vol. XI, North-Holland, Amsterdam, 1973.
3. D. R. Corson and P. Lorrain, *Introduction to Electromagnetic Fields and Waves*, W. H. Freeman and Co., San Francisco, 1962.
4. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley, Reading, Mass., 1964.
5. A. Ghatak, and K. Thyagarajan, *Optical Electronics*, Cambridge University Press, 1989. Reprinted by Foundation Books, New Delhi.
6. J. R. Heirtzler, "The Largest Electromagnetic Waves," *Scientific American*, Vol. 206, p.128, September 1962.
7. E. C. Jordon and K. G. Balmain, *Electromagnetic Waves and Radiating Systems*, Prentice-Hall, Englewood Cliffs, N.J., 1970.
8. W. K. H. Panofsky and M. Phillips, *Classical Electricity and Magnetism*, Addison-Wesley, Reading, Mass., 1962.
9. J. R. Reitz and F. J. Milford, *Foundations of Electromagnetic Theory*, Addison-Wesley, Reading, Mass., 1962.
10. H. S. Sandhu and G. B. Friendmann, "Change of Phase on Reflection," *American Journal of Physics*, Vol. 39, p. 388, 1971.

PART 6
Photons

This part consists of only one chapter, namely, Chap. 25 on the particle model of radiation. The photoelectric effect (discovered by Hertz in 1888) had certain peculiarities which cannot be explained on the basis of wave theory. In 1905, Einstein provided a simple explanation of the peculiarities by assuming that light consisted of quanta of energy $h\nu$ (where ν is the frequency) and that the emission of a photoelectron was the result of the interaction of a single quantum (i.e., of the photon) with an electron. For this Einstein received the 1921 Nobel Prize in Physics. Chapter 25 also discusses the Compton effect (for which Prof. Compton received the 1927 Nobel Prize in Physics) which established that the photon has a momentum equal to $h\nu/c$.

Chapter Twenty-Five

THE PARTICLE NATURE OF RADIATION

Are not the rays of light very small bodies emitted from shining substance?

—Isaac Newton, *Opticks*[1]

It is undeniable that there is an extensive group of data concerning radiation which shows that light has certain fundamental properties that can be understood much more readily from the standpoint of the Newton emission (particle) theory than from the standpoint of the wave theory. It is my opinion, therefore, that the next phase of the development of theoretical physics will bring us a theory of light that can be interpreted as a kind of fusion of the wave and emission theories.

—Albert Einstein (1909)[2]

Important Milestones

1887	*Heinrich Hertz, while receiving the electromagnetic waves in a coil with a spark gap, found that the maximum spark length was reduced when the apparatus was put in a black box.*
1897	*J. J. Thomson discovered the electron.*
1899	*J. J. Thomson showed that electrons are emitted when light falls on a metal surface; these are known as photoelectrons.*
1900	*To derive the blackbody radiation formula, Planck made a drastic assumption that the oscillators can only assume discrete energies.*
1902	*Philip Lenard observed that the kinetic energy of the emitted photoelectrons was independent of the intensity of the incident light and that the energy of the emitted electron increased when the frequency of the incident light was increased.*
1905	*In a paper entitled "On a Heuristic Point of View about the Creation and Conversion of Light," Einstein introduced the light quanta. In this paper, he wrote that for an explanation of phenomena such as blackbody radiation, production of electrons by ultraviolet light (which is the photoelectric effect), it is necessary to assume that when a light ray starting from a point is propagated, the energy is not continuously distributed over an ever-increasing volume. Rather it consists of a finite number of energy quanta, localized in space, which move without being divided and which can be absorbed or emitted only as a whole. Einstein received the 1921 Nobel Prize in Physics for his services to theoretical physics, and especially for his explanation of the photoelectric effect.*
1923	*Compton reported his studies on the scattering of X-rays by solid materials (mainly graphite) and showed that the shift of the wavelength of the scattered photon could be explained by assuming the photon having momentum equal to h/λ. Compton received the 1927 Nobel Prize in Physics for his discovery of the effect named after him.*
1926	*Gilbert Lewis, a U.S. chemist, coined the word* photon *to describe Einstein's localized energy quanta.*

[1] The author found this quotation in Ref. 1.

[2] The author found this quotation in Ref. 2.

25.1 INTRODUCTION

In earlier chapters, we discussed the interference, diffraction, and polarization of light. All these phenomena can be explained satisfactorily on the basis of the wave theory of light. We also discussed the electromagnetic character of light waves (see Chaps. 22 and 23) and showed that the electromagnetic theory can be successfully used to explain the origin of the refractive index (see Chap. 7), the phenomenon of double refraction (see Chap. 22), and many other experimental results. However, there exist a large number of experimental phenomena that can be explained only on the basis of the corpuscular nature of radiation. In this chapter, we will discuss the famous experiments on the photoelectric effect and the Compton effect which establish the particle nature of light—a wave model is totally inadequate to explain these effects. In Chap. 2, we briefly discussed how to reconcile the dual nature of radiation (i.e., the wave and the particle aspects) on the basis of the quantum theory.

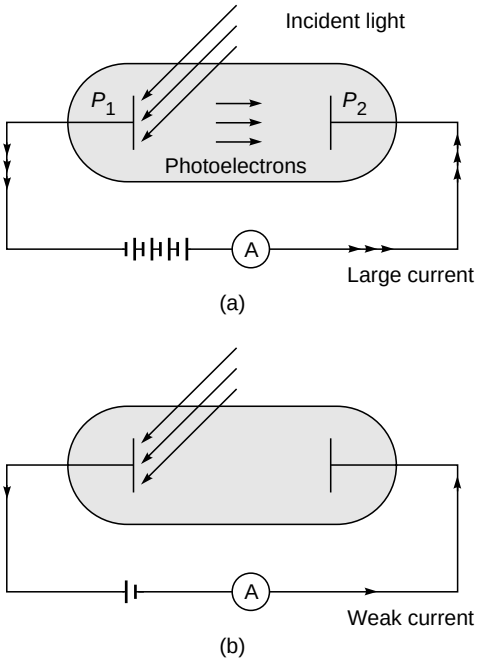


Fig. 25.1 If light (of a certain frequency) is allowed to fall on a metal such as sodium, electrons are emitted which can be collected by plate P_2. (a) and (b) correspond to positive and negative voltage applied to plate P_2. Even when the plate is kept at a low negative voltage, one can detect a small current.

25.2 THE PHOTOELECTRIC EFFECT

In 1887 Heinrich Hertz, while carrying out his experiments on electromagnetic waves, found that if the light emitted from one spark gap were blocked, it would reduce the maximum spark length in the other gap. After carrying out a series of experiments, he concluded that it was the ultraviolet radiation from the first spark that was helping the spark across the second gap. Hertz reported the observations but did not pursue further and also did not make any attempt to explain them. In 1897, J. J. Thomson discovered electrons, and in 1899, he showed that electrons are emitted when light falls on a metal surface; these are now known as photoelectrons. In 1902, Philip Lenard observed that (1) the kinetic energy of the emitted electrons was independent of the intensity of the incident light and (2) the energy of the emitted electron increased when the frequency of the incident light was increased. Later Millikan carried out very careful experiments on the photoelectric effect, and the apparatus that he used was similar to the one shown in Fig. 25.1; these photoelectrons constitute a current between plates P_1 and P_2 which can be detected by means of an ammeter A. When the voltage across the plates is varied, the current also varies; typical variations of the current with voltage are shown in Fig. 25.2. The figure corresponds to monochromatic light of a particular wavelength, and different curves correspond to different intensities of the beam. From the figure we can draw the following conclusions:

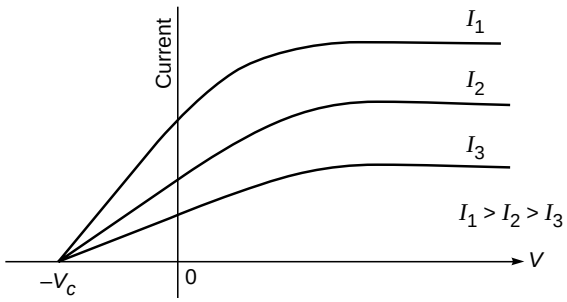


Fig. 25.2 Typical variation of the photocurrent with voltage. The curves correspond to light (of the same frequency) having different intensities.

1. At zero voltage there is a finite value of the current, implying that some of the emitted photoelectrons reach the metal surface P_2.
2. As the voltage is increased, the current increases until it reaches a saturation value; this will happen when plate P_2 collects all the emitted photoelectrons.
3. If plate P_2 is kept at a slightly negative potential, there is a weak current, implying that some of the photoelectrons do manage to reach plate P_2. However,

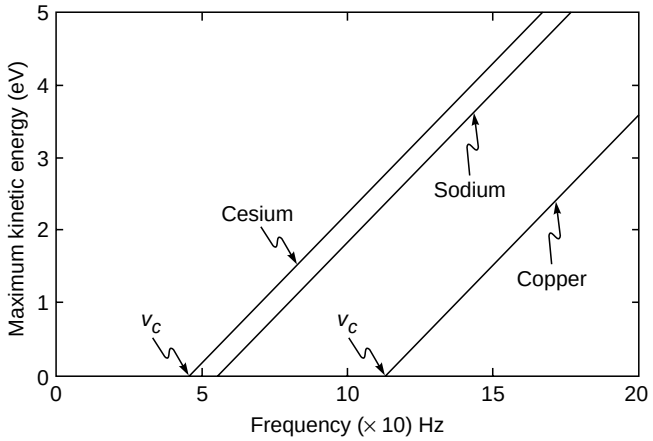


Fig. 25.3 The variation of the maximum kinetic energy of the electrons as a function of frequency of the incident light.

beyond a certain voltage (which is shown as $-V_c$ in the figure) the current is zero; V_c is known as the cutoff voltage, and the quantity $|q|V_c$ will represent the maximum kinetic energy of the photoelectrons (q represents the charge of the electron). For example, for sodium $V_c \approx 2.3$ V and for copper $V_c \approx 4.7$ V.

4. If we do not change the wavelength of the incident radiation but make it more intense, the magnitude of the current will become larger as shown in Fig. 25.2, implying a greater emission of photoelectrons. Notice that the value of the cutoff potential remains the same; this important result implies that the maximum kinetic energy of the emitted photoelectrons does not depend on the intensity of the incident radiation.
5. If the frequency of the incident radiation is increased, then the cutoff potential and hence the maximum kinetic energy of the electron ($= |q|V_c$) vary linearly with the frequency as shown in Fig. 25.3. Further, for frequencies less than a critical value (shown as ν_c in Fig. 25.3), there is *no* emission of photoelectrons no matter what the intensity of the incident radiation may be.

At first sight it appears that since electromagnetic waves carry energy, the wave model for light should be able to explain the emission of photoelectrons from a metal surface. However, there are certain peculiarities associated with the photoelectric effect which cannot be satisfactorily explained by means of a wave model:

1. The first peculiarity is the fact that the maximum kinetic energy of the electrons does not depend on the intensity of the incident radiation; it depends on only its frequency; further, a greater intensity leads to a larger number of electrons, constituting a larger current. Thus, a faint violet light would eject electrons of greater kinetic energy than an intense yellow light although the latter would produce a large number of electrons. A wave model would, however, predict that a large intensity of the incident radiation would result in a greater kinetic energy of the emitted electrons.
2. The second peculiarity is the fact that there is almost no time lag between the times of incidence of the radiation and the ejection of the photoelectron. For weak intensities of the incident beam, the wave theory predicts considerable time lag for the electrons to absorb enough energy to leave the metal surface. This can be illustrated by considering a specific example. One can observe a detectable photocurrent if the surface of sodium metal is illuminated by violet light of intensity as low as 10^{-10} W cm^{-2}. Now, 10 layers of sodium will contain about

$$\frac{6\times10^{23}\times10\times10^{-8}}{23} \approx 2\times10^{15} \text{ atoms per cm}^2$$

where we have assumed the density of sodium to be ≈ 1 g cm^{-3}. Assuming that the energy is uniformly absorbed by the upper 10 layers of sodium, each atom would receive energy at the rate of

$$\frac{10^{-10}}{2\times10^{15}} \approx 5\times10^{-26} \text{ J s}^{-1} \approx 3\times10^{-7} \text{ eV s}^{-1}$$

Assuming that an electron should acquire an energy of ~ 1 eV to escape from the metal, we should expect a time lag of order 10^7 s ($\sim$ few months). However, the experiments show that there is no detectable time lag between the incidence of the radiation and the emission of the photoelectrons. Indeed, in 1928, Lawrence and Beams devised an experiment to find out whether the time lag was $\leq 3\times10^{-9}$ s; the experiment gave a negative result.

In 1905, Einstein provided a simple explanation of the above-mentioned peculiarities. He argued that light consisted of quanta of energy $h\nu$ (where ν is the frequency) and that the emission of a photoelectron was the result of the interaction of a single quantum (i.e., of the photon) with an electron. In his 1905 paper (Ref. 3), Einstein wrote:

> Monocromatic radiation behaves as if it consists of mutually independent energy quanta of magnitude *[hν]*.

Einstein's theory gives a very satisfactory explanation of the photoelectric effect. According to this theory, a light beam (of frequency ν) essentially consists of individual

corpuscles called photons; the word *photon* was coined in 1926 by Gilbert Lewis, a U.S. chemist, to describe Einstein's *localized energy quanta*. Each photon carries an energy equal to $h\nu$. This corpuscular model can explain all the observations discussed above. Thus, for all frequencies below the cutoff ν_c, each photon will carry energy less than $h\nu_c$ which will not be sufficient to eject the electron from the metal. For $\nu > \nu_c$, a major fraction of the excess energy [$= h(\nu - \nu_c)$] appears as kinetic energy of the emitted electron. Further, the nonmeasurable time lag between the incidence of the radiation and the ejection of the electron follows immediately from the corpuscular nature of the radiation. Indeed, the observed maximum kinetic energy of the photoelectrons is linearly related to the frequency of the incident radiation, and one may write (see Fig. 25.3)

$$T_{\max} = -B + h\nu = h(\nu - \nu_c) \qquad \textbf{(1)}$$

where B ($= h\nu_c$) is a constant and h is Planck's constant ($= 6.627 \times 10^{-27}$ erg s). The frequency ν_c represents the cutoff frequency and is a characteristic of the metal. For example,

$$\text{For cesium } B \approx 1.9\,\text{eV} \quad \Rightarrow \quad \nu_c \approx 4.6 \times 10^{14}\,\text{Hz}$$
$$\text{For sodium } B \approx 2.3\,\text{eV} \quad \Rightarrow \quad \nu_c \approx 5.6 \times 10^{14}\,\text{Hz}$$
$$\text{For copper } B \approx 4.7\,\text{eV} \quad \Rightarrow \quad \nu_c \approx 11.4 \times 10^{14}\,\text{Hz}$$

In Fig. 25.3, ν_c is the intercept on the horizontal axis. In 1909, Einstein wrote (Ref. 1)

> It is undeniable that there is an extensive group of data concerning radiation which shows that light has certain fundamental properties that can be understood much more readily from the standpoint of the Newton emission (particle) theory than from the standpoint of the wave theory. It is my opinion, therefore, that the next phase of the development of theoretical physics will bring us a theory of light that can be interpreted as a kind of fusion of the wave and emission theories.

We may note the prediction of Einstein. Einstein received the 1921 Noble Prize in Physics for his discovery of the law of photoelectric effect. To quote Max Jammer (Ref. 4),

> Owing to Einstein's paper of 1905, it was primarily the photoelectric effect to which physicists referred as an irrefutable demonstration of the existence of photons and which thus played an important part in the conceptual development of quantum mechanics.

The validity of Eq. (1) was established in a series of beautiful experiments by Millikan who also made the first direct determination of Planck's constant h. In his Nobel lecture, Millikan (Ref. 5) said

> After ten years of testing and changing and learning and sometimes blundering, all efforts being directed from the first toward the accurate experimental measurement of the energies of emission of photoelectrons, now as a function of temperature, now of wavelength, now of material (contact e.m.f. relation), this work resulted, contrary to my own expectation, in the first direct experimental proof in 1914 of the exact validity, within narrow limits of experimental error, of the Einstein equation (Eq. (1)), and the first direct photoelectric determination of Planck's constant h.

Millikan further wrote:

> Einstein's equation is one of exact validity (always within the present small limits of experimental error) and of very general applicability, is perhaps the most conspicuous achievement of Experimental Physics during the past decade.

After Millikan's experiments, Duane and his associates found unambiguous proof of a relation which is just the inverse of Einstein's. They bombarded a metal target with electrons of known and constant energy and found that the maximum frequency of the emitted X-rays was given, with great precision, by

$$\tfrac{1}{2} m v^2 = h\nu \qquad \textbf{(2)}$$

In making this transition from Planck's *quantized oscillators* to *quanta of radiation*, Einstein made a very important conceptual transition; namely, he introduced the idea of corpuscular behaviour of radiation. Although Newton had described light as a stream of particles, this view had been completely superseded by the wave picture of light, a picture that culminated in the electromagnetic theory of Maxwell. The revival of the particle picture now posed a severe conceptual problem, one of reconciling wave- and particlelike behavior of radiation. It also soon became apparent that matter also exhibited wave-particle duality. For example, an electron with an accurately measured value of mass and charge could undergo diffraction in a manner similar to that of light waves—this led to the development of the uncertainty principle and quantum theory.

In the next section we will discuss a very important experiment carried out by Arthur Compton; this experiment could be explained by assuming the photon having momentum equal to $h\nu/c$.

25.3 THE COMPTON EFFECT

We have seen that Einstein's explanation for the photoelectric effect implies that quanta of light (photons) carry a definite amount of energy. The Compton effect provided an

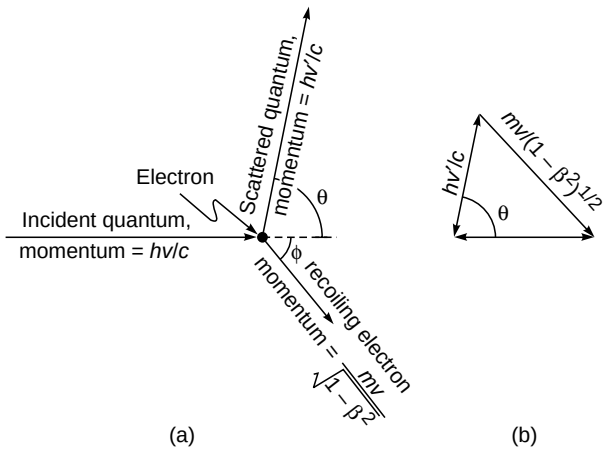


Fig. 25.4 The Compton scattering of a photon: the figure shows the incidence of a photon (of frequency ν) on an electron; the scattered photon (having a reduced frequency ν′) propagating along the direction which making an angle θ with the original direction; the electron also acquires a momentum. Figure adapted from the original paper of Arthur Compton (Ref. 9).

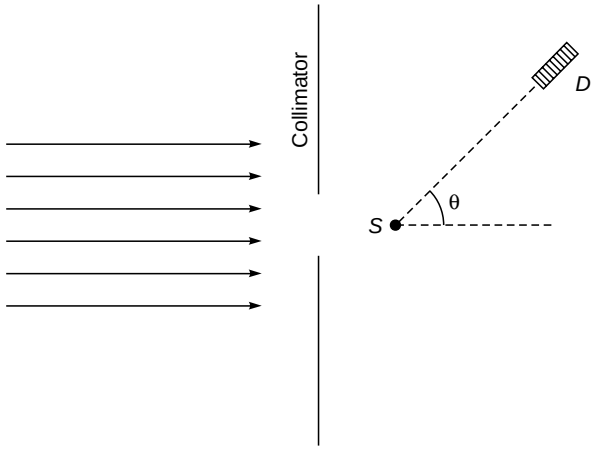


Fig. 25.5 Outline of the experimental arrangement for the measurement of the Compton shift. A collimated beam of monochromatic X-rays is scattered by the scatterer *S;* the wavelength of the scattered photon is measured by the detector *D*.

unambiguous example of a process in which a quantum of radiation carrying energy as well as momentum scatters off an electron (see Fig. 25.4). Now, if *u* represents the energy per unit volume associated with a plane electromagnetic wave, Maxwell's equations predict that the momentum per unit volume associated with the electromagnetic wave is *u/c*, where *c* represents the speed of light in free space (see Sec. 23.6). Since each photon carries an energy equal to $h\nu$, it should have a momentum given by

$$p = \frac{h\nu}{c} = \frac{h}{\lambda} \qquad \textbf{(3)}$$

In 1923, Compton investigated the scattering of X-rays by a block of paraffin and found that the wavelength of the radiation scattered at an angle of 90° is greater than the wavelength of the incident radiation. In other words, the frequency ν′ of the scattered wave is smaller than the frequency of the incident wave. Compton was able to explain the result[3] quantitatively as that of an elastic collision between a photon of energy $E = h\nu$ and the momentum given by Eq. (3). Compton was awarded the 1927 Nobel Prize in Physics for his discovery of the effect named after him.

The light quantum imparts some of its energy to the electron and emerges with less energy. Thus the scattered radiation has a lower frequency. The kinematics of this collision process can be worked out on elementary application of the laws of conservation of energy and momentum (see Sec. 25.3.1). These calculations give the following expression for the shift in the wavelength

$$\Delta\lambda = \frac{2h}{m_0 c}\sin^2\frac{\theta}{2} \qquad \textbf{(4)}$$

where θ is the angle of scattering of the light quantum (see Fig. 25.4) and m_0 represents the rest mass of the electron. If we substitute the values of *h*, m_0, and *c*, we obtain

$$\Delta\lambda = \lambda' - \lambda = 0.0485\sin^2\frac{\theta}{2} \qquad \textbf{(5)}$$

where λ is measured in Angstroms. Equation (5) shows that the maximum change in the wavelength is about 0.05 Å, and as such for a measurable shift one must use radiation of smaller wavelength. In Fig. 25.5 we have given the schematic of the experimental arrangement for the measurement of the Compton shift. A monochromatic beam of X-rays (or γ-rays) is allowed to fall on a sample scatterer, and the scattered photons were detected by means of a crystal spectrometer. The crystal spectrometer allows one to find the intensity distribution (as a function of λ) for a given value of θ. In Fig. 25.6

[3] According to the classical explanation of Compton scattering, the electron undergoes oscillatory motion because of the electric field associated with the incident electromagnetic radiation. The accelerated electron emits electromagnetic waves, and because of Doppler shifts due to the motion of the electron, the emitted wavelength differs from the wavelength of the incident radiation. However, classical theory predicts that for a given angle of scattering, a continuous range in the value of the scattered wavelength should be formed, which is contrary to experimental findings. The details of this analysis are given in Sec. 2.9 of Ref. 6.

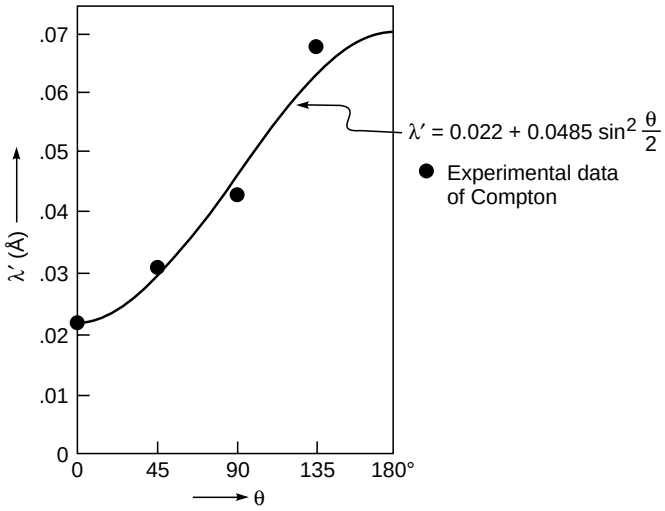


Fig. 25.6 The variation of wavelength of the scattered photon with the angle of scattering. The solid curve corresponds to Eq. (5) with $\lambda = 0.022$ Å. The dots represent the experimental points obtained by Compton. The figure has been adapted from the original paper of Compton (Ref. 9).

we have shown the wavelength of the scattered photon at different angles with respect to the primary beam as obtained by Compton in his original experiment (Ref. 7) in 1923. The solid curve corresponds to Eq. (5) with $\lambda = 0.022$ Å. Notice that the corresponding photon energy is

$$\sim \frac{6.6\times10^{-27}\ \text{erg s}\times 3\times10^{10}\ \text{cm s}^{-1}}{2.2\times10^{-10}\ \text{cm}}\ \text{erg} \simeq 0.5\ \text{MeV}$$

which corresponds to a γ-ray. The good agreement between theory and experiment proves that radiation behaves as if it consists of corpuscles of energy $h\nu$ having a momentum $h\nu/c$.

The experimental arrangement and findings of Compton are shown in Figs. 25.7 and 25.8; the experiment corresponds to the molybdenum K_α line ($\lambda = 0.711$ Å). The sample used was graphite. Notice that at each value of θ, there are two peaks; the first peak appears at almost the same wavelength as the primary beam. This peak is due to the fact that the photon may be scattered by the whole atom; consequently, the quantity m_0 appearing in Eq. (4) is not the electron mass but the mass of the carbon atom (which is about 22,000 times the mass of the electron). Thus the wavelength shift is negligible. The second peak corresponds to the Compton shift. In each figure, the two vertical lines correspond to the unmodified wavelength and the modified wavelength as given by Eq. (5), and one can see good agreement between the predicted and observed values.

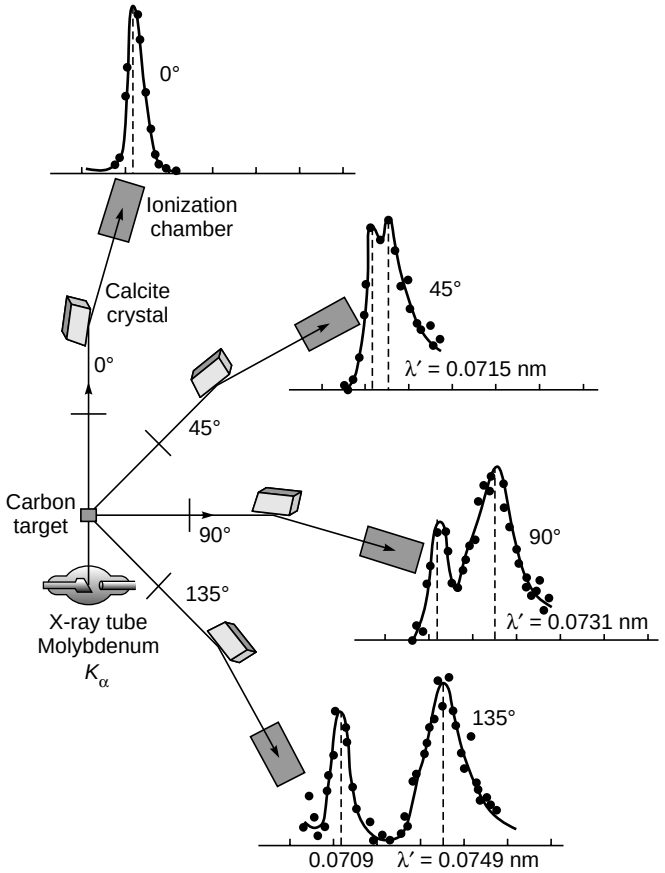


Fig. 25.7 Compton's original experiment made use of molybdenum K_α X-rays, which have a wavelength of 0.0709 nm. These were scattered from a block of carbon and observed at different angles with a Bragg spectrometer. The experimental data are from the original paper of Compton (see Fig. 25.8). Adapted from Ref. 10 and a diagram created by Professor Rod Nave at Georgia State University [Ref. http://hyperphysics.phy-astr.gsu.edu/hbase/hframe.html].

Further evidence of the validity of the above theory was provided by the experiments carried out by Compton and Simon who studied the scattering of X-rays through supersaturated water vapor. In the scattering process, the recoil electrons formed tracks of condensed droplets; however, the light quantum did not leave any track. Now, if the light quantum undergoes another Compton scattering, then from the track of the second recoil electron one can determine the path of the light quantum by simply joining the line of the starting points of the two recoil electrons. Although there was considerable uncertainty in the analysis of the experimental data (because of the presence of many tracks), Compton and Simon could establish agreement between theoretical results and experimental data.

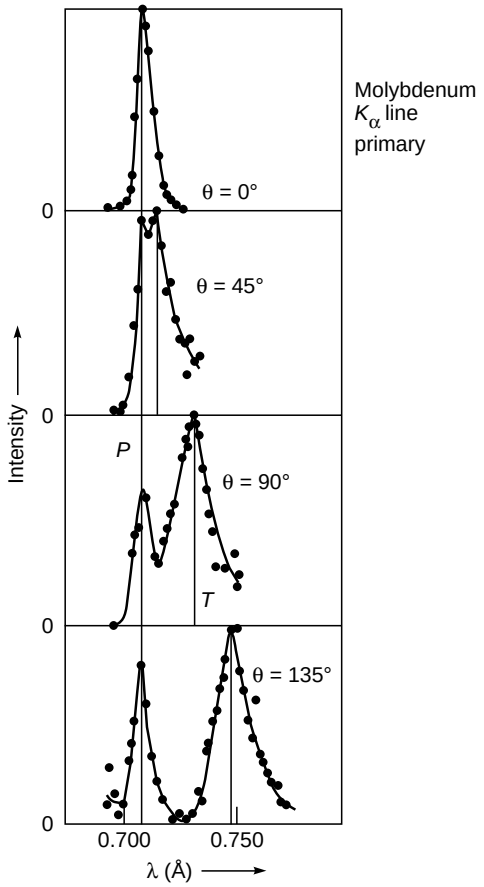


Fig. 25.8 The intensity variation as a function of the wavelength of the scattered photon. The vertical line (marked *P*) corresponds to the unmodified wavelength λ = 0.711 Å. The second vertical line (marked *T*) corresponds to the wavelength as predicted by Eq. (5). The figure has been adapted from the original paper of Compton (Ref. 10).

25.3.1 Kinematics of Compton Scattering

We next consider the scattering of a photon by an electron as shown in Fig. 25.4. The scattered photon is assumed to have a frequency ν′. Conservation of energy leads to

$$h\nu = h\nu' + E_k \tag{6}$$

where E_k represents the kinetic energy imparted to the electron. Conserving the x and y components of the momentum, we have

$$\frac{h\nu}{c} = \frac{h\nu'}{c}\cos\theta + p\cos\phi \tag{7}$$

and

$$0 = \frac{h\nu'}{c}\sin\theta - p\sin\phi \tag{8}$$

where p represents the momentum of the electron after collision and θ and φ represent the angles made by the scattered photon and electron with the original direction of the photon (see Fig. 25.4). It will be shown that for a measurable Compton effect, the frequency ν should be in the X-ray or in the γ-ray region (for X-rays λ ≤ 1 Å and $h\nu \geq 10^4$ eV). For such high-energy photons, the velocity imparted to the electron is comparable to the speed of light, and one must use proper relativistic expressions for E_k and p. Now, according to the theory of relativity, the kinetic energy E_k of the scattered electron is given by

$$E_k = E - m_0c^2 = mc^2 - m_0c^2 = \frac{m_0c^2}{\sqrt{1-\beta^2}} - m_0c^2 \tag{9}$$

where $\beta = v/c$, m_0 represents the rest mass of the electron, v is the speed of the electron, and c is the speed of light in free space; the quantities E and m_0c^2 are known as the total energy and the rest mass energy of the electron. Further, the relativistic momentum of the electron is given by

$$p = mv = \frac{m_0 v}{\sqrt{1-\beta^2}} \tag{10}$$

Now,

$$p^2c^2 + m_0^2c^4 = \frac{m_0^2v^2c^2}{1-v^2/c^2} + m_0^2c^4$$

$$= \frac{m_0^2c^4}{1-v^2/c^2} = m^2c^4$$

or

$$p^2c^2 + m_0^2c^4 = E^2 = (E_k + m_0c^2)^2$$

$$= E_k^2 + m_0^2c^4 + 2E_km_0c^2$$

Thus,

$$E_k^2 + 2E_km_0c^2 = p^2c^2$$

Substituting for E_k from Eq. (6), we get

$$h^2(\nu - \nu')^2 + 2h(\nu - \nu')\,m_0c^2 = p^2c^2 \tag{11}$$

Further, Eqs. (7) and (8) can be rewritten in the form

$$p\cos\phi = \frac{h\nu}{c} - \frac{h\nu'}{c}\cos\theta \tag{12}$$

and

$$p\sin\phi = \frac{h\nu'}{c}\sin\theta \tag{13}$$

To eliminate φ, we square and add to obtain

$$p^2 = \left(\frac{h\nu}{c}\right)^2 + \left(\frac{h\nu'}{c}\right)^2 - \frac{2h^2\nu\nu'}{c^2}\cos\theta \tag{14}$$

Substituting in Eq. (11), we obtain

$$h^2(\nu^2 - 2\nu\nu' + \nu'^2) + 2h(\nu - \nu')m_0c^2$$

$$= h^2\nu^2 + h^2\nu'^2 - 2h^2\nu\nu'\cos\theta$$

or

$$\frac{2h(\nu - \nu')m_0c^2}{2\nu\nu'} = h^2(1 - \cos\theta)$$

or $$\Delta\lambda = \lambda' - \lambda = \frac{h}{m_0c}(1 - \cos\theta)$$

or $$\Delta\lambda = \frac{2h}{m_0c}\sin^2\frac{\theta}{2} \quad \textbf{(15)}$$

which gives us the Compton shift.[4]

25.4 THE PHOTON MASS

Because the photon has energy (= $h\nu$ /c) we may assume it to have an inertial mass given by

$$m = \frac{h\nu}{c^2} \quad \textbf{(16)}$$

Thus when a light beam passes near a heavy star, its trajectory ought to get deflected. Indeed, the light coming from a distant star does get slightly deflected when passing near the Sun, which has been experimentally observed.

Also, we may expect that when a photon leaves a star, its energy should decrease because of the gravitation field. This indeed happens and manifests itself in a decrease in frequency which is usually referred as the *gravitational red shift*. One can approximately calculate the red shift by noting that the potential energy on the surface of the star is

$$V \approx -\frac{GMm}{R} = -\frac{GM}{R}\cdot\frac{h\nu}{c^2} \quad \textbf{(17)}$$

where M is the mass of the star, R its radius, and G the gravitational constant. Thus when the light beam reaches Earth, its frequency is

$$h\nu' = h\nu - \frac{GM}{R}\frac{h\nu}{c^2}$$

or

$$\frac{\Delta\nu}{\nu} = \frac{\nu - \nu'}{\nu} = \frac{GM}{Rc^2} \quad \textbf{(18)}$$

(we have neglected the effect of the Earth's gravitational field). From the above equation, we see that if the mass of the star is so large that the RHS exceeds unity, then the light beam will not be able to escape from the star—this is a black hole. In discussing black holes, we must use the general theory of relativity, which obtains the following value for the limiting radius of the star:

$$R_s = \frac{2GM}{c^2} \quad \textbf{(19)}$$

this is known as the Schwarschild radius. If the mass of the star is contained inside a sphere of radius

$$R < R_s$$

then a light beam will never leave the star and the star will be known as a black hole. Thus if

$$M \approx 10M_\odot \approx 2\times10^{34}\ \text{g}$$

where $M_\odot$ ($\approx 2\times10^{33}$ g) represents the mass of the Sun, then

$$R_s \approx \frac{2\times6.67\times10^{-8}\times2\times10^{34}}{(3\times10^{10})^2}\ \text{cm} \approx 30\ \text{km}$$

Indeed black holes with radius ~10 km have been detected!

25.5 ANGULAR MOMENTUM OF A PHOTON

We consider an electromagnetic wave (propagating along the x direction) to first pass through a Polaroid whose pass axis is along the y' axis (see Fig. 25.9). The electric field along the y and z directions will be given by

$$E_y = E_y'\cos\theta \qquad \text{and} \qquad E_z = E_y'\sin\theta$$

Thus when a y' polarized beam is passed through a Polaroid whose pass axis is along the y direction (see Fig. 25.9), the

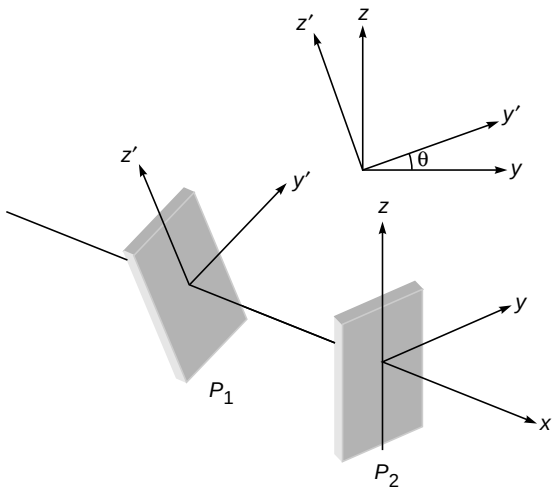


Fig. 25.9 Polaroid P_1 polarizes in the y' direction, and Polaroid P_2 polarizes in the y direction; the propagation is in the x direction.

[4] In the derivation of the Compton shift, we assume that the electron is free although we know that the electrons are bound to the atoms. The assumption of a free electron is justified because the binding energy (≈ few electron-volts) is usually very much smaller in comparison to the photon energy ($\gtrsim$ 1000 eV).

component of the electric field that passes through is cos θ and the intensity gets reduced by a factor of $\cos^2\theta$—which is nothing but the law of Malus (see Sec. 22.3). Similarly for a beam polarized along the z' direction

$$E_y = -E_z' \sin\theta \qquad \text{and} \qquad E_z = E_z' \cos\theta$$

In general, we may write

$$E_y = E_y' \cos\theta - E_z' \sin\theta$$

and

$$E_z = E_y' \sin\theta + E_z' \cos\theta$$

The above equations can be written in the matrix form

$$\begin{pmatrix} E_y \\ E_z \end{pmatrix} = \mathbf{S}(\theta) \begin{pmatrix} E_y' \\ E_z' \end{pmatrix} \tag{20}$$

where

$$\mathbf{S}(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{21}$$

is the (rotation) matrix which transforms from the y'-z' basis to the y-z basis. From the above equations we also obtain

$$E_y' = E_y \cos\theta + E_z \sin\theta$$
$$E_z' = -E_y \sin\theta + E_z \cos\theta$$

Thus

$$\begin{pmatrix} E_y' \\ E_z' \end{pmatrix} = \mathbf{S}^\dagger(\theta) \begin{pmatrix} E_y \\ E_z' \end{pmatrix} \tag{22}$$

where

$$\mathbf{S}^\dagger(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{23}$$

is the matrix which transforms from the y-z basis to the y'-z' basis. Obviously

$$\mathbf{S}^\dagger(\theta)\,\mathbf{S}(\theta) = \mathbf{S}(\theta)\,\mathbf{S}^\dagger(\theta) = \mathbf{1} \tag{24}$$

In Sec. 2.9, we discussed the polarization of a photon; a y-polarized photon can be represented by the "unit" vector (see Sec. 22.14):

$$|y> = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Similarly, a z-polarized photon is represented by the unit vector

$$|z> = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

In the rotated coordinates (see Fig. 25.9)

$$\begin{pmatrix} |y> \\ |z> \end{pmatrix} = \mathbf{S}(\theta) \begin{pmatrix} |y'> \\ |z'> \end{pmatrix}$$

and

$$\begin{pmatrix} |y'> \\ |z'> \end{pmatrix} = \mathbf{S}^\dagger(\theta) \begin{pmatrix} |y> \\ |z> \end{pmatrix}$$

Further, a right circularly polarized photon can be represented by the unit vector

$$|R> = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix} = \frac{1}{\sqrt{2}} [|y> + i|z>] \tag{25}$$

Similarly, a left circularly polarized photon can be represented by the unit vector

$$|L> = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix} = \frac{1}{\sqrt{2}} [|y> - i|z>] \tag{26}$$

Under rotation, the right circularly polarized states $|R>$ will transform to

$$\begin{aligned} |R'> &= \frac{1}{\sqrt{2}} [|y'> + i|z'>] \\ &= \frac{1}{\sqrt{2}} \{[\cos\theta\,|y> + \sin\theta\,|z>] \\ &\quad + i[-\sin\theta\,|y> + \cos\theta\,|z>]\} \\ &= \frac{1}{\sqrt{2}} e^{-i\theta} [|y> + i|z>] \end{aligned}$$

or

$$|R'> = e^{-i\theta}\,|R> \tag{27}$$

Thus in the rotated coordinate system, the right circularly polarized photon remains right circularly polarized except for a change in the phase. Now, in quantum mechanics, if $\mathcal{R}_x(\theta)$ represents the rotational operator corresponding to a rotation about the x axis through an angle θ, then (see, e.g., Refs. 11 and 12)

$$\mathcal{R}_x(\theta) = \exp\left(-\frac{i}{\hbar}\theta J_x\right) \tag{28}$$

where

$$\hbar = \frac{h}{2\pi}$$

with h being the Planck's constant and J_x representing the x component of the angular momentum operator. Thus

$$\mathcal{R}_x(\theta)\,|R> = |R'> = e^{-i\theta}\,|R> \tag{29}$$

or

$$\exp\left(-\frac{i}{\hbar}\theta J_x\right)|R> = e^{-i\theta}\,|R> \tag{30}$$

Now, the exponential of an operator O is defined by

$$e^{O} \equiv 1+\frac{O}{1!}+\frac{O^2}{2!}+\frac{O^3}{3!}+\cdots$$
$$=1+\frac{O}{1!}+\frac{O.O}{2!}+\frac{O.O.O}{3!}+\cdots$$

We expand the exponential on both sides of Eq. (30), and if we use the fact that Eq. (30) has to be valid for all values of θ, we must have[5]

$$J_x|R> = +\hbar\,|R> \qquad \textbf{(31)}$$

If we carry out a similar analysis for the left circularly polarized light, we obtain

$$J_x|L> = -\hbar\,|L> \qquad \textbf{(32)}$$

Equations (31) and (32) are known as *eigenvalue equations*; thus, the right and left circularly polarized states are said to be the *eigenstates* of J_x; the corresponding eigenvalues are $+\hbar$ and $-\hbar$, respectively. According to quantum mechanics, if we measure J_x of a right circularly polarized light photon, we will always obtain the value $+\hbar$; similarly, if we measure J_x of a left circularly polarized photon, we will obtain the value $-\hbar$. For an arbitrary state of polarization, if we measure J_x, we will obtain one of the eigenvalues; i.e., we obtain either the value $+\hbar$ or the value $-\hbar$. To obtain the probabilities of finding $+\hbar$ and $-\hbar$, the state should be expressed as a superposition of the eigenstates, which are the right circularly polarized state and the left circularly polarized state. For example, a *y*-polarized state can be represented as

$$|y> = \begin{pmatrix}1\\0\end{pmatrix} = \frac{1}{\sqrt{2}}[|R> + |L>]$$

Thus, if we make a measurement of J_x on a *y*-polarized (or on a *z*-polarized) state, then there will be 0.5 probability of obtaining $+\hbar$ and 0.5 probability of obtaining $-\hbar$; one can never predict the precise outcome of an experiment.[6] As discussed in Secs. 2.6 to 2.9, this is typical of quantum mechanics; physics has ceased to be deterministic—one can only predict the probabilities of a specific outcome of an experiment. As another example, for the left elliptically polarized state discussed in Sec. 22.14,

$$|\text{LEP}> = \begin{pmatrix}\frac{1}{2}\\ -i\frac{\sqrt{3}}{2}\end{pmatrix} = a|R> + b|L>$$
$$= \frac{a}{\sqrt{2}}\begin{pmatrix}1\\ i\end{pmatrix} + \frac{b}{\sqrt{2}}\begin{pmatrix}1\\ -i\end{pmatrix}$$

Simple manipulations will give

$$a = \frac{1}{2}\left(\frac{1}{\sqrt{2}} - \sqrt{\frac{3}{2}}\right) \approx -0.2588$$

and

$$b = \frac{1}{2}\left(\frac{1}{\sqrt{2}} + \sqrt{\frac{3}{2}}\right) \approx +0.9659$$

Thus, if we make a measurement of J_x on such an elliptically polarized state, we obtain one of the eigenvalues: the probability of obtaining $+\hbar$ will be about 0.0670, and the probability of obtaining $-\hbar$ will be about 0.933.

Summary

- In 1887, while receiving the electromagnetic waves in a coil with a spark gap, Hertz found that the maximum spark length was reduced when the apparatus was put in a black box; this is due to what is now known as the photoelectric effect, and the box absorbed the ultraviolet radiation which helped the electrons to jump across the gap. Hertz reported the observations but did not pursue further and also did not make any attempt to explain them. In 1897 J. J. Thomson discovered electrons, and in 1899 he showed that electrons are emitted when light falls on a metal surface; these are now known as photoelectrons, and the phenomenon is known as the *photoelectric effect*.
- There are certain peculiarities associated with the photoelectric effect which cannot be explained on the basis of wave theory. For example, a faint violet light ejects electrons of greater kinetic energy than an intense yellow light although the latter produces a larger number of electrons. In 1905, Einstein provided a simple explanation of the peculiarities by assuming that light consisted of quanta of energy $h\nu$ (where ν is the frequency) and that the emission of a photoelectron was the result of the interaction of a single quantum (i.e. of the photon) with an electron. In his 1905 paper, Einstein wrote *Monocromatic radiation behaves as if it consists of mutually independent energy quanta of magnitude [hν]*.
- In 1923, Compton reported his studies on the scattering of X-rays by solid materials (mainly graphite) and showed that the shift of the wavelength of the scattered photon could be explained by assuming the photon to have momentum equal to h/λ. The Compton effect provided an unambiguous example of a process in which a quantum of radiation carrying energy as well as momentum scatters off an electron. The kinematics of the

[5] Thus, for example, $J_x^3|R> = J_xJ_xJ_x|R> = \hbar^3|R>$.

[6] For an arbitrary state of polarization, we must express it as a linear superposition of $|R>$ and $|L>$.

scattering process gives the following expression for the shift in the wavelength

$$\Delta\lambda = \frac{2h}{m_0 c}\sin^2\frac{\theta}{2} \approx 0.0485 \sin^2\frac{\theta}{2}$$

where θ is the angle of scattering of the light quantum, m_0 represents the rest mass of the electron, and $\Delta\lambda$ is measured in angstroms. Compton found the above formula to be in agreement with his experimental measurements of $\Delta\lambda$.

Problems

25.1 (a) Calculate the number of photons emitted per second by a 5 mW laser, assuming that it emits light of wavelength 6328 Å.

[**Ans:** 1.6×10^{16}]

(b) The beam is allowed to fall normally on a plane mirror. Calculate the force acting on the mirror.

[**Ans:** 3.3×10^{-11} N]

25.2 Assume a 40 W sodium lamp ($\lambda \approx 5893$ Å) emitting light in all directions. Calculate the rate at which the photons cross a unit area placed normally to the beam at a distance of 10 m from the source.

[**Ans:** $\approx 10^{17}$ photons per m^2 per s]

25.3 In the photoelectric effect, a photon is completely absorbed by the electron. Show that the laws of conservation of energy and momentum cannot be satisfied simultaneously if a free electron is assumed to absorb the photon. (Thus the electron has to be bound to an atom, and the atom undergoes a recoil when the electron is ejected. However, since the mass of the atom is much larger than that of the electron, the atom picks up only a small fraction of the energy. This is somewhat similar to the case of a tennis ball hitting a heavy object—the momentum of the ball is reversed with its energy remaining almost the same.)

25.4 If photoelectrons are emitted from a metal surface by using blue light, can you say for sure that photoelectric emission will take place with yellow light and with violet light?

REFERENCES AND SUGGESTED READINGS

1. F. G. Smith and T. A. King, *Optics & Photonics*, Wiley, Chicester, 2000.
2. W. H. Cropper, *The Quantum Physicists and an Introduction to Their Physics*, Oxford University Press, New York, 1970.
3. A. Einstein, "On a Heuristic Point of View concerning the Production and Transformation of Light," *Annalen der Physik*, Vol. 17, p. 132, 1905.
4. M. Jammer, *The Conceptual Development of Quantum Mechanics*, McGraw-Hill, New York, 1965.
5. R. A. Millikan, "The Electron and the Light-Quanta from the Experimental Point of View," Nobel Lecture delivered in May 1924. Reprinted in *Nobel Lectures in Physics*, Elsevier Publishing Co., Amsterdam, 1965.
6. Bohm, *Quantum Theory*, Prentice-Hall, Englewood Cliffs, N. J., 1951.
7. R. S. Shankland (Ed.), *Scientific Papers of A. H. Compton: X-ray and Other Studies*, University of Chicago Press, Chicago, 1975. Most of the papers of A. H. Compton are reprinted here.
8. M. Born, *Atomic Physics*, Blackie & Son, London, 1962.
9. A. H. Compton, "A Quantum Theory on the Scattering of X-Rays by Light Elements," *Physical Review*, Vol. 21, p. 483. 1923, Reprinted in Ref. 7.
10. A. H. Compton, "The Spectrum of Scattered X-rays", *Physical Review*, Vol. 22, p. 409, 1923. Reprinted in Ref. 7.
11. J. Townsend, *A Modern Approach to Quantum Mechanics*, McGraw-Hill Inc., New York, 1992.
12. G. Baym, *Lectures on Quantum Mechanics*, W. A. Benjamin, Inc., New York, 1969.

PART 7
Lasers and Fiber Optics

This part consists of four chapters. Chapter 26 is on lasers whose discovery in 1960 led to numerous applications in many diverse areas; the chapter discusses the basic physics of lasers along with their special characteristics. Chapters 27 through 29 are on fiber optics and waveguide theory, an area which during the last 35 years has revolutionized communications.

Chapter Twenty-Six

LASERS: AN INTRODUCTION

In *The War of Worlds*, written before the turn of the century, H. G. Wells told a fanciful story of how Martians invaded and almost conquered the Earth. Their weapon was a mysterious "sword of heat," from which flickered "a ghost of a beam of light," it felled men in their tracks, made lead run like water and flashed anything combustible into masses of flame. Today Wells' sword of heat comes close to reality in the laser...

—Thomas Meloy

Important Milestones

1917	*The theory of stimulated emission was put forward by Albert Einstein.*
1954	*The phenomenon of stimulated emission was first used by Charles Townes in 1954 in the construction of a microwave amplifier device called the maser, which is an acronym for microwave amplification by stimulated emission of radiation. At about the same time, a similar device was also proposed by Prochorov and Basov in the U.S.S.R.*
1958	*The maser principle was later extended to the optical frequencies by Schawlow and Townes in 1958, which led to the realization of the device now known as the laser. Townes, Basov, and Prochorov were awarded the 1964 Nobel Prize in Physics for their "fundamental work in the field of Quantum Electronics, which has led to the construction of oscillators and amplifiers based on the laser-maser principle."*[1]
1959	*In a conference paper, Gordon Gould introduced the term LASER as an acronym for Light Amplification by Stimulated Emission of Radiation.*
1960	*The first successful operation of a laser device ($\lambda \sim 0.6943\ \mu m$) was demonstrated by Theodore Maiman in 1960 using a ruby crystal (see Sec. 26.3).*
1961	*Within a few months of the operation of the ruby laser, Ali Javan and his associates constructed the first gas laser, namely, the helium-neon laser (see Sec. 26.2).*
1961	*The first fiber laser (barium crown glass doped with Nd^{3+} ions) was fabricated by Elias Snitzer.*
1962	*Semiconductor lasers (which are now extensively used in fiber-optic communication systems) were discovered by four independent groups.*
1963	*C. K. N. Patel discovered the CO_2 laser ($\lambda \sim 10.6\ \mu m$).*
1964	*W. Bridges discovered the Ar-ion laser ($\lambda \sim 0.515\ \mu m$), and J. E. Geusic and coworkers discovered the Nd:YAG laser ($\lambda \sim 0.515\ \mu m$).*
	Since then, laser action has been obtained in a large variety of materials including liquids, ionized gases, dyes, and semiconductors.

26.1 INTRODUCTION

LASER is an acronym for *l*ight *a*mplification by *s*timulated *e*mission of *r*adiation. The light emitted from a laser often possesses some very special characteristics. Some of these are

1. Directionality. The divergence of the laser beam is usually limited by diffraction (see the figure in the insert at the back of the book), and the actual divergence can be less than 10^{-5} rad; this leads to the application of the laser in surveying, remote sensing, lidar etc.

[1] The Nobel lectures of Townes, Basov, and Prochorov (Refs. 1–3) give a nice perspective of the field. These are reprinted in Ref. 4.

2. High power. Continuous wave lasers having power levels of ~10^5 W and pulsed lasers having a total energy of ~50,000 J can have applications in welding, cutting, laser fusion, star wars, etc.
3. Tight focusing. Because of highly directional properties of the laser beams, they can be focused to areas of approximately few micrometers squared—this leads to very high intensities and therefore leads to applications in surgery, material processing, compact discs, etc. Laser pulses having very small cross-sectional area (and high energy) can be guided through special fibers leading to very interesting nonlinear effects (see Fig. 10.10).
4. Spectral purity. Laser beams can have an extremely small spectral width $\Delta\lambda \sim 10^{-6}$ Å. Because of high spectral purity, lasers find applications in holography, optical communications, spectroscopy, etc.

Because of such unique properties of the laser beam, it finds important applications in many diverse areas, and indeed one can say that after the discovery of the laser, optics has become an extremely important field of study. For example, in Sec. 18.4 we showed that a 2 mW diffraction-limited laser beam incident on the eye can produce an intensity of about 10^6 W m^{-2} at the retina—this would certainly damage the retina. Thus, whereas it is quite safe to look at a 500 W bulb, it is very dangerous to look directly into a 5 mW laser beam. Indeed, because a laser beam can be focused to very narrow areas, it has found applications in areas such as eye surgery and laser cutting.

The basic principle involved in the lasing action is the phenomenon of stimulated emission, which was predicted by Einstein in 1917 (Ref. 5); the original paper of Einstein is reprinted in Ref. 6. In Sec. 26.6.1 we will give the original argument of Einstein to obtain the relationship between the Einstein coefficients. This will be followed by brief discussions of the main components of a laser, and the underlying principle as to how the laser works. In Sec. 26.2 we will briefly discuss the working of a fiber laser, and in Sec. 26.3 we will discuss the working of the ruby laser, which was the first laser to be fabricated. In Sec. 26.4 we will discuss the working of the helium-neon laser. In Sec. 26.5 we will give a slightly more detailed account of resonators, and in Sec. 26.6 we will discuss Einstein coefficients and optical amplification. In Sec. 26.7 we will discuss the line shape function, and finally in Sec. 26.8 we will discuss the monochromaticity of the laser beam.

26.1.1 Spontaneous and Stimulated Emissions

Atoms are characterized by discrete energy states. According to Einstein, there are three different ways in which an atom can interact with electromagnetic radiation:

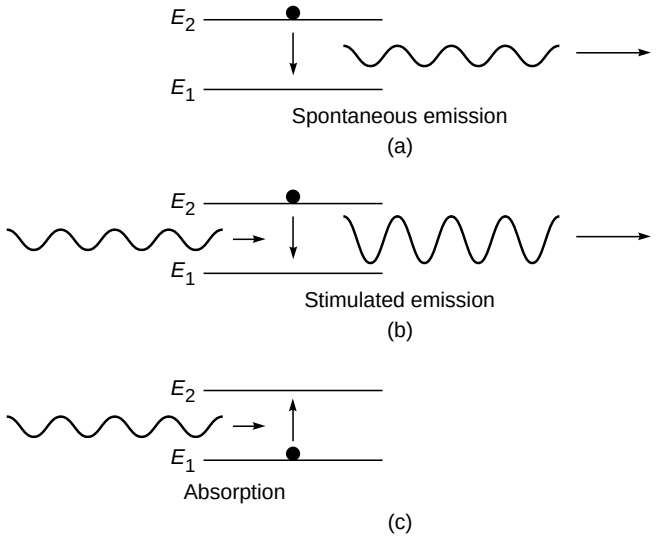


Fig. 26.1 (a) Spontaneous emission, (b) stimulated emission, and (c) stimulated absorption.

1. Spontaneous emission. Atoms in the energy state E_2 can make a (spontaneous) transition to the energy state E_1 with the emission of radiation of frequency

$$\omega = \frac{E_2 - E_1}{\hbar} \qquad \textbf{(1)}$$

where

$$\hbar = \frac{h}{2\pi} \approx 1.0546 \times 10^{-34} \text{ J s}$$

and h ($\approx 6.626 \times 10^{-34}$ J s) is known as Planck's constant. Since this process can occur even in the absence of any radiation, this is called spontaneous emission [see Fig. 26.1(a)]. The rate of spontaneous emission is proportional to the number of atoms in the excited state.
2. Stimulated emission. As put forward by Einstein, when an atom is in the excited state, it can also make a transition to a lower energy state through what is known as stimulated emission, in which an incident signal of appropriate frequency triggers an atom in an excited state to emit radiation—this results in the amplification of the incident beam [see Fig. 26.1(b)]. The rate of stimulated emission depends on both the intensity of the external field and the number of atoms in the excited state.
3. Stimulated absorption. Stimulated absorption (or simply absorption) is the process in which the electromagnetic radiation of an appropriate frequency (corresponding to the energy difference of the two atomic levels) can pump the atom to its excited state [see Fig. 26.1(c)]. The rate of stimulated absorption depends both on the intensity of the external field and on the number of atoms in the lower energy state.

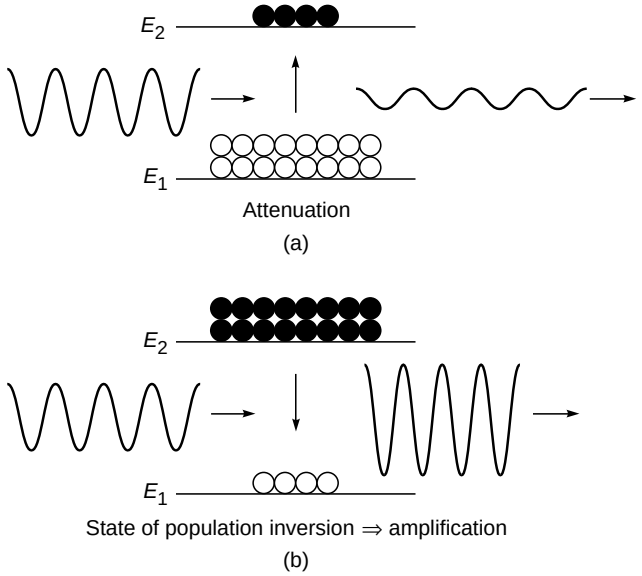


Fig. 26.2 (a) A larger number of atoms in the lower state result in the attenuation of the beam. (b) A larger number of atoms in the upper state (which is known as population inversion) result in the amplification of the beam.

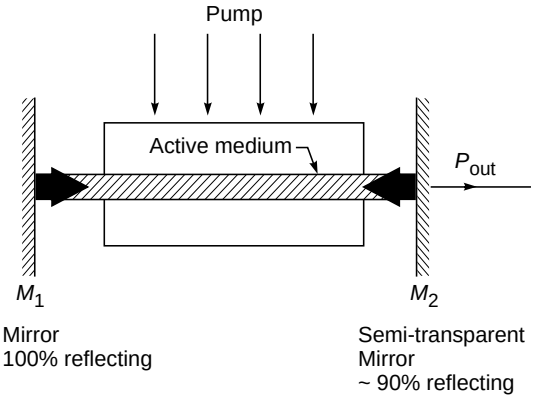


Fig. 26.3 The three basic components of a laser are (1) the active medium (which provides amplification), (2) the optical resonator (which provides frequency selection and optical feedback), and (3) the pump (which supplies power to the active medium to achieve population inversion).

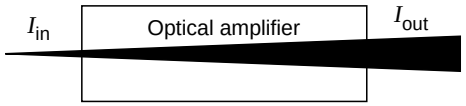


Fig. 26.4 The active medium essentially consists of a collection of atoms in a state of population inversion which can amplify the input light beam (or spontaneously emitted light) by stimulated emission. This is known as optical amplification.

When the atoms are in thermodynamic equilibrium, there are larger number of atoms in the lower state, implying that the number of absorptions exceeds the number of stimulated emissions; this results in the attenuation of the beam [see Fig. 26.2(a)]. On the other hand, if we are able to create a state of population inversion in which there are larger number of atoms in the upper state, then the number of stimulated emissions exceeds the number of absorptions, resulting in the (optical) amplification of the beam [see Fig. 26.2(b)]. The amplification process due to stimulated transitions is *phase-coherent*, i.e. [quoting Townes (Ref. 1)], "the energy delivered by the molecular system has the same field distribution and frequency as the stimulating radiation."

26.1.2 Main Components of the Laser

The three main components of any laser are (see Fig. 26.3)

1. Active medium. The active medium consists of a collection of atoms, molecules, or ions (in solid, liquid, or gaseous form) which is capable of amplifying light waves. Under normal circumstances, there are always a larger number of atoms in the lower energy state than in the excited energy state. An electromagnetic wave passing through such a collection of atoms is attenuated; this is discussed in detail in Sec. 26.6. To have optical amplification, the medium has to be kept in a state of *population inversion*, i.e., in a state in which the number of atoms in the upper energy level is greater than that in the lower energy level—this is achieved by means of the pump.
2. Pumping source. The pump enables us to obtain such a state of population inversion between a pair of energy levels of the atomic system. When we have a state of population inversion, the input light beam can get amplified by stimulated emission (see Fig. 26.4).
3. Optical resonator. A medium with population inversion is capable of amplification; however, for it to act as an oscillator, a part of the output energy must be fed back into the system.[2] Such feedback is brought about by placing the active medium in a resonator; the resonator could be just a pair of mirrors facing each other.

[2] Since some of the energy is coupled back to the system, it is said to act as an oscillator. Indeed, in the early stages of the development of the laser, there was a move to change its name to LOSER which is an acronym for *l*ight *o*scillation by *s*timulated *e*mission of *r*adiation. Since it would have been difficult to obtain a research grant for LOSERs, it was decided to retain the name LASER.

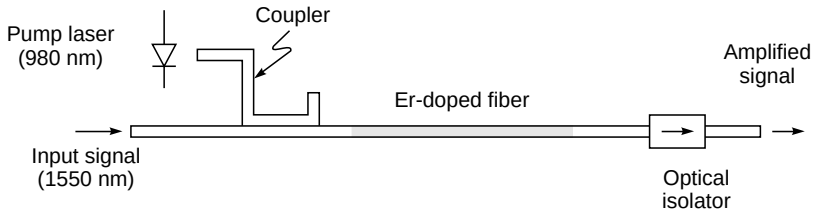


Fig. 26.5 The erbium-doped fiber amplifier (EDFA) in which the input optical pulses (at 1550 nm) are amplified by stimulated emission of radiation.

Although Einstein proposed the theory of stimulated emission in 1917, the concept of population inversion to amplify the light beam came much much later. According to Charles Townes,[3]

> The laser invention happened because I wanted very much to be able to make an oscillator at frequencies as high as the infrared in order to extend the field of microwave spectroscopy in which I was working. I had tried several ideas, but none worked very well. At the time I was also chairman of a committee for the navy that was examining ways to obtain very short-wave oscillators. In 1951, on the morning before the last meeting of this committee in Washington, I woke up early worrying over our lack of success. I got dressed and stepped outside to Franklin Park, where I sat on a bench admiring the azaleas and mulling over our problem.
>
> Why couldn't we think of something that would work at high frequencies? I went through the possibilities, including, of course, molecules, which oscillate at high frequencies. Although I had considered molecules before, I had dismissed them because of certain laws of thermodynamics.[4] But suddenly I recognized, "Hey, molecules don't have to obey such a law if they are not in equilibrium." And I immediately took a piece of paper out of my pocket and wrote equations to see if selection of excited molecules by molecular beam methods could produce enough molecules to provide a feedback oscillator. Wow! It looked possible.
>
> I went back to my hotel and told Art Schawlow about the idea, since he was staying at the same place. . . . Its extension to waves as short as light came a few years later, after much excitement over the maser and as a result of my continued collaboration with Schawlow, then at Bell Labs. An essential element in this discovery, I believe, was my experience in both engineering and physics: I knew both quantum mechanics and the workings and importance of feedback oscillators.

26.1.3 Understanding Optical Amplification: EDFA

Perhaps the easiest way to understand optical amplification is to discuss the working principle of an EDFA (erbium-doped fiber amplifier), which is shown in Fig. 26.5. The EDFA essentially consists of about 20 to 40 m of a silica optical fiber the core of which is doped with erbium oxide (Er_2O_3). We will give a detailed discussion on the optical fiber in Chaps. 27 and 29; it suffices here to say that light is guided through the optical fiber because of total internal reflection (see Fig. 27.8). The radius of the core of the optical fiber is typically about 4 to 5 μm. The erbium concentration is about 10^{25} ions m^{-3}. Figure 26.6 shows the first three energy levels of Er^{3+} ion in silica host glass. Actually, each level shown in the diagram consists of a large number of very closely spaced levels—but to keep the analysis simple, we have shown them as single levels. The energy difference between E_1 (the ground state) and E_3 corresponds to a wavelength of about 980 nm, and the energy difference between E_1 and E_2 corresponds to a wavelength of about 1530 nm; thus $E_3 - E_1 \approx 1.3$ eV and $E_2 - E_1 \approx 0.81$ eV.

Now, when a laser beam corresponding to the wavelength 980 nm is passed through the erbium-doped fiber, then the erbium atoms in the ground state E_1 absorb this radiation

[3] "LASERS and Fiber Optics Essay," by Charles H. Townes, http://www.greatachievements.org/?id=3717.

[4] In his Nobel lecture (reprinted in Ref. 4) Townes writes, "Why not use the atomic and molecular oscillators already built for us by nature? This had been one recurring theme which was repeatedly rejected. Thermodynamic arguments tell us that the interaction between electromagnetic waves and matter at any temperature cannot produce amplification." However, Townes realized that if population inversion is somehow achieved, then the radiation can be amplified. Quoting Townes again, "This condition is of course one of nonequilibrium for the group of molecules, which hence successfully obviates the limits set by blackbody radiation."

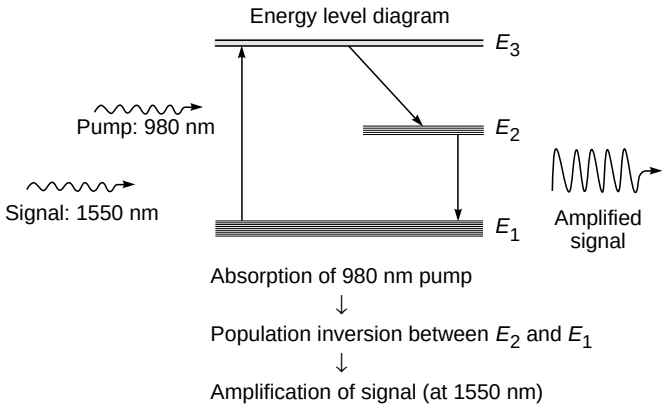


Fig. 26.6 The energy level diagram of the erbium atom in host silica.

and get excited to energy state E_3. This laser beam is usually referred to as a *pump* because it pumps the atoms to the higher energy state E_3. The atom in the energy state E_3 makes an almost immediate nonradiative transition to state E_2; in a nonradiative transition, a photon is not emitted—the energy released could, for example, add to the vibrational energy of the host medium, resulting in its heating. State E_2 is a metastable state characterized by a long lifetime (~ few milliseconds). The erbium atom in state E_2 can undergo a spontaneous transition to state E_1. However, because of the large lifetime of state E_2 (in comparison to that of E_3), the population of the erbium atoms in state E_2 grows with time, and if the pump power is high, the rate at which the erbium atom goes over to state E_2 can be so high that we may have a state of population inversion between E_1 and E_2; i.e., the number of erbium atoms in state E_2 is greater than that in E_1. When this happens, a signal beam at 1550 nm can get amplified by stimulated emission of radiation—this is the underlying principle of optical amplification which is nothing but *light amplification through stimulated emission of radiation* (see Fig. 26.4). Conversely, if the population of level E_2 is less than that of level E_1, the number of stimulated absorptions will exceed stimulated emission, resulting in the attenuation of the signal beam at 1550 nm. The variation of the pump and signal powers with distance along the doped fiber is shown schematically in Fig. 26.7. We notice that because of absorption by erbium atoms, the pump power gets attenuated as it propagates through the erbium-doped fiber. Because of this absorption, the erbium atoms are in a state of population inversion, and the signal at 1550 nm gets amplified. However, as we propagate through the erbium-doped fiber, the pump power decreases, and the erbium atoms are no more in a state of population inversion and the signal starts attenuating because of absorption by erbium atoms. Thus, for a given pump power, there is always an optimum length of the erbium-doped

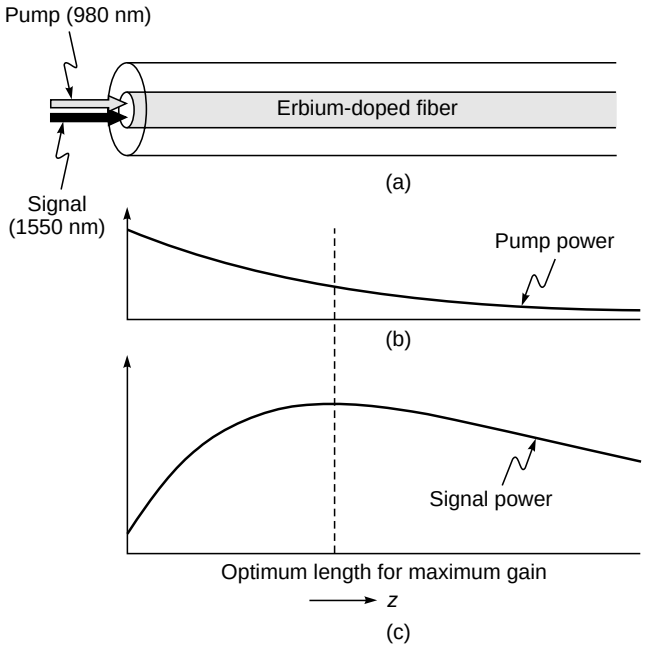


Fig. 26.7 (a) The pump (corresponding to 980 nm wavelength) and the signal (corresponding to 1550 nm wavelength) propagate in the core of an erbium-doped fiber. (b) and (c) represent the schematic variation of the pump and signal power as the two beams propagate through the erbium-doped fiber.

fiber for which maximum amplification occurs. For a typical erbium-doped fiber, we may have

$$\mathrm{Er}^{3+}\ \text{concentration} \approx 7 \times 10^{24}\ \text{ions m}^{-3} \qquad \text{pump power} \approx 5\ \text{mW}$$

and, the optimum length of the erbium doped fiber ≈ 7m.

A typical gain spectrum of an EDFA (using a 50 mW pump at 980 nm) is shown in Fig. 26.8(a). The gain is usually measured in dB which is defined as

$$\text{Gain (dB)} = 10 \log \frac{P_{\text{output}}}{P_{\text{input}}}$$

The gain (corresponding to the optimum length) is usually between 20 and 30 dB; a 20 dB gain implies a power amplification of 100; and a 30 dB gain implies a power amplification of 1000. If the pump power is higher, the optimum length and also the gain will be higher. The gain spectrum can be made flat over a certain wavelength region by a variety of techniques (e.g., by putting an appropriate filter after the EDFA). Figure 26.8(b) shows an almost flat gain (of about 28 dB) of an EDFA for wavelengths lying between 1530 and 1560 nm; a 28 dB gain corresponds to a power amplification of about 631. The wavelength region 1530 nm < λ < 1560 nm is extremely important for optical communications (see Chaps. 27 and 29). For more details on erbium-doped fiber amplifiers, see Refs. 7

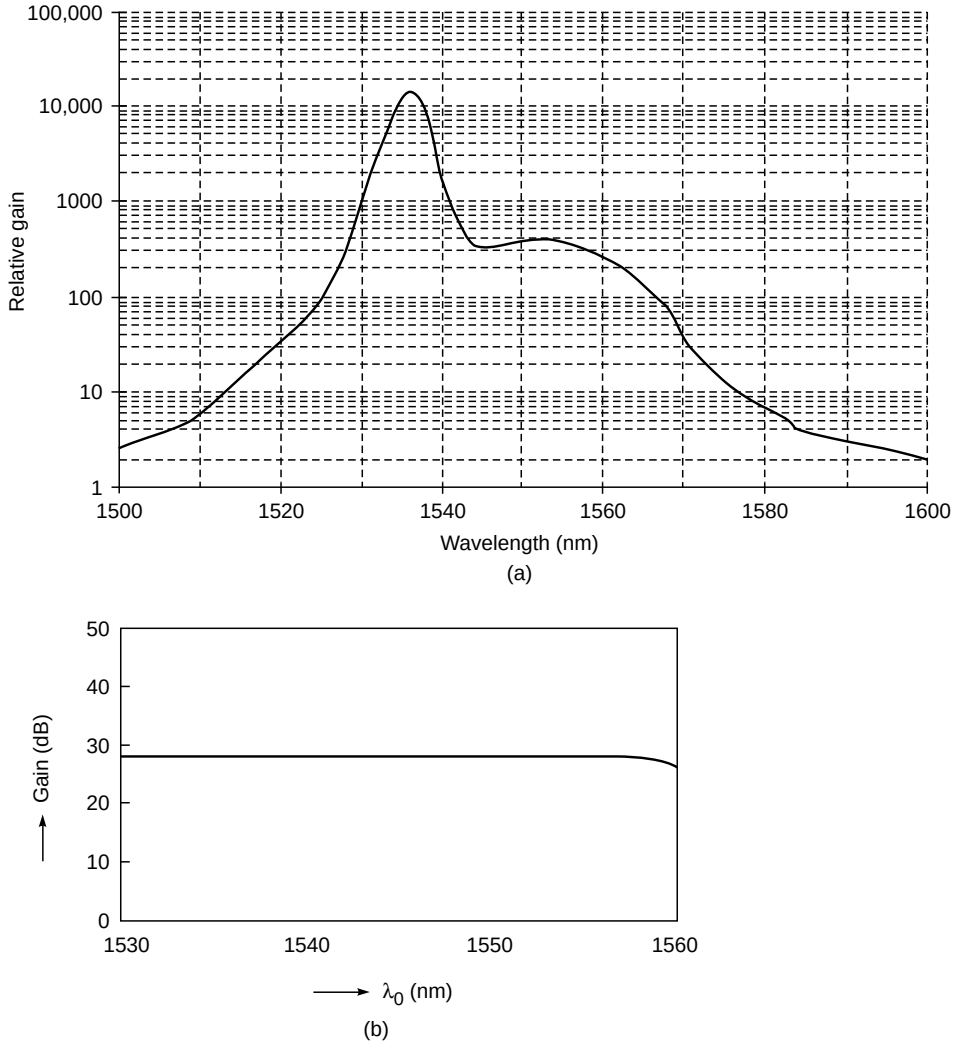


Fig. 26.8 (a) The gain spectrum of a typical erbium-doped fiber amplifier using a 50 mW pump at 980 nm (Adapted from Ref. 10). (b) Through various mechanisms, the gain spectrum of an EDFA can be made almost flat. The above figure corresponds to an EDFA which has an almost flat gain (of about 28 dB) in the wavelength region 1530 to 1560 nm (Adapted from Ref. 11).

and 8. There can be two laser diodes providing the pump power for the erbium-doped fiber (see Fig 26.9). A commercially available EDFA, along with its main characteristics, is shown in Fig. 26.10.

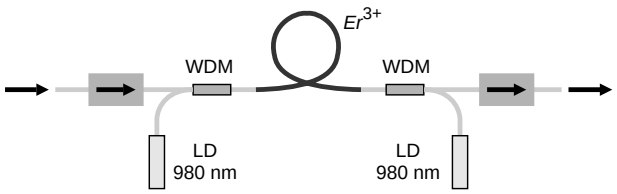


Fig. 26.9 Schematic setup of a simple erbium-doped fiber amplifier with two laser diodes (LDs) providing the pump power for the erbium-doped fiber. Figure adapted from http://www.rpphotonics.com/erbium_doped_fiber_amplifiers.html.

26.1.4 The Resonator

As mentioned earlier, a medium with population inversion is capable of amplification, but in order that it act as an oscillator, a part of the output energy must be fed back into the system. Such feedback is brought about by placing the active medium between a pair of mirrors facing each other (see Fig. 26.3). Such a system formed by a pair of mirrors is referred to as a resonator, a slightly more detailed account of which will be given in Sec. 26.5. The sides of the cavity are usually open, and hence such resonators are also referred to as open resonators. A resonator is characterized by various modes of

Fig. 26.10 An erbium-doped optical fiber amplifier for telecommunication application developed jointly by CGCRI, Kolkata, and NeST, Cochin. The main characteristics are as follows: 32 wavelengths can be simultaneously amplified in the wavelength region from 1532 to 1565 nm. The input power (of each channel) can be between –4 dBm (≈ 0.4 mW) and +3 dBm (≈ 2 mW), and the output power is always 18 dBm (≈ 63 mW) with a gain flatness of ±0.5 dB [Photo courtesy Dr. Shyamal Bhadra of CGCRI and Dr. Suresh Nair of NeST].

oscillation with different field distributions and frequencies (for more details see Ref. 9). One can visualize a mode as a wave having a well-defined transverse amplitude distribution which forms a standing wave pattern. The transverse intensity distribution of the fundamental mode is usually a Gaussian [see Eq. (13)]. Because of the open nature of the resonator, all modes have a finite loss due to the diffraction spillover of energy at the mirrors. In addition to this basic loss, scattering from the laser medium, absorption at the mirrors, and output coupling at the mirrors also contribute to the cavity loss. In an actual laser, the modes that keep oscillating are those for which the gain provided by the laser medium compensates for the losses. When the laser oscillates in steady state, the losses are exactly compensated for by the gain. Since the gain provided by the medium depends on the extent of population inversion, for each mode there is a critical value of population inversion (known as the threshold population inversion) below which that particular mode would cease to oscillate in the laser (see Sec. 26.6).

26.1.5 The Lasing Action

The onset of oscillations in a laser cavity can be understood as follows: Through a pumping mechanism, one creates a state of population inversion in the laser medium placed inside the resonator system. Thus the medium is prepared to be in a state in which it is capable of coherent amplification over a specified band of frequencies.The spontaneous emission occurring inside the resonator cavity excites the various modes of the cavity. For a given population inversion, each mode is characterized by a certain amplification coefficient due to the gain and a certain attenuation coefficient due to the losses in the cavity. The modes for which the losses in the cavity exceed the gain die out. On the other hand, the modes whose gain is higher than the losses get amplified by drawing energy from the laser medium. The amplitude of the mode increases rapidly until the upper level population reaches a value when the gain equals the losses, and the mode oscillates in steady state. When the laser oscillates in steady state, the losses are exactly compensated for by the gain provided by the medium, and the wave coming out of the laser can be represented as a continuous wave.

26.2 THE FIBER LASER

If we put the doped fiber between two mirrors (which act as a resonator), then with an appropriate pump we have a fiber laser (see Fig. 26.11). Indeed in 1961, Elias Snitzer wrapped a flash lamp around a glass fiber (having a 300 μm core doped with Nd^{3+} ions clad in a lower-index glass) and when suitable feedback was applied, the first fiber laser was born (Ref. 12). Thus the fiber laser was fabricated within a year of the demonstration of the first ever laser by Theodore Maiman. These days fiber lasers are commercially available in the market which have applications in many diverse areas because of their flexibility and high power levels. The lower curve in Fig. 26.12 corresponds to the output spectrum of an EDFA just before it starts lasing. As we increase the pump power, the EDFA starts lasing and the spikes correspond to the various resonator modes; the ends of the fiber act as the resonator. Fiber lasers now find widespread applications in welding, cutting, drilling, and in medical surgery.

26.2.1 The MOPA[5]

The term *master oscillator power amplifier* (MOPA) refers to a configuration consisting of a master laser (or seed laser) and an optical amplifier to boost the output power. A special

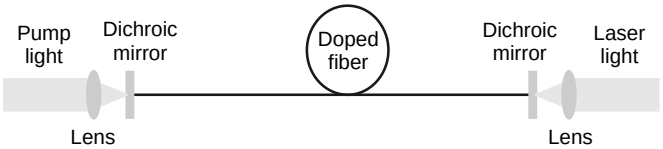


Fig. 26.11 Setup of a simple fiber laser. Pump light is launched from the left side through a dichroic mirror into the core of the doped fiber. The generated laser light is extracted on the right side. Figure adapted from http://www.rpphotonics.com/fiber_lasers.html.

[5] The writeup for this section and Figs. 26.14 to 26.16 have been kindly provided by Mrinmay Pal and Kamal Dasgupta, CGCRI, Kolkata.

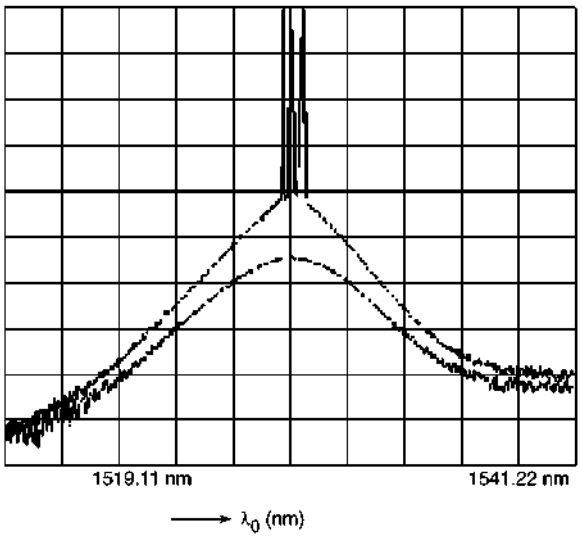


Fig. 26.12 The lower and upper curves show the output of an EDFA just before and after it starts lasing (Photograph courtesy Prof. Thyagarajan and Mr. Mandeep Singh).

case is the master oscillator fiber amplifier (MOFA), where the power amplifier is a fiber device. Although a MOPA configuration is in principle more complex than a laser which directly produces the required output power, the MOPA concept can have the advantage of the ease to achieve the required performance, e.g., in terms of line width, beam quality, or pulse duration if the required power is very high. In the MOFA configuration (shown in Fig. 26.13), the seed laser consists of a 54.7 cm length of EDF (erbium-doped fiber) comprising of two high reflective FBGs (fiber Bragg gratings) written directly on both ends of the EDF. We discussed FBGs in Sec. 15.6 and showed that they are characterized with high reflectivity at a particular wavelength with a very small bandwidth; thus the two FBGs form a resonator. The important characteristics of both the gratings are given in Table 26.1.

Table 26.1 Characteristics of Two Fiber Bragg Gratings Used in MOPA.

Parameter	FBG I	FBG II
Peak wavelength	1549.456 nm	1549.168 nm
3-dB bandwidth	0.344 nm	0.216 nm
Reflectivity	99%	90%

The EDF's numerical aperture is 0.18 NA, and it has 500 ppm Er-ion in the fiber core; the numerical aperture of a fiber is defined in Sec. 27.7. The EDF in the cavity is pumped through a WDM coupler by a 976 nm laser diode of pump power 100 mW. Lasing emission starts at the peak wavelength when the threshold is achieved. Since there is a small offset in the peak wavelengths of the two FBGs, FBG II is slightly stretched to match the peak wavelength with that of the FBG I. When these two wavelengths coincide, laser emission is obtained from the FBG II with maximum output power and very good beam quality. In this MOFA, a seed laser (at 1549.45 nm wavelength) with 1 mW of output power is generated (see Fig. 26.14). To amplify the laser output power, an extra length of 15 m EDF is spliced to the cavity. This extra EDF is pumped by the residual pump power of a 976 nm laser diode. An optical isolator is placed after the amplifier to prevent the back reflection which otherwise degrades the noise figure. In the output, 16.05 dBm (≈ 40 mW) of laser power is obtained (shown in Fig. 26.15). This power can be further enhanced by increasing the pump power.

26.3 THE RUBY LASER

In the first laser fabricated by Maiman in 1960 (Ref. 13), the population inversion was achieved in the following manner. It was made from a single cylindrical crystal of ruby whose ends were flat, with one of the ends completely silvered and the other partially silvered (see Figs. 26.16 and 26.17). Ruby

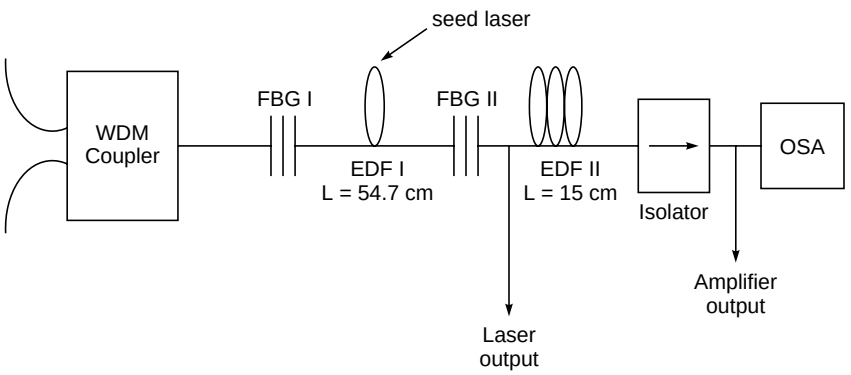


Fig. 26.13 Schematic of the master oscillator power amplifier (MOPA) configuration.

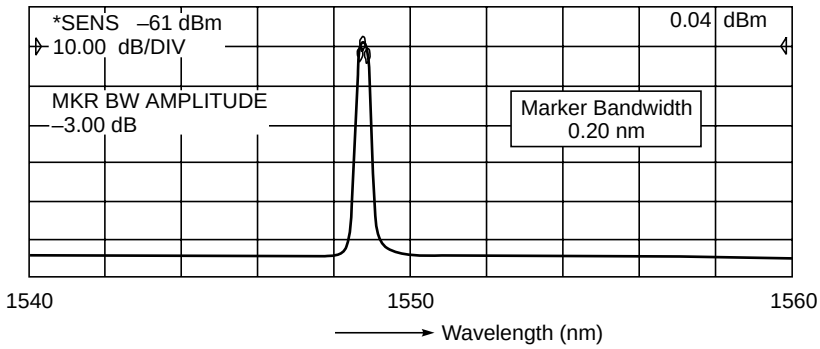


Fig. 26.14 Spectrum of the seed laser. The peak wavelength is 1548.73 nm with peak power of –0.05 dBm and bandwidth of 0.225 nm. Figure courtesy Mrinmay Pal and Kamal Dasgupta, CGCRI, Kolkata.

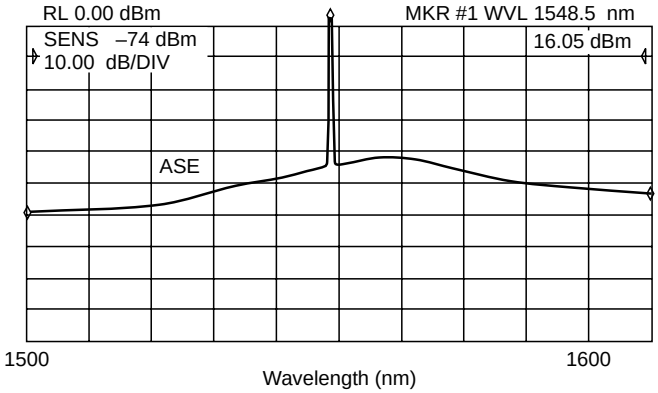


Fig. 26.15 Laser output spectrum from MOPA configuration.

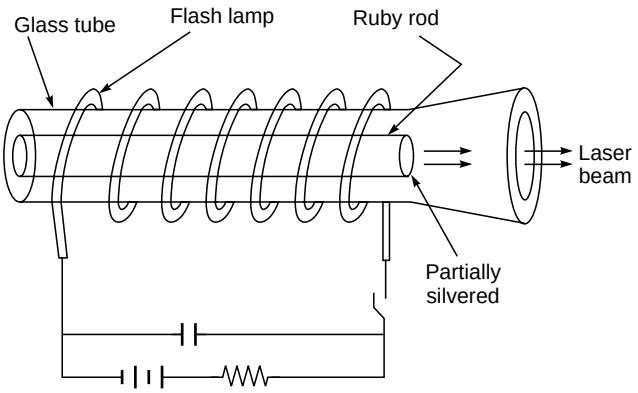


Fig. 26.16 The ruby laser.

consists of Al_2O_3 with some of the aluminum atoms replaced by chromium.[6] The energy states of the chromium ion are shown in Fig. 26.18. The chief characteristic of the energy levels of a chromium ion is the fact that the bands labeled E_1 and E_2 have a lifetime of $\sim 10^{-8}$ s whereas the state marked M has a lifetime of $\sim 3 \times 10^{-3}$ s—the lifetime represents the average time an atom spends in an excited state before making a transition to a lower energy state. A state characterized by such a long lifetime is termed a *metastable state.*

The chromium ion in its ground state can absorb a photon (whose wavelength is around 6600 Å) and make a transition to one of the states in the band E_1. It could also absorb a photon of $\lambda \sim 4000$ Å and make a transition to one of the states in the band E_2—this is known as optical pumping, and the photons which are absorbed by the chromium ions are

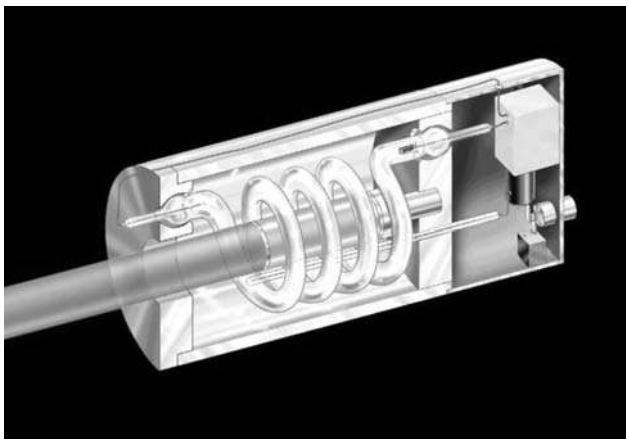


Fig. 26.17 The first ruby laser.

[6] The Al_2O_3 crystal which serves as a medium to suspend the chromium ions is known as the host crystal. The characteristics of the host crystal affect the laser action and also the broadening of the energy levels of the activator atoms which in this case are chromium. For a good lasing action, the ruby crystal consists of about 0.05% (by weight) of chromium; however, higher concentrations of chromium have also been used. For a detailed discussions of host crystals, see Ref. 14.

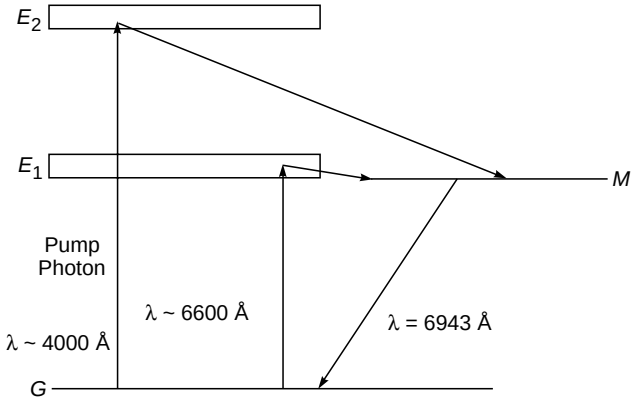


Fig. 26.18 The energy levels of the chromium ion; *G* and *M* represent the ground and metastable states, respectively.

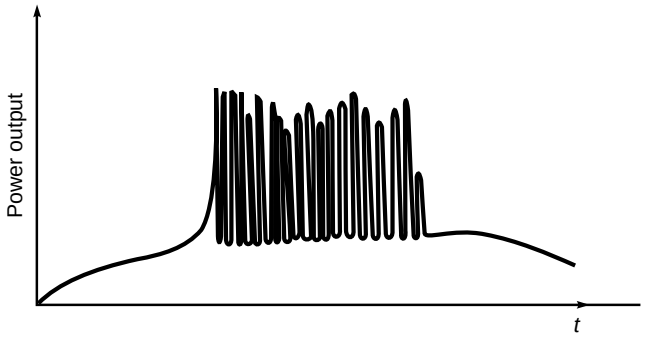


Fig. 26.19 The characteristic spiking of a ruby laser.

produced by the flash lamp (see Fig. 26.16). In either case, it immediately makes a nonradiative transition (in a time ~ 10^{-8} s) to the metastable state *M*—in a nonradiative transition, the excess energy is absorbed by the lattice and does not appear in the form of electromagnetic radiation. Also since state *M* has a very long life, the number of atoms in this state keeps increasing and one may achieve population inversion between states *M* and *G*. Thus we may have a larger number of atoms in states *M* and *G*. Once population inversion is achieved, light amplification can take place, with two reflecting ends of the ruby rod forming a cavity. The ruby laser is an example of a three-level laser.

In the original setup of Maiman, the flash lamp (filled with xenon gas) was connected to a capacitor (see Fig. 26.16) which was charged to a few kilovolts. The energy stored in the capacitor (~ a few thousand Joules) was discharged through the xenon lamp in a few milliseconds. This results in a power which is about a few megawatts. Some of this energy is absorbed by the chromium ions, resulting in their excitation and subsequent lasing action.

26.3.1 Spiking in Ruby Laser

The flash operation of the lamp leads to a pulsed output of the laser. Even in the short period of a few tens of microseconds in which the ruby is lasing, one finds that the emission is made up of spikes of high-intensity emissions as shown in Fig. 26.19. This phenomenon is known as *spiking* and can be understood as follows. When the pump is suddenly switched on to a value much above the threshold, the population inversion builds up and crosses the threshold value, as a consequence of which the photon number builds up rapidly to a value much higher than the steady-state value. Since the photon number is higher than the steady-state value, the rate at which the upper level depletes (because of stimulated transitions) is much higher than the pump rate. Consequently, the inversion becomes below threshold, and the laser action ceases. Thus the emission stops for a few microseconds, within which time the flash lamp again pumps the ground-state atoms to the upper level, and laser oscillations begin again. This process repeats itself till the flash lamp power falls below the threshold value and the lasing action stops (see Fig. 26.19).

26.4 THE He–Ne LASER

We will now briefly discuss the He-Ne laser which was first fabricated by Ali Javan and coworkers at Bell Telephone Laboratories in the United States (Ref. 15). This was also the first gas laser to be operated successfully.

The He-Ne laser consists of a mixture of He and Ne in a ratio of about 10 : 1, placed inside a long, narrow discharge tube (see Figs. 26.20 and 26.21). The pressure inside the rube is about 1 torr.[7] The gas system is enclosed between a pair of plane mirrors or a pair of concave mirrors so that a resonator system is formed. One of the mirrors is of very high reflectivity while the other is partially transparent so that energy may be coupled out of the system.

The first few energy levels of He and Ne atoms are shown in Fig. 26.22. When an electric discharge is passed through the gas, the electrons traveling down the tube collide with the He atoms and excite them (from the ground state F_1) to the levels marked F_2 and F_3. These levels are metastable; i.e., He atoms excited to these states stay in these levels for a sufficiently long time before losing energy through collisions. Through these collisions, the Ne atoms are excited to the levels marked E_4 and E_6 which have nearly the same energy as the levels F_2 and F_3 of He. Thus when the atoms in levels F_2 and F_3 collide with unexcited Ne atoms, they raise

[7] 1 torr = 1 mm of Hg = 133 Pa = 133 N m^{-2}; the unit torr is named after Torricelli, the seventh-century Italian mathematician who invented the mercury manometer.

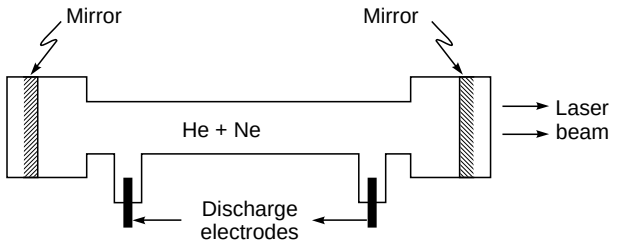


Fig. 26.20 The helium-neon laser.

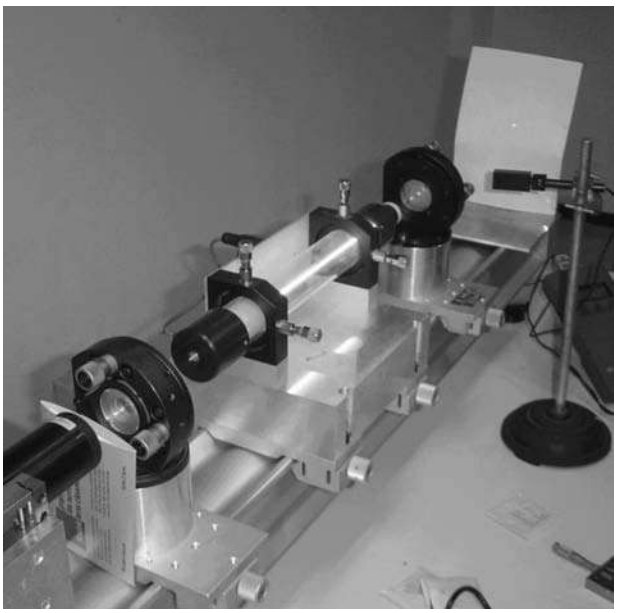


Fig. 26.21 A helium-neon laser demonstration at the Kastler-Brossel Laboratory at Univ. Paris 6. The glowing ray in the middle is an electric discharge producing light in much the same way as a neon light. It is the gain medium through which the laser passes, *not* the laser beam itself, which is visible there. The laser beam crosses the air and marks a red point on the screen to the right. Photograph by Dr. David Monniaux; used with kind permission of Dr. Monniaux. A color photo appears in the insert at the end of the book.

them to the levels E_4 and E_6, respectively. Thus, we have the following two-step process:

1. Helium atom in the ground state F_1 + collision with electron

   $\rightarrow$ Helium atom in the excited state (F_2 or F_3) + electron with lesser kinetic energy.

2. The excited states of He (F_2 or F_3) are metastable[8]—they would not readily lose energy through spontaneous emissions (the radioactive lifetime of these excited states would be about 1 h). However, they can readily lose energy through collisions with Ne atoms:

   He atom in excited state F_3 + Ne atom in ground state

   $\rightarrow$ He atom in ground state + Ne atom in excited state E_6

   Similarly,

   He atom in excited state F_2 + Ne atom in ground state

   $\rightarrow$ He atom in ground state + Ne atom in excited state E_4

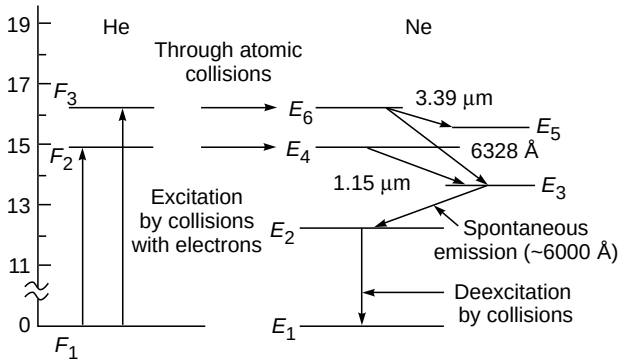


Fig. 26.22 Relevant energy levels of helium and neon.

This results in a sizeable population of the levels E_4 and E_6. The population in these levels happens to be much more than those in the lower levels E_3 and E_5. Thus a state of population inversion is achieved, and any spontaneously emitted photon can trigger laser action in any of the three transitions shown in Fig. 26.22. The Ne atoms then drop down from the lower laser levels to the level E_2 through spontaneous emission. From the level E_2 the Ne atoms are brought back to the ground state through collision with the walls. The transitions from E_6 to E_5, E_4 to E_3, and E_6 to E_3 result in the emission of radiation having wavelengths of 3.39 μm, 1.15 μm, and 6328 Å, respectively. Note that the laser transitions corresponding to 3.39 and 1.15 μm are not in the visible region. The 6328 Å transition corresponds to the well-known red light of the He-Ne laser. A proper selection of different frequencies may be made by choosing end mirrors having high reflectivity over only the required wavelength range. The pressures of the two gases must be chosen so that the condition of population inversion is not quenched. Thus the conditions must be such that there is an efficient transfer of energy from He to Ne atoms. Also, since the level marked E_2 is metastable, electrons colliding with atoms in level E_2 may excite them to level E_3, thus decreasing the

[8] The spectroscopic states corresponding to states F_1, F_2, and F_3 are 1^1S_0, 2^3S_1, and 2^1S_0, respectively.

population inversion. The tube containing the gaseous mixture is also made narrow so that Ne atoms in level E_2 can get de-excited by collision with the walls of the tube. Actually there are a large number of levels grouped around E_2, E_3, E_4, E_5, and E_6 (see Fig. 26.22). Only those levels are shown in the figure which correspond to the important laser transitions. Further details on the He-Ne laser can be found in Refs. 16 and 17.

Gas lasers are, in general, found to emit light, which is more directional and more monochromatic. This is so because of the absence of such effects as crystalline imperfection, thermal distortion, and scattering, which are present in solid-state lasers. Gas lasers are capable of operating continuously without need for cooling.

26.5 OPTICAL RESONATORS

In Sec. 26.1 we briefly discussed that a light beam passing through a suitable medium with population inversion may be amplified. To construct an oscillator which can supply light energy and act as a source of light, one must couple a part of the output back into the medium. This can be achieved by placing the active medium between two mirrors which reflect most of the output energy back to the system; see Fig. 26.3. Such a system of two mirrors represents a *resonant cavity.*

Now, to obtain an output beam, one of the mirrors is made partially reflecting. Thus, imagine a wave that starts from one of the mirrors and travels toward the other. In passing through the active medium, it gets amplified. If the second mirror is partially reflecting, then the wave is partially transmitted and the rest is reflected back toward the first mirror. In traveling to the first mirror, it again gets amplified and returns to the position it has started from. Thus, in between the two mirrors, we have waves propagating along both directions. For resonance, when a wave returns after one round trip, it must be in phase with the existing wave. For this to happen, the total phase change suffered by the wave in one complete round trip must be an integral multiple of 2π, so that standing waves are formed in the cavity. Thus if d represents the length of the cavity, then we may write

$$\frac{2\pi}{\lambda}\, 2d = 2m\pi \qquad m = 1, 2, 3, \ldots \tag{2}$$

where λ is the wavelength of the radiation in the medium enclosed by the cavity, if n_0 represents the refractive index of the medium enclosed by the cavity, then

$$\lambda = \frac{\lambda_0}{n_0}$$

If we put $\lambda_0 = c/\nu$, Eq. (2) gives

$$\nu = \nu_m = m\frac{c}{2n_0 d} \tag{3}$$

which gives the discrete frequencies of oscillation of the modes. If we assume

$$n_0 \approx 1$$

(as in a He-Ne laser), then Eq. (3) simplifies to

$$\nu = \nu_m = m\frac{c}{2d} \tag{4}$$

Different values of m lead to different oscillation frequencies, which constitute the longitudinal modes of the cavity; for further details and for reasons why they are known as longitudinal modes, see any textbook on lasers (see, e.g., Refs. 4, 9, 14, and 16). The frequency difference between adjacent longitudinal modes is given by

$$\delta\nu = \frac{c}{2d} \tag{5}$$

Returning to Eq. (3), we note that for a practical optical resonator, m is a very large number. For example, for an optical resonator of length $d \approx 60$ cm operating at an optical frequency of $\nu \approx 5 \times 10^{14}$ Hz (corresponding to $\lambda \approx 6000$ Å), we obtain

$$m \approx \frac{5\times10^{14}\times 2\times 60}{3\times10^{10}} = 2\times10^{6}$$

Equation (3) tells us that the cavity will support only those frequencies for which the round-trip phase shift is an integral multiple of 2π.

An open resonator consisting of two plane mirrors facing each other is nothing but the Fabry–Perot interferometer discussed in Chap. 16; the main difference is that in a Fabry–Perot interferometer, the spacing between the mirrors is small compared to the transverse dimension of the mirrors while in an optical resonator, the converse is true. Now, in Sec. 16.3 we showed that for a light beam incident normally on a Fabry–Perot interferometer, transmission resonances occur when

$$\delta = \frac{4\pi d}{\lambda_0} = 2m\pi \qquad m = 1, 2, 3, \ldots \tag{6}$$

where we have assumed $n_0 = 1$ and $\cos\theta = 1$ since we have assumed normal incidence. Comparing Eqs. (2) and (6), we readily observe that transmission resonances occur for the modes of the cavity.

Example 26.1 Consider a light beam of central frequency $\nu = \nu_0 = 6 \times 10^{14}$ Hz and a spectral width of 7000 MHz that is

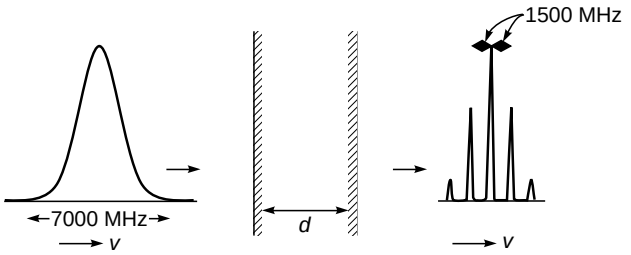


Fig. 26.23 A light beam of central frequency $\nu = \nu_0 = 6 \times 10^{14}$ Hz and a spectral width of 7000 MHz is incident normally on a resonator. The output beam corresponds to the resonant frequencies of the optical cavity.

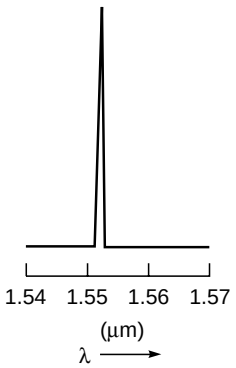


Fig. 26.24 The output of a single longitudinal mode laser.

incident normally on a resonator as shown in Fig. 26.23 with $n_0 = 1$, and $d = 10$ cm. Thus the spacing of two adjacent modes will be

$$\delta\nu = \frac{c}{2d} = 1500 \text{ MHz}$$

Thus the output beam will have frequencies

$$\nu_0 - 2\delta\nu,\ \nu_0 - \delta\nu,\ \nu_0,\ \nu_0 + \delta\nu, \text{ and } \nu_0 + 2\delta\nu$$

corresponding to

m = 399,998; 399,999; 400,000; 400,001; and 400,002

respectively. In the above example, if the reflectivity of one of the mirrors $R = 0.95$ and if output power corresponding to one of the modes is 1 mW, then

Corresponding power incident on mirror **inside** of the cavity will be

$$= 1 \text{ mW}/(1 - 0.95) = 20 \text{ mW}$$

Figures 26.24 and 26.25 show respectively the output of a typical single mode and a typical multilongitudinal mode (MLM) lasers. The wavelength spacing of two adjacent modes in the latter case is about 0.005 μm.

In obtaining Eq. (4) for the various oscillating frequencies, we have assumed that a plane wave can propagate to and fro unmodified inside the resonator. This would not be true in practice since the mirrors of any practical resonator

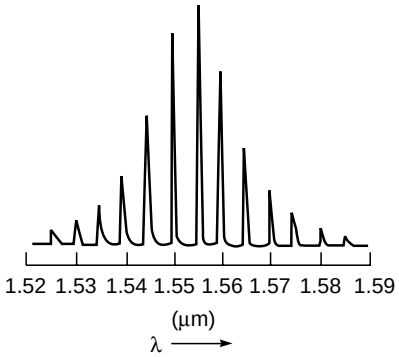


Fig. 26.25 The output of a typical multilongitudinal mode (MLM) laser (Adapted from Ref. 18).

system have finite transverse dimensions and hence only that portion of the wave which strikes the mirror would get reflected; the portion of the wave lying outside the transverse dimension of the mirror will be lost from the resonator. The wave which travels back to the first mirror has now finite transverse dimensions, determined by the transverse dimensions of the mirror. As we have seen in Chap. 18, a beam with a finite transverse dimension diffracts as it propagates. Thus, when the beam comes back to the first mirror, it will have a larger transverse dimension than the mirror. Further, since only that portion of the wave that is intercepted by the mirror is reflected, the remaining portion lying outside the mirror is lost. This loss constitutes a basic loss mechanism and is referred to as diffraction loss.

If we consider a resonator made of mirrors of transverse dimension a and separated by a distance d, then from Eq. (26) of Chap. 18 we see that the wave after reflection at one of the mirrors undergoes diffraction divergence at an angle $\sim\lambda/a$. The angle subtended by one of the mirrors at the other mirror is $\sim a/d$. Hence for diffraction losses to be low,

$$\frac{\lambda}{a} << \frac{a}{d}$$

or

$$\frac{a^2}{\lambda d} >> 1 \qquad \textbf{(7)}$$

The quantity $a^2/\lambda d$ is known as the Fresnel number. As an example, if the resonator mirrors have transverse dimension of 1 cm and are separated by 60 cm, then for a wavelength of 5000 Å, we have

$$\frac{a^2}{\lambda d} \approx 330 >> 1$$

and hence the diffraction losses will be extremely small. The losses in a resonator formed by the plane parallel mirrors would be extremely sensitive to the parallelism of the two mirrors because a slight angular misalignment would cause a

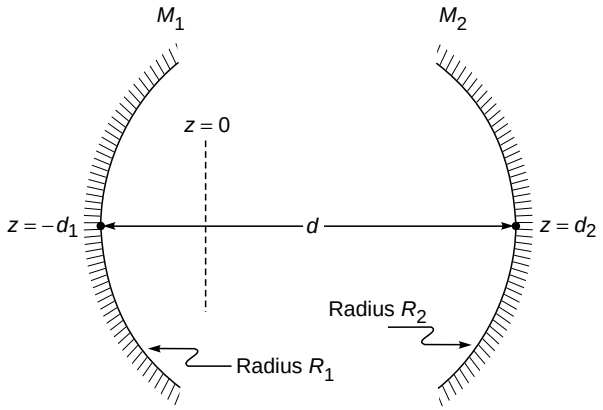


Fig. 26.26 A resonator consisting of two spherical mirrors.

large amount of light energy to escape from the resonator. The loss can be reduced by using spherical mirrors to form the resonant cavity (see Fig. 26.26). The spherical mirrors help in focusing which leads to much less loss due to diffraction spillover. In a stable optical resonator, one can show that there are specific transverse field configurations which maintain their field distribution after successive round trips. These field configurations are referred to as the transverse modes of the resonator. Since oscillations of the laser would occur in modes which do not have excessive loss on successive reflections, it is clear that the modes that would oscillate would be the ones which propagate more or less along the axis of the resonator and which do not diffract appreciably as they propagate between the mirrors.

We will show below that under certain conditions, a Gaussian beam with the appropriate spot size will resonate between the mirrors of the resonator system shown in Fig. 26.26. Let the poles of the mirrors M_1 and M_2 be at $z = z_1 = -d_1$ and at $z = z_2 = +d_2$, respectively. We are assuming the origin somewhere between the mirrors so that both d_1 and d_2 are positive quantities. Thus the distance between the two mirrors is given by

$$d = d_1 + d_2$$

We assume a Gaussian beam propagating along the z direction whose amplitude distribution on the plane $z = 0$ is given by

$$u(x, y) = a \exp\left(-\frac{x^2 + y^2}{w_0^2}\right) \qquad \textbf{(8)}$$

implying that the phase front is plane at $z = 0$; the parameter w_0 is the spot size and also called the *beam waist.* In Sec. 20.5 we showed that as the Gaussian beam propagates along the z direction, the spot size and the radius of curvature of the wave front change and are given by

$$w(z) = w_0 \sqrt{1 + \frac{z^2}{\alpha}} \qquad \text{and} \qquad R(z) = z + \frac{\alpha}{z} \qquad \textbf{(9)}$$

wherer

$$\alpha = \frac{\pi^2 w_0^4}{\lambda^2} \qquad \textbf{(10)}$$

For the Gaussian beam to resonate between the two mirrors, the radii of the phase front (at the mirrors) should be equal to the radii of curvatures of the mirrors:

$$-R_1 = -d_1 - \frac{\alpha}{d_1} \qquad \text{and} \qquad R_2 = d_2 + \frac{\alpha}{d_2}$$

We have used the sign convention such that for the type of mirrors shown in Fig. 26.26, both R_1 and R_2 are positive. Thus

$$\alpha = d_1 (R_1 - d_1) = d_2 (R_2 - d_2)$$

If we use the relation $d_2 = d - d_1$, we readily get

$$d_1 = \frac{(R_2 - d)d}{R_1 + R_2 - 2d} \qquad \text{and} \qquad d_2 = \frac{(R_1 - d)d}{R_1 + R_2 - 2d}$$

We define

$$g_1 = 1 - \frac{d}{R_1} \qquad \text{and} \qquad g_2 = 1 - \frac{d}{R_2} \qquad \textbf{(11)}$$

From the above equations we may write $R_1 = d/(1 - g_1)$ and $R_2 = d/(1 - g_2)$, and we get

$$d_1 = \frac{g_2 (1 - g_1) d}{g_1 + g_2 - 2g_1 g_2} \qquad \text{and} \qquad d_2 = \frac{g_1 (1 - g_2) d}{g_1 + g_2 - 2g_1 g_2}$$

Thus

$$\alpha = d_1 (R_1 - d_1) = \frac{g_1 g_2 d^2 (1 - g_1 g_2)}{(g_1 + g_2 - 2g_1 g_2)^2}$$

Since $\alpha = \pi^2 w_0^4 / \lambda^2$, we get for the spot size at the waist

$$w_0^2 = \frac{\lambda d}{\pi |g_1 + g_2 - 2g_1 g_2|} \sqrt{g_1 g_2 (1 - g_1 g_2)} \qquad \textbf{(12)}$$

For w_0 to be real, we must have $0 \le g_1 g_2 \le 1$, or

$$0 \le \left(1 - \frac{d}{R_1}\right)\left(1 - \frac{d}{R_2}\right) \le 1 \qquad \textbf{(13)}$$

where R_1 and R_2 are the radii of curvature of the mirrors. The above equation represents the stability condition for a resonator consisting of two spherical mirrors. Figure 26.27 shows different resonator configurations. In Figure 26.28 the stability

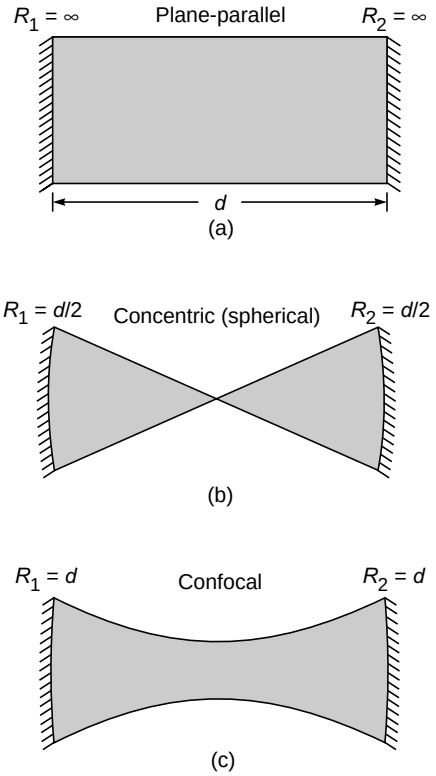


Fig. 26.27 Different configurations of the optical resonator.

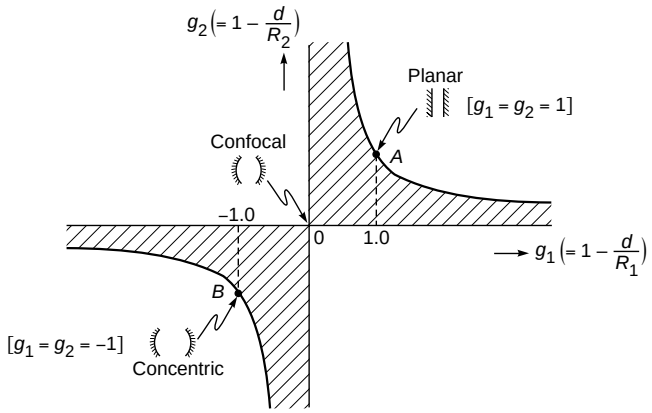


Fig. 26.28 The stability diagram for optical resonators. The shaded region corresponds to stable configurations.

diagram and the shaded region correspond to stable resonator configurations. In a stable resonator, a ray of light can keep bouncing back and forth between the two mirrors without ever escaping from it.

The spot size of the Gaussian beam at the mirrors is given by (see Prob. 26.11)

$$w^2(z_1) = \frac{\lambda d}{\pi}\sqrt{\frac{g_2}{g_1(1-g_1g_2)}} \qquad \textbf{(14)}$$

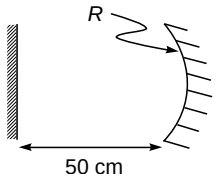


Fig. 26.29 A simple resonator consisting of a plane mirror and a concave mirror of radius R_2 [given by Eq. (15)].

and a similar expression for $w^2(z_2)$. As can be easily seen, when $g_1g_2 \to 0$ or $g_1g_2 \to 1$, $w(z_1)$ or $w(z_2)$ or both become very large and our analysis would not remain valid.

Example 26.2 We consider a simple resonator configuration consisting of a plane mirror and a spherical mirror separated by a distance d (see Fig. 26.29); indeed such a configuration is used to produce a single transverse mode oscillation in a ruby laser. Thus $R_1 = \infty$ and $R_2 = R$, giving $g_1 = 1$ and $g_2 = 1 - d/R$. Simple manipulation of the above equation gives

$$w_0^2 = \frac{\lambda d}{\pi}\sqrt{\left(\frac{R}{d}-1\right)} \qquad \textbf{(15)}$$

Example 26.3 For a typical He-Ne laser ($\lambda = 0.6328$ μm) we may have $d \approx 50$ cm and $R \approx 100$ cm (see Fig. 26.29), giving $g_1 = 1$, and $g_2 = 0.5$, and the resonator configuration is well within the shaded region of Fig. 26.28 and is very much stable. Further, $g_1g_2 = 0.5$, and $w_0 \approx 0.32$ mm. If we increase R to 200 cm, we will get $w_0 \approx 0.38$ mm. For $R < d$, w_0 will become imaginary and the resonator will become unstable.

Example 26.4 We next consider another resonator configuration consisting of two spherical mirrors separated by a distance d = 1.5 m with $R_1 = 1.0$ m and $R_2 = 0.75$ m, giving $g_1 = -0.5$, $g_2 = -1.0$, and $g_1g_2 = 0.5$. Thus the values of g_1 and g_2 are such that the resonator configuration is well within the shaded region of Fig. 26.28 and is very much stable. For $\lambda = 1$ μm one can readily show that $w_0 \approx 0.38$ mm.

Example 26.5 When $g_1 = g_2 = g$, Eq. (12) simplifies to

$$w_0^2 = \frac{\lambda d}{2\pi}\sqrt{\frac{1+g}{1-g}} \qquad \textbf{(16)}$$

The symmetric concentric resonator will correspond to $R_1 = R_2 = d/2$ so that the centers of curvature of both mirrors are at the center [see Fig. 26.27(b)]. Thus $g_1 = g_2 = -1$ and $g_1g_2 = 1$ and w_0 becomes zero!

The symmetric confocal resonator will correspond to $R_1 = R_2 = d$ so that the centers of curvature of both mirrors are at the pole of

the other mirror [see Fig. 26.27(c)]. Thus $g_1 = g_2 = 0$ and $g_1 g_2 = 0$ and

$$w_0 = \sqrt{\frac{\lambda d}{\pi}}$$

Finally for plane parallel mirrors [see Fig. 26.27(a)] $R_1 = R_2 = \infty$, $g_1 = g_2 = 1$ and w_0 becomes infinity!

All three configurations discussed above (concentric, confocal, and planar) lie on the boundary of the stability diagram so that a small variation of the parameters can make the system unstable and will have a very large loss. Thus, it is better to operate inside the shaded region using configurations which have small diffraction loss.

If one chooses a closed resonator system, then the number of modes (which can get amplified and which can oscillate in a resonator of practical dimensions) becomes so large that the output is far from monochromatic. To overcome this problem, one uses open resonators where the number of modes (which can oscillate) is only a few and even single-mode oscillation is possible; furthermore, the open sides of the resonator can be used for optical pumping as in ruby lasers. Because of the open nature of the resonator, all modes have a finite loss due to the diffraction spillover of energy at the mirrors. In addition to this basic loss, scattering from the laser medium, absorption at the mirrors, and output coupling at the mirrors contribute to the cavity loss. One can visualize a mode as a wave having a well-defined transverse amplitude distribution which forms a standing wave pattern. In an actual laser, the modes that keep oscillating are those for which the gain provided by the laser medium compensates for the losses. When the laser oscillates in steady state, the losses are exactly compensated for by the gain. Since the gain provided by the medium depends on the extent of population inversion, for each mode there is a critical value of population inversion (known as the threshold population inversion) below which that particular mode would cease to oscillate in the laser.

26.6 EINSTEIN COEFFICIENTS AND OPTICAL AMPLIFICATION

The consideration which led Einstein to the prediction of stimulated emission was the description of thermodynamic equilibrium between atoms and the radiation field. Consider an atom having two states. Let N_1 and N_2 be the number of atoms (per unit volume) in states 1 and 2, respectively; the levels correspond to energies E_1 and E_2 (see Fig. 26.30). As mentioned earlier, an atom in the lower energy level can absorb radiation and get excited to level E_2. This excitation process can occur only in the presence of radiation. The rate of absorption depends on the density of radiation at the particular frequency corresponding to the energy separation of the two levels. Thus, if

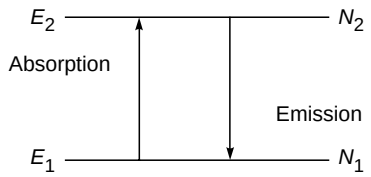


Fig. 26.30 E_1 and E_2 represent the energy levels of an atom. N_1 and N_2 represent the number of atoms (per unit volume) in the energy levels E_1 and E_2, respectively.

$$E_2 - E_1 = \hbar\omega \qquad \textbf{(17)}$$

then the absorption process depends on the energy density of radiation at the frequency ω; this energy density is denoted by $u(\omega)$ and is defined such that

$u(\omega)\,d\omega$ = radiation energy per unit volume within frequency interval ω and $\omega + d\omega$

The rate of absorption is proportional to N_1 and to $u(\omega)$. Thus, we may write

Number of absorptions per unit volume per unit time = $N_1 B_{12} u(\omega)$ **(18)**

where B_{12} is the coefficient of proportionality and is a characteristic of the energy levels.

Let us now consider the reverse process, namely, the emission of radiation at a frequency ω when the atom de-excites from level E_2 to E_1. As mentioned in Sec. 26.1, an atom in an excited level can make a radiative transition to a lower energy level either through spontaneous emission or through stimulated emission. In spontaneous emission, the probability per unit time of the atom making a downward transition is independent of the energy density of the radiation field and depends only on the levels involved in the transition. The rate of spontaneous transitions (per unit volume) from level E_2 to E_1 is proportional to N_2 and thus

$$\frac{dN_2}{dt} = -A_{21} N_2 = -\frac{N_2}{t_{sp}} \qquad \textbf{(19)}$$

where A_{21} represents the coefficient of proportionality and is known as the Einstein A coefficient and depends on the energy level pair and

$$t_{sp} = \frac{1}{A_{21}} \qquad \textbf{(20)}$$

represents the spontaneous lifetime of the upper level. The solution of Eq. (19) is given by

$$N_2(t) = N_2(0)\, e^{-t/t_{sp}} \qquad \textbf{(21)}$$

implying that the population of level 2 reduces by $1/e$ in a time t_{sp}. For example, for the $2P \rightarrow 1S$ transition in a hydrogen atom $A \approx 6 \times 10^8\ s^{-1}$, giving a mean lifetime ($\approx 1/A$) of about 1.6×10^{-9} s (see, e.g., Chap. 21 of Ref. 19). In the case of stimulated emission, the rate of transition to the lower energy level is directly proportional to the number of atoms in the upper energy level as well as to the energy density of the radiation at frequency ω. Thus

Number of stimulated emissions (per unit time per unit volume) = $N_2 B_{21} u(\omega)$

with B_{21} representing the corresponding proportionality constant. The quantities A_{21}, B_{12}, and B_{21} are known as Einstein coefficients and are determined by the atomic system. At thermal equilibrium, the number of upward transitions must be equal to the number of downward transitions. Thus, we may write (at thermal equilibrium)

$$N_1 B_{12} u(\omega) = N_2 A_{21} + N_2 B_{21} u(\omega)$$

or

$$u(\omega) = \frac{A_{21}}{N_1/N_2 B_{12} - B_{21}} \qquad \textbf{(22)}$$

Now according to a fundamental principle in thermodynamics, at thermal equilibrium we have the following expression for the ratio of the populations of two levels:

$$\frac{N_1}{N_2} = \exp\left(\frac{E_2 - E_1}{k_B T}\right) = \exp\left(\frac{\hbar\omega}{k_B T}\right) \qquad \textbf{(23)}$$

where k_B ($= 1.38 \times 10^{-23}$ J K^{-1}) represents the Boltzmann constant and T represents the absolute temperature. Equation (21) is known as Boltzmann's law. Thus, we may write

$$u(\omega) = \frac{A_{21}}{B_{12} e^{\hbar\omega/k_B T} - B_{21}} \qquad \textbf{(24)}$$

Now, at thermal equilibrium, the radiation energy density is given by Planck's law:

$$u(\omega) = \frac{\hbar\omega^3 n_0^3}{\pi^2 c^3} \frac{1}{e^{\hbar\omega/k_B T} - 1} \qquad \textbf{(25)}$$

where n_0 represents the refractive index of the medium. Comparing Eqs. (24) and (25), we obtain[9]

$$B_{12} = B_{21} = B \qquad \text{(say)} \qquad \textbf{(26)}$$

and

$$\frac{A_{21}}{B_{21}} = \frac{\hbar\omega^3 n_0^3}{\pi^2 c^3} \qquad \textbf{(27)}$$

Notice that if we had not assumed the presence of stimulated emission, we would not have been able to arrive at an expression for $u(\omega)$; Einstein in 1917 had predicted the existence of stimulated emission which was later confirmed by rigorous quantum theory.

At thermal equilibrium, the ratio of the number of spontaneous to stimulated emissions is given by

$$\frac{A_{21}}{B_{21} u(\omega)} = e^{\hbar\omega/k_B T} - 1 \qquad \textbf{(28)}$$

We note the following two important points:

1. For a normal optical source, $T \sim 10^3$ K with $\omega \approx 3 \times 10^{15}\ s^{-1}$ (corresponding to $\lambda \approx 6000$ Å) we have

$$\frac{\hbar\omega}{k_B T} \approx \frac{1.054 \times 10^{-34}\ \text{J s} \times 3 \times 10^{15}\ \text{s}^{-1}}{1.38 \times 10^{-23}\ \text{J K}^{-1} \times 10^3} \approx 23$$

giving

$$\frac{A_{21}}{B_{21} u(\omega)} = 10^{10}$$

Thus, when the atoms are in thermal equilibrium, the emission (at optical frequencies) is predominantly due to spontaneous transitions and hence the emission from ordinary light sources is incoherent.

2. From Eq. (27), one can see that the coefficient B_{21} is inversely proportional to ω^3, implying that laser action will become more difficult as we go to higher frequencies.

26.6.1 Population Inversion

In the previous section we assumed that the atom is capable of interacting with radiation of a particular frequency ω. However, if one observes the spectrum of the radiation due to spontaneous emissions from a collection of atoms, one finds that the radiation is not monochromatic but is spread over a certain frequency range. This would imply that energy levels have widths and atoms can interact over a range of frequencies. As an example, in Fig. 26.31 we have shown that the $2P$ level of hydrogen atom has a certain width ΔE ($= \hbar\,\Delta\omega$) so that the

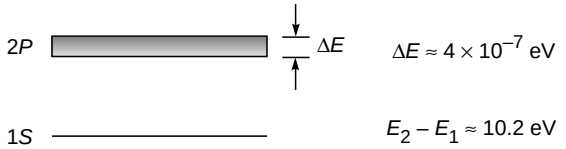


Fig. 26.31 The $2P$ level of hydrogen atom has a certain width ΔE ($= \hbar\,\Delta\omega$) so that the atom can absorb/emit radiation over a range of frequencies $\Delta\omega$.

[9] If levels 1 and 2 are g_1 and g_2 fold degenerate, then $N_1/N_2 = (g_1/g_2) \exp(\hbar\omega/k_B T)$, $B_{12} = B_{21}\, g_2/g_1$, and $A_{21}/B_{21} = n_0^3\ \hbar\omega^3/\pi^2 c^3$.

atom can absorb/emit radiation over a range of frequencies $\Delta\omega$. For the $2P \rightarrow 1S$ transition

$$\Delta E \approx 4 \times 10^{-7} \text{ eV} \quad \Rightarrow \quad \Delta\omega \approx 6 \times 10^{-8} \text{ s}^{-1}$$

Since $\omega_0 \approx 1.55 \times 10^{16}$ s^{-1}, we get

$$\frac{\Delta\omega}{\omega_0} \approx 4 \times 10^{-8}$$

Thus, in general, $\Delta\omega << \omega_0$, showing the spectral purity of the source. We introduce the normalized line shape function $g(\omega)$ such that

- Number of spontaneous emissions per unit time per unit volume with emitted frequency lying between ω and $\omega + d\omega = N_2 A_{21} g(\omega)\, d\omega$
- Number of stimulated emissions per unit time per unit volume with emitted frequency lying between ω and $\omega + d\omega = N_2 B_{21} u(\omega) g(\omega)\, d\omega$
- Number of stimulated absorptions per unit time per unit volume with absorbed frequency lying between ω and $\omega + d\omega = N_1 B_{12} u(\omega) g(\omega)\, d\omega$

Obviously

$$\int_0^\infty g(\omega)\, d\omega = 1$$

Thus the total number of stimulated emissions per unit time per unit volume is given by

$$W_{21} = N_2 \int_0^\infty B_{21} u(\omega) g(\omega)\, d\omega$$

$$= N_2 \frac{\pi^2 c^3}{\hbar t_{sp} n_0^3} \int_0^\infty \frac{u(\omega)}{\omega^3} g(\omega)\, d\omega$$

where we have used Eqs. (27) and (20). Now, for a near monochromatic radiation field (as is indeed the case for the laser), $u(\omega)$ is very sharply peaked at a particular value of ω (say, ω'), and in carrying out the above integration, $g(\omega)/\omega^3$ can be assumed to be essentially constant over the region where $u(\omega)$ is appreciable, to give

$$W_{21} \approx N_2 \frac{\pi^2 c^3}{\hbar t_{sp} n_0^3} \frac{g(\omega')}{\omega'^3} U \qquad \textbf{(29)}$$

where $g(\omega')$ represents the value of the line shape function evaluated at the radiation frequency ω' and U represents the energy density associated with the radiation field.[10]

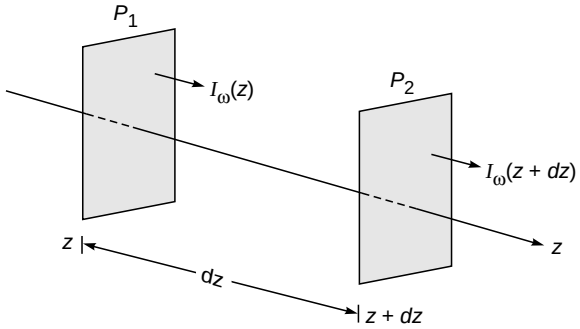


Fig. 26.32 Electromagnetic wave propagating along the z direction through a collection of atoms.

$$U = \int_0^\infty u(\omega)\, d\omega \qquad \textbf{(30)}$$

Now the energy density U and the intensity I_ω are related through the following equation[11] [see Eq. (78) of Chap. 23]

$$I_\omega = vU = \frac{c}{n_0} U \qquad \textbf{(31)}$$

where v (= c/n_0) represents the velocity of the radiation field in the medium, n_0 being its refractive index. (The quantity I_ω represents energy per unit area per unit time, so the mks units of I_ω are therefore J m^{-2} s^{-1}; the quantity U is denoted by $\langle u \rangle$ in Sec. 23.5.) Thus the total number of stimulated emissions per unit time per unit volume is given by

$$W_{21} = N_2 \frac{\pi^2 c^2}{\hbar t_{sp} n_0^2} \frac{g(\omega)}{\omega^3} I_\omega \qquad \textbf{(32)}$$

where we have dropped the prime on ω. Similarly, the number of stimulated absorptions per unit time per unit volume is given by

$$W_{21} = N_1 \frac{\pi^2 c^2}{\hbar t_{sp} n_0^2} \frac{g(\omega)}{\omega^3} I_\omega \qquad \textbf{(33)}$$

We next consider a collection of atoms and let a near monochromatic beam of frequency ω be propagating through it along the z direction. To obtain an expression for the rate of change of the intensity of the beam as it propagates, we consider two planes of area S perpendicular to the z direction at z and $z + dz$ (see Fig. 26.32). The volume of the medium between planes P_1 and P_2 is $S\, dz$, and hence the number of

[10] The argument essentially implies

$$u(\omega) \approx U\delta(\omega - \omega')$$

where $\delta(\omega - \omega')$ represents the Dirac-delta function.

[11] This is analogous to the equation $J = \rho v$, where ρ represents the number of particles per unit volume (all propagating with velocity v) and J represents the number of particles crossing a unit area perpendicular to the direction of propagation per unit time. This can be easily seen from the fact that the number of particles crossing a unit area per unit time is those contained in a cylinder of length v units with unit area of cross section.

stimulated absorptions per unit time is $W_{12}S\,dz$. Since each photon has an energy $\hbar\omega$, the energy absorbed per unit time in the volume element $S\,dz$ is

$$W_{12}\hbar\omega S\,dz$$

Similarly, the corresponding energy gain (because of stimulated emissions) is

$$W_{21}\hbar\omega S\,dz$$

where we have neglected the radiation arising out of spontaneous emissions, because such radiation propagates in random directions and is, in general, lost from the beam. Thus, the net amount of energy absorbed per unit time in the volume element $S\,dz$ is

$$(W_{12} - W_{21})\,\hbar\omega S\,dz$$

If $I_\omega(z)$ represents the intensity of the beam in plane P_1, then the total energy entering the volume element $S\,dz$ per unit time is

$$I_\omega(z)\,S$$

Similarly, if $I_\omega(z + dz)$ represents the intensity in plane P_2, then the total energy leaving the volume element per unit time is

$$I_\omega(z + dz)S = I_\omega(z)S + \frac{\partial I_\omega}{\partial z}dz\,S$$

Hence the net amount of energy leaving the volume element per unit time is

$$\frac{\partial I_\omega}{\partial z}dz\,S$$

This must be equal to the negative of the energy absorbed by the medium between z and $z + dz$. Thus,

$$\frac{\partial I_\omega}{\partial z}S\,dz = -(W_{12} - W_{21})\,\hbar\omega S\,dz$$

$$= -\frac{\pi^2 c^2}{\hbar\, t_{sp}\omega^3 n_0^2}\,g(\omega)\,I_\omega\,\hbar\omega S\,dz\,(N_1 - N_2)$$

or

$$\frac{1}{I_\omega}\frac{\partial I_\omega}{\partial z} = \gamma \qquad \textbf{(34)}$$

where

$$\gamma = \frac{\pi^2 c^2}{\omega^2\, t_{sp}\, n_0^2}(N_2 - N_1)\,g(\omega) \qquad \textbf{(35)}$$

Since the line shape function $g(\omega)$ is very sharply peaked (see Sec. 26.7), the function γ is also sharply peaked. Equation (34) can be readily integrated to give[12]

$$I_\omega(z) = I_\omega(0)e^{\gamma z} \qquad \textbf{(36)}$$

Thus if $N_1 > N_2$, then γ is negative and the intensity of the beam decreases exponentially with z, with the intensity decreasing to $1/e$ of its value at $z = 0$ in a distance $1/\gamma$. Hence at thermal equilibrium, since the number of atoms in the lower level is greater than that in the upper level, the intensity of the beam (as it propagates through the medium) decreases exponentially. On the other hand, if there are more atoms in the excited level than in the lower level (i.e., there is a population inversion), then $\gamma > 0$ and there will be an exponential increase in intensity of the beam; this is known as light amplification.

26.6.2 Cavity Lifetime

In an actual laser system, the active medium (which is capable of amplification) is placed between a pair of mirrors, forming what is known as a resonator (see Sec. 26.5). In order that oscillations be sustained in the cavity, it is essential that the net losses suffered by the beam be compensated for by the gain of the medium. At threshold and under steady-state operation, the two are exactly compensated. To obtain the threshold condition, we first calculate the passive cavity lifetime t_c, which is the time in which the energy $W(t)$ in the (passive) cavity decreases by a factor $1/e$ in the absence of the amplifying medium:

$$W(t) = W(0)\;e^{-t/t_c} \qquad \textbf{(37)}$$

Let d represent the length of the active medium. In one round trip the beam traverses a distance $2d$ through the active medium and gets attenuated by a factor

$$R_1R_2\;e^{-2\alpha_c d}$$

where R_1 and R_2 are the reflectivities of the mirrors at the two ends of the resonator and the term $e^{-2\alpha_c d}$ represents losses caused by absorption, scattering, diffraction, etc. Now the time taken for one round trip is given by

$$t = \frac{2d}{c/n_0}$$

Thus

$$\exp\left[-\frac{2d}{(c/n_0)t_c}\right] = R_1R_2\,e^{-2\alpha_c d} \qquad \textbf{(38)}$$

[12] In obtaining Eq. (36) from Eq. (34), it has been assumed that $N_1 - N_2$ (and hence γ) is independent of I_ω. Such an approximation is valid only for small values of I_ω. For intense light beams (when I_ω becomes very large) saturation of the levels sets in and the attenuation is linear rather than exponential (see, for example, Ref. 4, Sec. 4.2 and 4.3).

giving the following expression for the passive cavity lifetime:

$$\frac{1}{t_c} = \frac{c/n_0}{2d}\,[2\alpha_c\, d - \ln\,(R_1 R_2)] \qquad \textbf{(39)}$$

The cavity lifetime can also be expressed as

$$t_c = \frac{2n_0 d}{c\,[\ln(1/1-x)]} \qquad \textbf{(40)}$$

where

$$x = 1 - R_1 R_2 e^{-2\alpha_c d} \qquad \textbf{(41)}$$

is the fractional loss per round trip.

26.6.3 Threshold Condition

Now, because of population inversion, in one round trip, the beam gets amplified by a factor $e^{2\gamma d}$, and therefore, for the laser oscillation to begin, we must have

$$e^{2\gamma d}\left[R_1\, R_2\, e^{-2\alpha_c d}\right] \geq 1 \qquad \textbf{(42)}$$

which can be rewritten in the form

$$e^{2\gamma\, d} \cdot \exp\left[-\frac{2d}{(c/n_0)\,t_c}\right] \geq 1$$

or

$$\gamma \geq \frac{1}{(c/n_0)\,t_c} \qquad \textbf{(43)}$$

Substituting for γ from Eq. (35), we get

$$N_2 - N_1 \geq \frac{\omega^2 n_0^3 t_{sp}}{\pi^2 c^3 t_c\, g(\omega)} \qquad \textbf{(44)}$$

The equality sign in the above equation gives the threshold population inversion required for the oscillation of the laser. Thus, for the frequency ω, the threshold population inversion is given by

$$(N_2 - N_1)_{th} = \frac{\omega^2 n_0^3 t_{sp}}{\pi^2 c^3 t_c\, g(\omega)} \qquad \textbf{(45)}$$

Now, as we will show in the next section, for a He-Ne laser, $g(\omega)$ is given by

$$g(\omega)\, d\omega = \frac{2}{\Delta\omega_D}\left(\frac{\ln 2}{\pi}\right)^{1/2} \exp\left[-4\ln 2\,\frac{(\omega-\omega_0)^2}{(\Delta\omega_D)^2}\right] d\omega \qquad \textbf{(46)}$$

where

$$\Delta\omega_D = 2\omega_0\left(\frac{2k_B T}{Mc^2}\ln 2\right)^{1/2} \qquad \textbf{(47)}$$

represents the FWHM (full width at half maximum) of the line; in Eq. (47), T represents the absolute temperature of the gas, and M represents the mass of the atom responsible for the lasing transition (neon in the case of a He-Ne laser). Equation (47) describes the line shape function due to the Doppler effect and is shown in Fig. 26.33. Figure 26.34 shows the actual spectrum of a helium-neon laser; the figure shows the very high spectral purity intrinsic to most lasers.

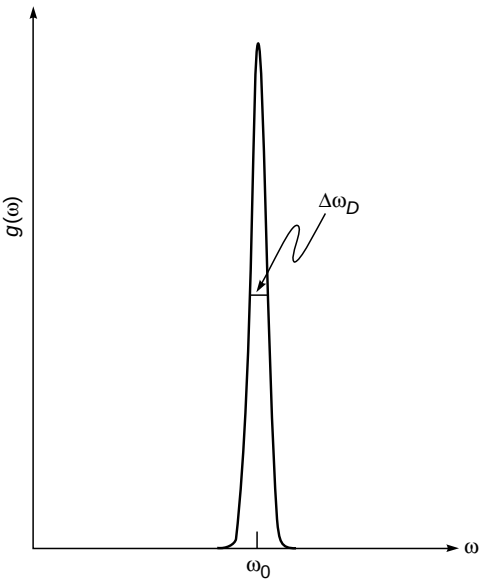


Fig. 26.33 The Gaussian line shape function corresponding to a He-Ne laser; $\Delta\omega_D$ represents the FWHM and usually $\Delta\omega/\omega_0 <<< 1$.

We note that

- The minimum threshold value of $N_2 - N_1$ corresponds to the center of the line where $g(\omega)$ is a maximum, and for the case of Doppler broadening the maximum value is given by

$$g(\omega_0) = \frac{2}{\Delta\omega_D}\left(\frac{\ln 2}{\pi}\right)^{1/2} \qquad \textbf{(48)}$$

  Thus smaller values of $\Delta\omega_D$ will give rise to a smaller threshold value of $(N_2 - N_1)$. Further, as the laser medium is pumped harder and harder, the population inversion between the two levels goes on increasing. The mode that lies nearest to the resonance frequency of the atomic system reaches threshold first and begins to oscillate. As the pumping is further increased, the nearby modes may also reach the threshold and start oscillating.
- From Eq. (45) it also follows that for smaller values of the threshold population inversion $N_2 - N_1$, one must have a small value of t_{sp}, implying strongly allowed transitions; however, for strongly allowed transitions, larger pumping power will be required. In general, for a large value of t_{sp}, population inversion is more easily obtained.

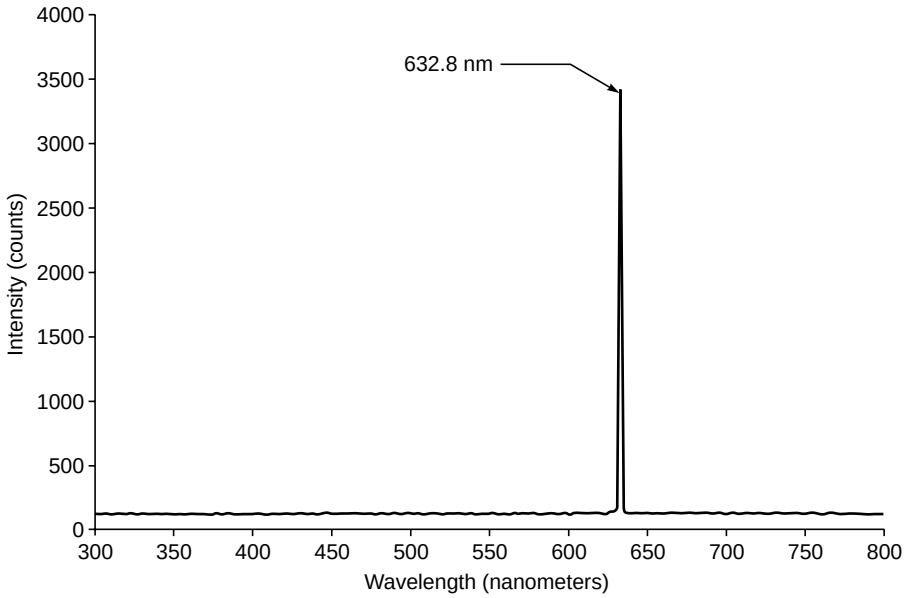


Fig. 26.34 Spectrum of a helium-neon laser showing the very high spectral purity intrinsic to most lasers.

Example 26.6 **Typical parameters for a He-Ne laser**

We consider a He-Ne laser; we assume $T \approx 300$ K. Thus, for the $\lambda_0 \approx 6328$ Å radiation

$$\Delta\omega_D = \frac{2\omega_{21}}{c}\left(\frac{2k_BT}{M}\ln 2\right)^{1/2}$$

$$= \frac{4\pi}{\lambda_0}\left(\frac{2\times 1.38\times 10^{-23}\ \text{J K}^{-1}\times 300\ \text{K}\times 0.693}{20\times 1.67\times 10^{-27}\ \text{kg}}\right)^{1/2}$$

$$\approx 8230\ \text{MHz}$$

implying

$$\Delta\nu_D = \frac{\Delta\omega_D}{2\pi}$$

$$\approx 1310\ \text{MHz}$$

where we have assumed

$$M_{\text{Ne}} \approx 20M_H \approx 3.34\times 10^{-26}\ \text{kg}$$

The frequency variation of $g(\omega)$ is shown in Fig. 26.33. For $\lambda_0 = 6328$ Å,

$$\omega = \frac{2\pi c}{\lambda_0} \approx 2.98\times 10^{15}\ \text{s}^{-1}$$

Thus

$$\frac{\Delta\omega_D}{\omega} \approx 2.8\times 10^{-6}$$

showing that the line shape function is usually a very sharply peaked function. Further,

$$g(\omega_0) = \frac{2}{\Delta\omega_D}\left(\frac{\ln 2}{\pi}\right)^{1/2} \approx 1.1\times 10^{-10}\ \text{s} \qquad \textbf{(49)}$$

(In Sec. 26.7.4 we will show that for a He-Ne laser, the Doppler broadening dominates over natural broadening and collisional broadening). If we assume a cavity with the following values of various parameters

$d = 60$ cm $\quad n_0 \approx 1 \quad R_1 \approx 1 \quad R_2 \approx 0.98 \quad \alpha_c \approx 0$

we would get

$$t_c \approx 2\times 10^{-7}\ \text{s}$$

Further, for the He-Ne laser

$$t_{sp} \approx 10^{-7}\ \text{s} \qquad n_0 \approx 1 \qquad \lambda_0 \approx 6328\ \text{Å}$$

giving

$$(N_2 - N_1)_{\text{th}} \approx 1.5\times 10^8\ \text{cm}^{-3}$$

For a given value of $N_2 - N_1$ (which is greater than the threshold value), a typical gain curve $\gamma(\nu)$ (which has a bandwidth of about 1300 MHz) is shown in Fig. 26.35. The horizontal line represents the value of

$$\frac{1}{(c/n_0)t_c} \qquad \textbf{(50)}$$

For $n_0 \approx 1$ and $t_c \approx 2\times 10^{-7}$ s, the above value is $\approx 1.7\times 10^{-4}$ cm^{-1}. If we assume a 60 cm long He-Ne laser, then the longitudinal mode spacing is given by

$$\delta\nu = \frac{c}{2d} \approx 250\ \text{MHz} \qquad \textbf{(51)}$$

and, as shown in Fig. 26.35, there will be seven longitudinal modes for which gain will exceed loss and which will oscillate. On the other hand, if d is only 10 cm, then

$$\delta\nu = 1500\ \text{MHz}$$

and we may have single mode oscillation; (the value of t_c and hence the position of the horizontal line in Fig. 26.35 would have changed slightly).

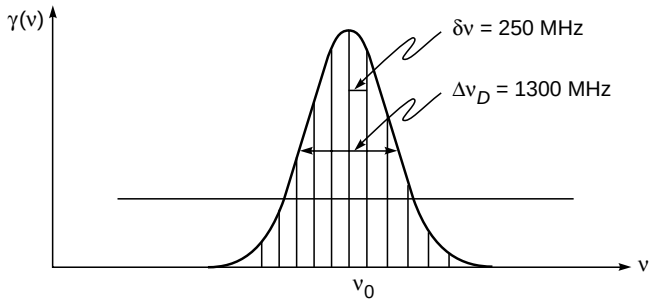


Fig. 26.35 For a given value of $(N_2 - N_1)$ a typical variation of the gain curve $\gamma(\nu)$. The vertical lines show the longitudinal modes of the cavity.

26.7 THE LINE SHAPE FUNCTION

Since the line shape function $g(\omega)$ determines the threshold population inversion [see Eq. (45)], we digress here to discuss some of the typical forms of $g(\omega)$ corresponding to different conditions.

We first consider the Doppler broadening which is due to the thermal motion of gas atoms. Also, in the He-Ne laser (which is probably the most popular laser), the line broadening mechanism is mainly due to Doppler broadening.

26.7.1 Doppler Broadening

In astronomy, we can determine how fast the stars or galaxies are moving (either directly away from or directly toward us) by measuring the Doppler shift of spectral lines. For $v/c << 1$,

$$\omega - \omega_0 = \pm \omega_0 \frac{v}{c} \tag{52}$$

the + sign corresponds to when the source of light is moving toward the observer and the – sign corresponds to when the source of light is moving away from the observer (see Sec. 31.3). Thus when the star is moving away from the observer, the measured frequency is slightly less than the actual value, leading to the well-known *red shift* of spectral lines. Now, the probability that an atom has a *z* component of velocity lying between v_z and $v_z + dv_z$ is given by the Maxwell distribution

$$P(v_z)\,dv_z = \left(\frac{M}{2\pi k_B T}\right)^{1/2} \exp\left(-\frac{M v_z^2}{2k_B T}\right) dv_z \tag{53}$$

where M is the mass of the atom and T the absolute temperature of the gas. Obviously

$$\int_{-\infty}^{+\infty} P(v_z)\,dv_z = 1$$

as indeed it should be. Now, the probability $g(\omega)d\omega$ that the transition frequency lies between ω and $\omega + d\omega$ is equal to the probability that the z component of the velocity of the atom lies between v_z and $v_z + dv_z$, where [using Eq. (52)]

$$v_z = \frac{\omega - \omega_0}{\omega_0} c \tag{54}$$

Thus

$$g(\omega)\,d\omega = \frac{c}{\omega_0}\left(\frac{M}{2\pi k_B T}\right)^{1/2} \exp\left[-\frac{Mc^2}{2k_B T}\frac{(\omega-\omega_0)^2}{\omega_0^2}\right] d\omega \tag{55}$$

which corresponds to a Gaussian distribution. The line shape function peaks at ω_0, and the FWHM is given by

$$\Delta\omega_D = 2\omega_0\left(\frac{2k_B T}{Mc^2}\ln 2\right)^{1/2} \tag{56}$$

where the subscript D implies that we are considering Doppler broadening. In terms of $\Delta\omega_D$, Eq. (55) can be written as

$$g(\omega)\,d\omega = \frac{2}{\Delta\omega_D}\left(\frac{\ln 2}{\pi}\right)^{1/2} \exp\left[-4\ln 2\frac{(\omega-\omega_0)^2}{(\Delta\omega_D)^2}\right] d\omega \tag{57}$$

A typical plot of the Gaussian line shape function corresponding to the He-Ne laser is shown in Fig. 26.33. Since $g(\omega)$ is a very sharply peaked around $\omega = \omega_0$

$$\int_0^{\infty} g(\omega)d\omega \approx \int_{-\infty}^{+\infty} g(\omega)d\omega = 1$$

26.7.2 Natural Broadening

The frequency spectrum associated with spontaneous emission is described by the Lorentzian line shape function

$$g(\omega) = \frac{1}{2\pi t_{sp}} \cdot \frac{1}{(\omega-\omega_0)^2 + \dfrac{1}{4t_{sp}^2}} \tag{58}$$

where

$$t_{sp} = \frac{1}{A_{21}} \tag{59}$$

represents the spontaneous emission lifetime. The FWHM of the Lorentzian is

$$\Delta\omega = \frac{1}{t_{sp}} = A_{21} \tag{60}$$

Thus, in terms of $\Delta\omega$, Eq. (58) can be written in the form

$$g(\omega) = \frac{\Delta\omega}{2\pi} \frac{1}{(\omega-\omega_0)^2 + (\Delta\omega/2)^2} \tag{61}$$

giving

$$g(\omega_0) = \frac{2}{\pi(\Delta\omega)} \tag{62}$$

Further

$$\int_0^{\infty} g(\omega)d\omega \approx \int_{-\infty}^{+\infty} g(\omega)d\omega = 1 \tag{63}$$

26.7.3 Collisional Broadening

In a gas, random collisions occur between the atoms. In such a collision process, when the atoms are very close to one another, the energy levels of the atoms change due to their mutual interaction. This leads to a Lorentzian line shape function given by

$$g(\omega) = \frac{\tau_0}{\pi}\frac{1}{1+(\omega-\omega_0)^2\tau_0^2} \tag{64}$$

where τ_0 represents the mean time between collisions; the derivation of Eq. (64) is given in many places—see, e.g., Sec. 8.8.2 of Ref. 17. The FWHM will be

$$\Delta\omega_c = \frac{2}{\tau_0}$$

In a typical gas laser $\tau_0 \sim 10^{-6}$ s, giving

$$\Delta\omega_c \approx 2 \text{ MHz}$$

or

$$\Delta\nu_c \approx 0.3 \text{ MHz}$$

For the He-Ne laser, the Doppler line width is about 1300 MHz (see Example 26.6); on the other hand, the natural broadening is about 20 MHz, and the collision broadening at 0.5 torr is about 0.64 MHz. Thus, for He-Ne laser parameters, the Doppler broadening dominates over natural broadening and collision broadening.

The various line broadening mechanisms can be broadly classified under homogeneous and inhomogeneous broadening. Certain line broadening mechanisms, such as collision broadening or natural broadening, act to broaden the response of each atom in an identical fashion; such broadening mechanisms come under the class of homogeneous broadening. On the other hand, Doppler broadening or broadening produced due to local inhomogeneities in a crystal lattice act to shift the central frequency of the response of individual atoms by different amounts and thereby lead to an overall broadening of the response of the atomic system. Such a form of broadening is referred to as inhomogeneous broadening. If the effects which cause the inhomogeneous broadening are random in origin, then the broadened line is Gaussian in shape. In contrast, homogeneous broadening in general results in a Lorentzian line shape.

We return to Eq. (45) and notice that to have a low threshold value of population inversion:

1. The value of t_c should be large; i.e., the losses in the cavity must be small.
2. The value of $g(\omega)$ at the center of the line is $\approx 0.64/\Delta\omega$ for a Lorentzian line and $\approx 0.94/\Delta\omega$ for a Gaussian line [see Eqs. (62) and (48)]. Thus, smaller the value of $\Delta\omega$ (the width of the line), the smaller the threshold population inversion.
3. Smaller values of t_{sp} (i.e., strongly allowed transitions) also lead to smaller values of threshold inversion. Note that, for shorter relaxation rates, larger pumping power is required to maintain a given amount of population inversion. In general, population inversion is more easily obtained on transitions that have longer relaxation times.
4. The value of $g(\omega)$ at the center of the line is inversely proportional to $\Delta\omega$, which, for example, in the case of Doppler broadening is proportional to ω [see Eq. (47)]. Thus, the threshold population inversion increases approximately in proportion to the third power of ω (apart from the frequency dependence of the other terms). Hence it is much easier to obtain laser action at infrared wavelengths than in the ultraviolet region.

26.8 TYPICAL PARAMETERS FOR A RUBY LASER

To get an idea of the magnitude of population inversion required for oscillation, we consider a ruby laser (see Sec. 26.3). Let us consider the laser to be oscillating at the frequency corresponding to the peak of the emission line. We assume a concentration of 0.05% of Cr^{3+} ions in the crystal; this corresponds to a population of $N = 1.6 \times 10^{19}$ Cr^{3+} ions cm^{-3}. For the case of ruby, the line is homogeneously broadened, and the value of $g(\omega)$ at the peak of the line is $2/(\pi\Delta\omega)$. Hence the threshold population inversion density is

$$(N_2 - N_1)_{th} = \frac{\omega^2 n_0^3 t_{sp}}{\pi^2 c^3 t_c g(\omega)}$$

$$= \frac{4\pi^2 n_0^3}{\lambda_0^3} \cdot \frac{\Delta\omega}{\omega} \cdot \frac{t_{sp}}{t_c} \tag{65}$$

where λ_0 is the free space wavelength, t_{sp} is the spontaneous relaxation time of the upper laser level, and t_c is the cavity

lifetime. For ruby laser transition, one has

$$\lambda_0 = 6943\ \text{Å} \quad \Rightarrow \quad \omega \approx 2.715 \times 10^{15}\ \text{s}^{-1}$$

$$\Delta\omega \approx 9.4 \times 10^{11}\ \text{s}^{-1} \qquad t_{sp} \approx 3 \times 10^{-3}\ \text{s} \qquad n_0 \approx 1.76$$

where n_0 (= 1.76) represents the refractive index of ruby. If we assume a cavity length of 5 cm and a loss per round trip of 10%, then $x = 0.1$ and using Eq. (40), we get

$$t_c \approx 6 \times 10^{-9}\ \text{s}$$

Substituting all these values in Eq. (65), we get for the threshold population inversion density

$$(N_2 - N_1)_{th} \approx 1.1 \times 10^{17} \quad \text{Cr}^{3+}\ \text{ions cm}^{-3}$$

Since the total density of Cr^{3+} ions in ruby is about 1.6×10^{19} cm^{-3}, the fractional excess population required is very small.

We will next calculate approximately the minimum power required to maintain population inversion. Since t_{sp} represents the spontaneous relaxation time of the upper laser level, the number of atoms decaying per unit time from the upper laser level is approximately N_2/t_{sp}. For each atom lifted to level 2, one has to supply at least an amount of energy given by $h\nu_p$, where ν_p represents the average pump frequency. Hence to maintain N_2 atoms in level 2, the minimum power P to be spent (per unit volume of the active material) is given by

$$P = \frac{N_2 h\nu_p}{t_{sp}} \tag{66}$$

Now, since $(N_2 - N_1)_{th} << N$ (where N represents the total number of atoms per unit volume), we may write

$$N_2 \approx \frac{N}{2} \tag{67}$$

Thus, the minimum pumping power per unit volume required to maintain population inversion in a three-level laser system is

$$P_{th} \approx \frac{N}{2}\frac{h\nu_p}{t_{sp}} \tag{68}$$

Taking the average pumping frequency as $\nu_p \approx 6.25 \times 10^{14}$ Hz (which is averaged over the green and violet absorption bands), we obtain

$$P_{th} \approx \frac{1.6 \times 10^{19}}{2} \times \frac{6.6 \times 10^{-34} \times 6.25 \times 10^{14}}{3 \times 10^{-3}}$$

$$\approx 1100\ \text{W cm}^{-3}$$

If we assume that the efficiency of the pumping source is 25% and also that only 25% is absorbed in passage through the ruby rod, then the electrical threshold power comes out to be about 18 kW cm^{-3} of the active material. This is consistent with the threshold powers determined experimentally.

The threshold power calculation is particularly simple for the ruby laser where only three levels are involved. In general, to calculate the steady-state population difference between the actual levels involved in the laser transition (for a given pumping rate) and to know whether an inversion of population is achievable in a transition—and if so, what would be the minimum pump power required to maintain a steady population inversion for continuous wave operation of the laser—it is necessary to solve equations that govern the rate at which populations of various levels change under the action of a pump and in the presence of laser radiation. These equations are referred to as *rate equations* and have been discussed at many places; see, for example, Refs. 4, 9, 16, and 17. Even for a three-level laser system, the equation $N_2 = N/2$ [see Eq. (67)] is only approximately valid, and to obtain a more accurate expression, it is necessary to solve the rate equations.

26.9 MONOCHROMATICITY OF THE LASER BEAM

Figure 26.36 shows the various line widths associated with a laser. The broad solid curve represents the spectral width due to Doppler broadening of the laser medium. As an example, if we consider the He-Ne laser operating at 6328 Å, the Doppler broadened line width is about 1300 MHz. Inside the broad curve the cavity modes are shown as sharp peaks. The frequency separation between two adjacent cavity modes is $c/2d$ [see Eqs. (6) and (51)] which for a typical laser cavity 60 cm long corresponds to 250 MHz; this is much less than the Doppler width (see Example 26.6). As we discussed earlier, the cavity modes are also broadened due to the various losses in the cavity. Thus, for a 60 cm long cavity specified by a fractional loss per round trip of 4×10^{-2}, the width of the

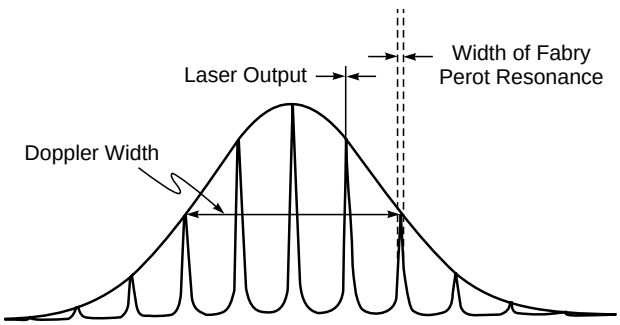


Fig. 26.36 The solid curve represents a typical Doppler broadened spectral line. The closely spaced cavity modes are shown as narrow peaks inside the curve. The sharp line represents the output of the laser (Ref. 21).

cavity mode is about 1.5 MHz. This is much smaller than the spacing between adjacent cavity modes. When the losses in the cavity are compensated for by the active medium placed inside the cavity, the resultant emission becomes extremely narrow and is limited due to the presence of spontaneous emissions (which are random) and the fluctuations in the resonator parameters. The ultimate line width of an oscillating laser determined solely by random spontaneous emissions can be shown to be given by (see, e.g., Ref. 22)

$$(\delta\nu)_{sp} \approx \frac{2\pi\left(\Delta\nu_p\right)^2 h\nu_0}{P^0}$$

where ν_0 is the frequency of oscillation, P^0 is the output power, and

$$\Delta\nu_p = \frac{1}{2\pi t_c}$$

is known as the passive cavity line width, t_c being the cavity lifetime (see Sec. 26.6.2). The subscript "sp" refers to the fact that the line width is due to spontaneous emissions. The decrease in $(\delta\nu)_{sp}$ with an increase in power output is due to the fact that for a given mirror transmittance, an increase in P^0 corresponds to an increase in laser power inside the resonator cavity, and this leads to the dominance of stimulated emissions over spontaneous emission.

As a typical example, $\Delta\nu_p \approx 1$ MHz, $P^0 = 1$ mW $= 10^{-3}$ W, and $h\nu = 2 \times 10^{-19}$ J (corresponding to the red region of the spectrum) so that

$$(\delta\nu)_{sp} \approx 10^{-3}\ \text{Hz}$$

an extremely small quantity indeed! Thus the ultimate monochromaticity is determined by the spontaneous emissions occurring inside the cavity because the radiation coming out due to spontaneous emission is incoherent. However, in practice, the monochromaticity is limited by external factors such as temperature fluctuations, mechanical vibrations of the optical cavity, etc. For example, if we assume the oscillation frequency of a mode is given by Eq. (4), then the change in frequency $\Delta\nu$ caused by a change in length Δd is given by

$$\frac{\Delta\nu}{\nu} = \frac{\Delta d}{d}$$

Thus for $d \approx 50$ cm, if we assume a stability of $\Delta d \approx 1$ Å then for $\nu \approx 5 \times 10^{14}$ Hz

$$\Delta\nu \approx 10^5\ \text{Hz}$$

which is much much larger than $(\delta\nu)_{sp}$. Note that $\Delta\nu \approx 10^5$ Hz corresponds to $\Delta\lambda \approx 10^{-6}$ Å. Indeed, for a single-mode He-Ne laser, we can have $\Delta\nu \approx 10^5$ Hz. On the other hand, for a multimode He-Ne laser $\Delta\lambda \sim 0.02$ Å, implying a coherence length of about 20 cm.

26.10 RAMAN AMPLIFICATION AND RAMAN LASER

We will first discuss the physics of Raman effect. When a monochromatic light beam gets scattered by a transparent substance, one of the following may occur:

1. Over 99% of the scattered radiation has the same frequency as that of the incident light beam; this is known as Rayleigh scattering, discussed in Sec. 7.6. The sky looks blue because of Rayleigh scattering, and the light that comes out from the side of the optical fiber (see Fig. 27.2) is also due to Rayleigh scattering.
2. A very small portion of the scattered radiation has a frequency different from that of the incident beam—this may arise due to one of the following three processes:
   (i) The incident radiation may lead to translatory motion of the molecules–this would result in shift of frequency which is usually very small and difficult to measure. This is known as Brillouin scattering.
   (ii) A part of the energy $h\nu$ of the incident photon is taken over by the molecule in the form of rotational (or vibrational) energy, and the scattered photon has a smaller energy $h\nu'$. This leads to what are known as Raman–Stokes lines [see Figs. 26.37(a) and 26.38].
   (iii) On the other hand, the photon can undergo scattering by a molecule which is already in an excited state. The molecule can de-excite to one of the lower energy states, and in the process, the incident photon can take up this excess energy and come out with a higher frequency. This leads to what are known as Raman anti-Stokes lines [see Figs. 26.37(b) and 26.38].

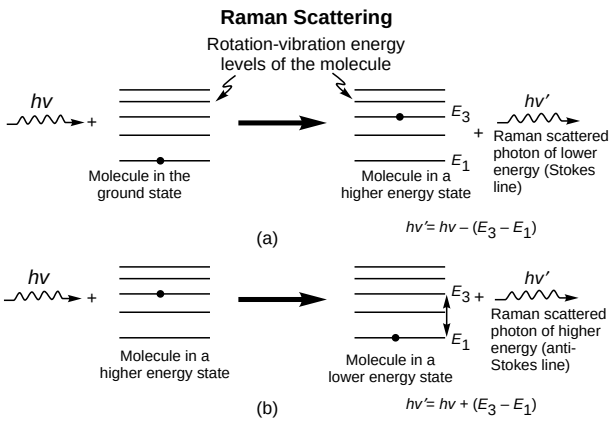


Fig. 26.37 The generation of the Raman–Stokes and the Raman anti-Stokes lines.

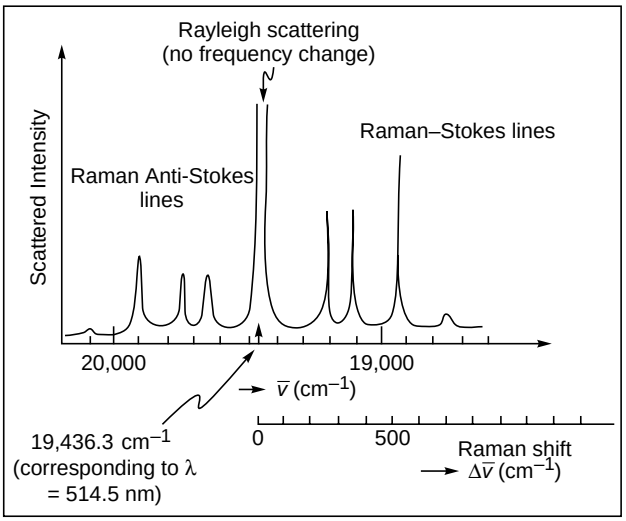


Fig. 26.38 Raman spectra of CCl_4 excited by 514.5 nm line of an Argon-ion laser.

The difference energy, which is $h\nu - h\nu'$ for the Raman–Stokes line and $h\nu' - h\nu$ for the Raman anti-Stokes line, would therefore correspond to the energy difference between the rotational (or vibrational) energy levels of the molecule and would therefore be a characteristic of the molecule itself.

The quantity $h\nu - h\nu'$ or $h\nu' - h\nu$ is usually referred to as the *Raman shift* (see Fig. 26.38) and is independent of the frequency of the incident radiation. Through a careful analysis of the Raman spectra, one can determine the structure of molecules; there lies the tremendous importance of the Raman effect. The intensity distribution of a typical Raman spectrum for the CCl_4 molecule is shown in Fig. 26.38.

In spectroscopy, the energy levels of atoms or molecules and also the energy of a photon are measured in wave number units which are obtained by dividing the energy by hc, where h ($\approx 6.56 \times 10^{-27}$ erg s) is Planck's constant and c ($\approx 3 \times 10^{10}$ cm s^{-1}) is the speed of light in free space—in spectroscopy everyone uses cgs units! In the case of molecular (or atomic) energy levels, these are usually denoted by the symbol T_n:

$$T_n = \frac{E_n}{hc}$$

The photon's energy is $h\nu$ and therefore, in wave number units

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

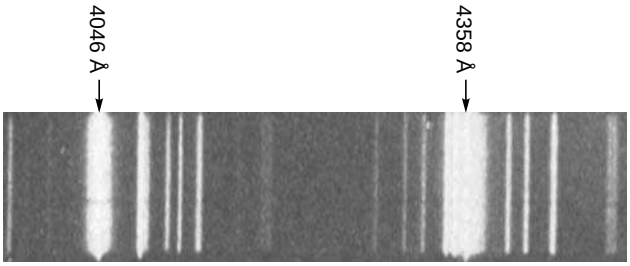


Fig. 26.39 The observed Raman spectra of CCl_4 for the 4046Å and 4358Å lines of mercury lamp. The photograph is adapted from the 1930 Nobel lecture of C.V.Raman.

is just the inverse of the wavelength and is usually denoted by the symbol $\bar{\nu}$. Thus

$$\bar{\nu} = \frac{1}{\lambda}$$

Now, the energy levels of the hydrogen atom in wave number units are given by

$$T_n = \frac{E_n}{hc} = -\frac{R}{n^2} \qquad n = 1, 2, 3, \ldots$$

where R ($\approx 109{,}678$ cm^{-1}) is known as the Rydberg constant and n (= 1, 2, 3, . . .) is the total quantum number of the state. Thus corresponding to the $n = 3$ to $n = 2$ transition (one of the lines of the Balmer series) we will get a photon of wave number

$$\bar{\nu} = -R\left(\frac{1}{9} - \frac{1}{4}\right) = \frac{5}{36} \times 109{,}678 \approx 15{,}233 \text{ cm}^{-1}$$

The inverse of the above number ($\approx 6.56 \times 10^{-5}$ cm) represents the wavelength of the emitted photon.

Figure 26.38 shows the intensity distribution of the Raman spectrum of CCl_4 molecule[13] when the incident radiation corresponds to the argon-ion laser line having a wavelength of 5.145×10^{-5} cm; in wave number units the value is 19,436.3 cm^{-1}. The central peak in the figure corresponds to this wavelength and is due to Rayleigh scattering. The Raman shift for the Stokes lines is the same as for the anti-Stokes lines although the latter is much weaker. This is so because at room temperature, the number of molecules in the ground state is much larger than the molecules present in excited states. This leads to very low intensities of the Raman anti-Stokes lines. The actual Raman spectrum of the CCl_4 molecule for the 4046 Å lines of mercury lamp is shown in Fig. 26.39. The photograph is adapted from the 1930 Nobel lecture of C. V. Raman. On February 28, 1928, K. S. Krishnan

[13] The Raman spectrum from a mixture of hydrogen and deuterium molecules (when the mixture is illuminated by a laser beam at l = 488 nm) is discussed in Ref. 23.

and C. V. Raman observed the Raman effect in several organic vapors such as pentane, which they called "the new scattered radiation." Raman made a newspaper announcement on February 29 and on March 8, 1928; he communicated a paper entitled "A Change of Wavelength in Light Scattering" to *Nature*; the paper was published on April 21, 1928. Although in the paper, he acknowledged that the observations were made by K. S. Krishnan and himself, the paper had Raman as the author and therefore the phenomenon came to be known as the Raman effect although many scientists (particularly in India) kept on referring it as the Raman–Krishnan effect. Subsequently, several papers were written by Raman and Krishnan. Raman got the Nobel Prize in 1930 for "his work on the scattering of light and for the discovery of the effect named after him." At about the same time, Landsberg and Mandel′shtam (in Russia) were also working on light scattering, and according to Mandel'shtam, they observed the "Raman lines" on February 21, 1928. But the results were presented in April 1928, and it was only on May 6, 1928, that Landsberg and Mandel'shtam communicated their results to the journal *Naturwissenschaften*. But by then it was too late! Much later, scientists from Russia kept referring to Raman scattering as Mandel'shtam–Raman scattering. For a very nice historical account of the Raman effect, we refer the reader to a book by G. Venkataraman, *Journey into Light: Life and Science of C.V. Raman*, published by Penguin Books (1994).

In 1958, thirty years after the discovery of the Raman effect, Raman wrote an article on the Raman effect in *Encyclopaedia Britannica*. In that article he wrote, "The rotations of the molecules in gases give more readily observable effects, viz., a set of closely spaced but nevertheless discrete Raman lines located on either side of the incident line. In liquids, only a continuous wing or band is usually observed in the same region, indicating that the rotations in a dense fluid are hindered by molecular collisions. The internal vibrations of the molecules, on the other hand, give rise in all cases to large shifts of wave length. The Raman lines attributed to them appear well separated from the parent line and are therefore easily identified and measured."

In stimulated Raman emission, the radiation emitted in the ordinary Raman effect is made to stimulate further Raman emission. This can lead to what is usually referred to as the *Raman amplification* of the beam.

Now, in fused silica, because of interaction between adjacent SiO_2 molecules, the vibrational bands are very broad; this leads to a very broad Raman shift lying between 430 and 470 cm^{-1} [this corresponds to a Raman frequency shift between 13 and 14 THz (1 THz = 10^{12} Hz)]. Thus if we have a pump laser at 1450 nm ($\bar{\nu}$ = 6897 cm^{-1}), then an incoming beam at 1550 nm ($\bar{\nu}$ = 6452 cm^{-1}) will get amplified by stimulated Raman scattering

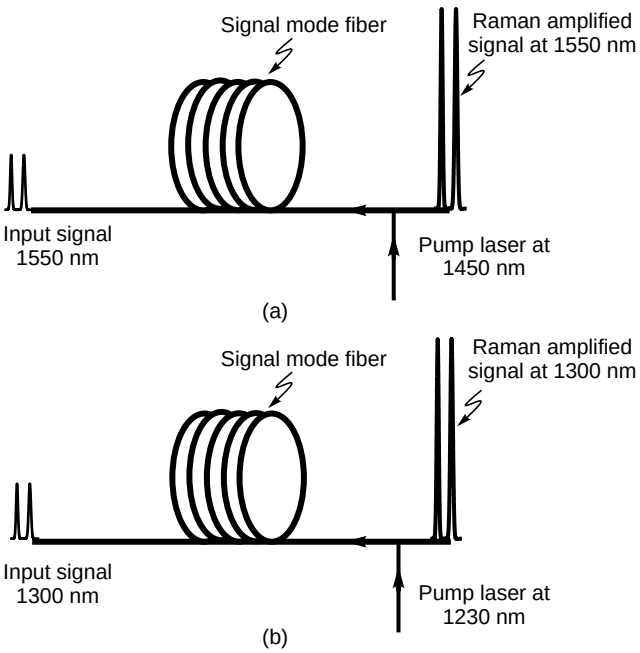


Fig. 26.40 Raman fiber amplifiers at 1550 and 1300 nm wavelengths [Figure adapted from Ref. 23].

($\Delta\bar{\nu}$ = 445 cm^{-1}) as shown in Fig. 26.40(a). In an actual commercially available single-mode fiber of length about 30 km, one can obtain a Raman gain of about 15 dB (i.e., a power amplification by a factor of about 30) by using a pump laser of 500 mW power.

Similarly, if we want to amplify an incoming beam at 1300 nm ($\bar{\nu}$ = 7692 cm^{-1}), then we must use a pump laser at about 1230 nm wavelength ($\bar{\nu}$ = 8130 cm^{-1}) as shown in Fig. 26.40(b). This is the great advantage of the Raman fiber amplifier. One can amplify signal at *any* wavelength provided we choose the pump laser frequency separated by about 13.5 THz (equivalent to a wave number shift of about 450 cm^{-1}). On the other hand, as we may recall, in erbium-doped fiber amplifiers (EDFAs), one can amplify signals only around 1550 nm wavelength; however, the laser power required is much smaller.

The above principle can be used to build the cascaded Raman laser (see Fig. 26.41). The vertical bars represent FBGs (fiber bragg gratings) which are strongly reflecting at the wavelengths written on the top (see Sec. 15.6.1 for a brief account on FBGs). Thus the input wavelength of 1100 nm (≈ 9091 cm^{-1}) produces Raman scattered line at 1155 nm (≈ 8658 cm^{-1}, implying a Raman shift of about 433 cm^{-1}); this resonates between two FBGs having peak reflectivity at 1155 nm. Now, this 1155 nm (≈ 8658 cm^{-1}) beam produces Raman scattered line at 1218 nm (≈ 8210 cm^{-1} implying a Raman shift of about 448 cm^{-1}) which resonates between two

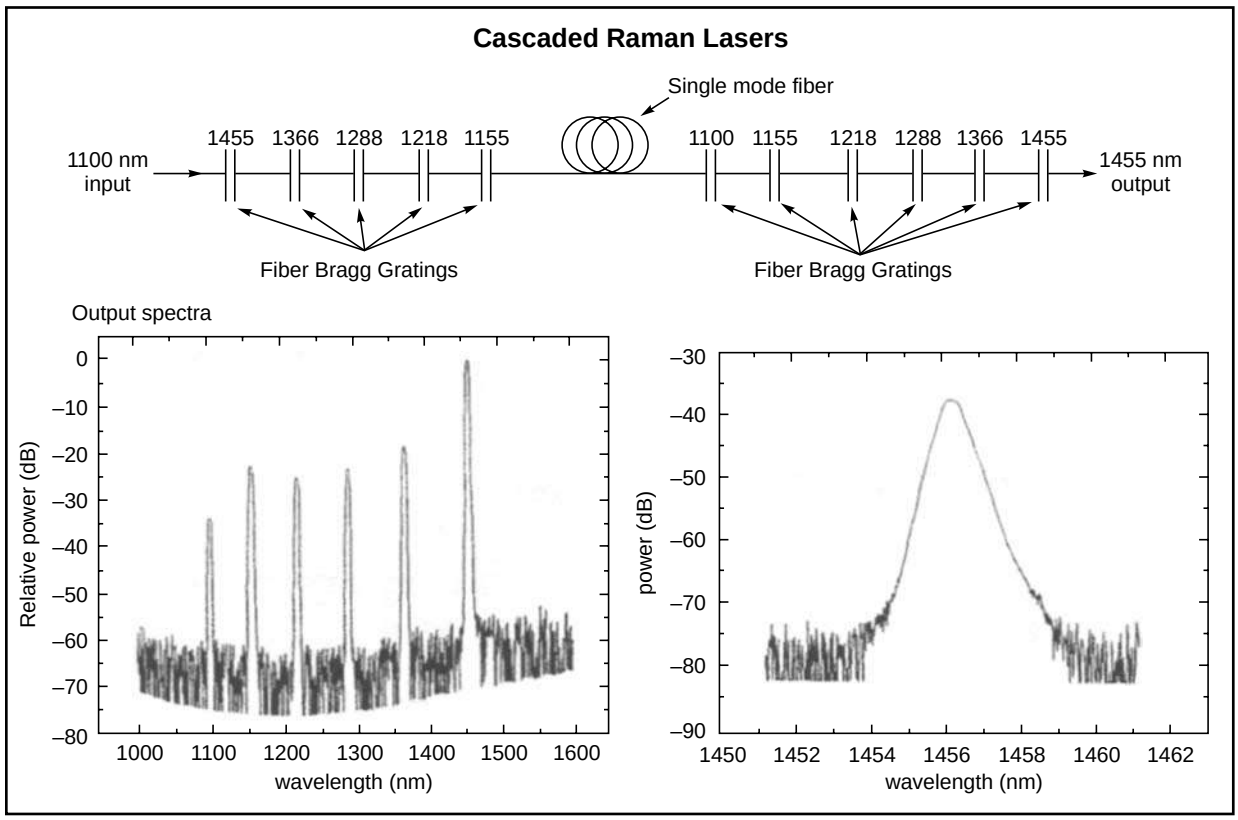


Fig. 26.41 The cascaded Raman laser; output can be generated anywhere from 1100 to 1600 nm [Adapted from the lecture notes of K. Rottwitt on "Raman Amplification Using Optical Fibers," CGCRI, Kolkata].

FBGs having peak reflectivity at 1218 nm etc. This way laser output can be generated anywhere from 1100 to 1600 nm (see Fig. 26.41).

Summary

- LASER is an acronym for *l*ight *a*mplification by *s*timulated *e*mission of *r*adiation. The light emitted from a laser often possesses some very special characteristics. Some of these are (1) *directionality*: because of which a laser beam can be focused to areas ~ few $(\mu m)^2$ leading to applications in surgery, material processing, compact discs, etc.; (2) *high power*: continuous wave lasers having power levels ~ 10^5 W and pulsed lasers having a total energy ~ 50,000 J have applications in welding, cutting, laser fusion etc.; and (3) *spectral purity*: laser beams can have an extremely small spectral width $\Delta\lambda$, because of which lasers find applications in holography, optical communications, spectroscopy etc.
- As put forward by Einstein, when an atom is in the excited state, then, in addition to the spontaneous emission, it can make a transition to a lower energy state by what is known as stimulated emission in which an incident signal of appropriate frequency triggers an atom in an excited state to emit radiation. This results in the amplification of the incident beam. If we are able to create a state of population inversion in which there are a larger number of atoms in the upper state, then the number of stimulated emissions will exceed the number of stimulated absorptions, resulting in the (optical) amplification of the beam.
- The three main components of any laser are
  (i) The active medium which consists of a collection of atoms, molecules, or ions (in solid, liquid, or gaseous form), which is capable of amplifying light waves,
  (ii) The pumping mechanism which allows us to obtain a state of population inversion between a pair of energy levels of the atomic system,
  (iii) The optical resonator, which provides the feedback.
- Through a pumping mechanism, one creates a state of population inversion in the laser placed inside the resonator system. The spontaneous emission occurring inside the resonator cavity excites the various modes of the cavity. The modes for which the gain is higher than the losses get amplified by drawing energy from the laser medium. The amplitude of the mode increases rapidly until the upper level

population reaches a value when the gain equals the losses and the mode oscillates in steady state.

- Two mirrors facing each other form a resonant cavity. The discrete frequencies of the resonator modes are given by $\nu = \nu_m = m\,c/2d$. Different values of m lead to different oscillation frequencies, which constitute the longitudinal modes of the cavity. For example, for an optical resonator of length $d \approx 60$ cm operating at an optical frequency of $\nu \approx 5 \times 10^{14}$ Hz (corresponding to $\lambda \approx 6000$ Å), we obtain $m \approx 2 \times 10^6$.
- The first successful operation of a laser device ($\lambda \sim 0.684$ μm) was demonstrated by Theodore Mainman in 1960 using a ruby crystal. Within a few months of the operation of the ruby laser, Ali Javan and his associates constructed the first gas laser ($\lambda \sim 0.633$ μm), namely, the helium-neon laser.
- If we put a fiber (doped with erbium or neodymium) between two mirrors (which act as a resonator), then with an appropriate pump we would have a fiber laser. In 1961, the first fiber laser (barium crown glass doped with Nd^{3+} ions) was fabricated by Elias Snitzer.
- The threshold population inversion required for the oscillation of the laser is given by

$$(N_2 - N_1)_{\text{th}} = \frac{\omega^2 n_0^3 t_{\text{sp}}}{\pi^2 c^3 t_c g(\omega)}$$

where t_{sp} is the spontaneous emission lifetime, t_c is the passive cavity lifetime, and $g(\omega)$ is the line shape function. For a He-Ne laser

$$g(\omega) = \frac{2}{\Delta\omega_{\text{D}}}\left(\frac{\ln 2}{\pi}\right)^{1/2} \exp\left[-4\ln 2\,\frac{(\omega-\omega)^2}{\Delta\omega_{\text{D}}}\right]$$

where $\Delta\omega_{\text{D}} = 2\omega_0 (2k_B T/Mc^2)^{1/2} \ln 2$ represents the FWHM (full width at half maximum) of the line k_B the Boltzmann constant, T represents the absolute temperature of the gas, and M represents the mass of the atom responsible for the lasing transition (neon in the case of a He-Ne laser). Notice that the minimum threshold value of $N_2 - N_1$ would correspond to the center of the line where $g(\omega)$ is a maximum and for He-Ne laser at $T = 300$ K, $\Delta\omega_{\text{D}} \approx 8230$ MHz giving $g(\omega_0) \approx 1.1 \times 10^{-10}$ s. Assuming $M = 20M_H \approx 3.3 \times 10^{-23}$ g, $t_c \approx 10^{-7}$ s $\approx t_{\text{sp}}$, $n_0 \approx 1$, we get $(N_2 - N_1)_{\text{th}} \approx 4 \times 10^8$ cm^{-3}.

Problems

26.1 Determine the mks units of $u(\omega)$, u_ω, A, and B.

[**Ans:** J s m^{-3}; m^{-3}; s^{-1}; m^{-3} J^{-1} s^{-2}]

26.2 For the $2P \rightarrow 1S$ transition in the hydrogen atom, calculate ω. Assuming the spontaneous emission lifetime of the $2P$ state to be 1.6 ns, calculate the Einstein B coefficient. Assume $n_0 \approx 1$.

[**Ans:** $\omega \approx 1.5 \times 10^{16}$ Hz; $B_{21} \approx 4.2 \times 10^{20}$ m^{-3} J^{-1} s^{-2}]

26.3 **(a)** Consider a He-Ne laser with cavity lifetime $t_c \approx 5 \times 10^{-8}$ s. If $R_1 = 1.0$ and $R_2 = 0.98$, calculate the cavity length d; assume $n_0 \approx 1$.

(b) Calculate $\Delta\nu_p$ and compare with the longitudinal mode spacing $\delta\nu$.

[**Ans:** (a) $d \approx 15$ cm; (b) $\Delta\nu_p \approx 3.2$ MHz; $\delta\nu \approx 1$ GHz]

26.4 In a typical He-Ne laser ($\lambda = 6328$ Å) we have $d \approx 20$ cm, $R_1 \approx R_2 \approx 0.98$, $\alpha_c \approx 0$, $t_{\text{sp}} \approx 10^{-7}$ s, $\Delta\nu_D \approx 1.3 \times 10^9$ Hz, and $n_0 = 1$. Calculate t_c and $(N_2 - N_1)_{\text{th}}$.

[**Ans:** 33 ns; 8.8×10^8 cm^{-3}]

26.5 Consider the D_1 line of Na ($\lambda \approx 5890$ Å).

(a) The spontaneous emission lifetime $t_{\text{sp}} \approx 16$ ns. Calculate $\Delta\nu_N$ and $\Delta\lambda_N$.

(b) Assume $T = 500$ K. Calculate $\Delta\nu_D$ and $\Delta\lambda_D$. ($k_B \approx 1.38 \times 10^{-23}$ J K^{-1}, $M_{\text{Na}} \approx 23M_H$; $M_H \approx 1.67 \times 10^{-27}$ kg).

[**Ans:** $\Delta\lambda_N \approx 10^{-4}$ Å; $\Delta\lambda_D \approx 0.02$ Å]

26.6 In a CO_2 laser ($\lambda_0 \approx 10.6$ μm), the laser transition occurs between the vibrational states of the CO_2 molecule. At $T \approx 300$ K, calculate the Doppler line width $\Delta\nu_D$ and $\Delta\lambda_D$ ($M_{\text{CO}_2} \approx 44M_H$).

[**Ans:** $\Delta\nu_D \approx 53$ MHz; $\Delta\lambda_D \approx 0.2$ Å]

26.7 Consider a light beam of all frequencies lying between $\nu = \nu_0 = 5.0 \times 10^{14}$ Hz and $\nu = 5.00002 \times 10^{14}$ Hz incident normally on a resonator (see Fig. 26.23) with $R = 0.95$, $n_0 = 1$, and $d = 25$ cm. Calculate the frequencies (in the above frequency range) and the mode number which will correspond to transmission resonances.

[**Ans:** $\nu = \nu_0 + 400$ MHz ($m = 833{,}334$), $\nu_0 + 1000$ MHz ($m = 833{,}335$) and $\nu_0 + 1600$ MHz ($m = 833{,}336$)]

26.8 Referring to Fig. 26.26, if $d = 2R_1 = 2R_2$, show that all rays passing through the common center of curvature of the mirrors will retrace their path and hence be trapped inside the cavity.

26.9 Consider a He-Ne laser ($\lambda_0 = 0.6328$ μm) with $d = 30$ cm, $n_0 \approx 1$, $R_1 \approx 1$, and $R_2 \approx 0.99$. Calculate the passive cavity line width $\Delta\nu_p$ and the passive cavity lifetime t_c. Assume $\alpha_c \approx 0$.

[**Ans:** 0.8 MHz, 0.2 μs]

26.10 **(a)** For the He-Ne laser described in Prob. 26.9, if the power level is 0.5 mW, calculate the ultimate line width $(\delta\nu)_{\text{sp}}$.

(b) Discuss the stability of the mirror position Δd to obtain the ultimate line width.

[**Ans:** (a) $(\delta\nu)_{\text{sp}} \approx 2.5 \times 10^{-3}$ Hz;
(b) $\Delta d \leq 1.6 \times 10^{-6}$ m]

26.11 The spot size of a propagating Gaussian beam is given by

$$w^2(z_1) = w_0^2 + \frac{\lambda^2 z_1^2}{\pi^2 w_0^2}$$

Substitute the expressions for w_0^2 [see Eq. (12)], and using the results derived in Sec. 26.5, show that the spot sizes at the mirrors are given by

$$w^2(z_2) = \frac{\lambda d}{\pi}\sqrt{\frac{g_1}{g_2(1-g_1g_2)}}$$

and

$$w^2(z_1) = \frac{\lambda d}{\pi}\sqrt{\frac{g_2}{g_1(1-g_1g_2)}}$$

REFERENCES AND SUGGESTED READINGS

1. C. H. Townes, "Production of Coherent Radiation by Atoms and Molecules," in *Nobel Lectures in Physics* (1963–1970), Elsevier Publishing Company, Amsterdam. Reprinted in Ref. 4.
2. N. G. Basov, "Semiconductor Lasers," in *Nobel Lectures in Physics* (1963–1970), Elsevier Publishing Company, Amsterdam. Reprinted in Ref. 4.
3. A. M. Prochorov, "Quantum Electronics," in *Nobel Lectures in Physics* (1963–1970), Elsevier Publishing Company, Amsterdam. Reprinted in Ref. 4.
4. K. Thyagarajan and A. K. Ghatak, *Lasers: Theory and Applications*, Plenum Press, New York, 1981.
5. A. Einstein, "On the Quantum Theory of Radiation," *Physikalische Zeitschrift* Vol. 18, p. 121, 1917. Reprinted in Ref. 6.
6. D. Ter Haar, *The Old Quantum Theory*, Pergamon Press, Oxford, 1967.
7. A. K. Ghatak and K. Thyagarajan, *Introduction to Fiber Optics*, Cambridge University Press, 1998.
8. E. Desurvire, *Erbium Doped Fiber Amplifiers*, Wiley, New York, 1994.
9. A. E. Siegman, *Lasers*, Oxford University Press, Oxford, 1986.
10. W. Johnstone, "Erbium Doped Fiber Amplifiers," unpublished lecture notes.
11. S. Yoshida, S. Kuwano, and K. Iwashita, "Gain Flattened EDFA with High Al-Concentration for Multistage Repeated WDM Transmission Experiments," *Electronics Letters*, Vol. 31, p. 1765, 1995.
12. E. Snitzer, "Optical Maser Action of Nd^{3+} in a Barium Crown Glass," *Phys. Rev. Letts.* Vol. 7, pp. 444–446, 1961.
13. T. H. Maiman, "Stimulated Optical Radiation in Ruby," *Nature*, Vol. 187, pp. 493–494, 1960.
14. R. Brown, *Lasers, A Survey of Their Performance and Applications*, Business Books, London, 1969.
15. A. Javan, W. R. Bennett, Jr., and D. R. Herriott, "Population Inversion and Continuous Optical Maser Oscillation in a Gas Discharge Containing a He-Ne Mixture," *Phys. Rev. Letts.* Vol. 6, pp. 106–110, 1961.
16. C. C. Davis, *Lasers and Electro-Optics*, Cambridge University Press, Cambridge, 1996.
17. J. T. Verdeyen, *Laser Electronics*, Prentice-Hall, Englewood Cliffs, N.J., 1989.
18. C. Lin, "*Optical Communications: Single Mode Optical Fiber Transmission Systems,"* in *Optoelectronic Technology and Lightwave Communication Systems*, Ed. C. Lin, Van Nostrand Reinhold, New York, 1989.
19. Ajoy Ghatak and S. Lokanathan, *Quantum Mechanics*, 5th Edition, Macmillan India, New Delhi (2005); Reprinted by Kluwer Academic Publishers, Dordrecht (2005).
20. A. K. Ghatak and K. Thyagarajan, *Optical Electronics*, Cambridge University Press, Cambridge, 1989.
21. D. R. Herriot, "Optical Properties of a Continuous He-Ne Optical Maser," *J. Opt. Soc. Am.*, Vol. 52, p. 31, 1962.
22. A. Maitland and M. H. Dunn, *Laser Physics*, North-Holland Publishing Co., Amsterdam, 1969.
23. Ajoy Ghatak and K. Thyagarajan, *Fiber Optics & Lasers: The Two Revolutions*, Macmillan India, New Delhi (2006).

Chapter Twenty-Seven

OPTICAL WAVEGUIDES I: OPTICAL FIBER BASICS USING RAY OPTICS

I have heard a ray of light laugh and sing. We may talk by light to any visible distance without any conducting wire.

—Alexander Graham Bell (1880),
after succeeding in transmitting a
voice signal over 200 m using light as the signal carrier

Important Milestones[1]

1841	*Daniel Colladon demonstrates (in Geneva) light guiding in water jets.*
1842	*Jacques Babinet demonstrates (in Paris) light guiding in water jets and also in bent glass rods.*
1854	*John Tyndall demonstrates light guiding in water jets, duplicating but not acknowledging Babinet.*
1880	*Alexander Graham Bell invents the Photophone in Washington.*
1926	*C. W. Hansell outlines the principles of fiber-optic imaging bundle.*
1930	*Heinrich Lamm, a medical student in Munich, first assembled a bundle of transparent fibers to transmit an image. van Heel in the Netherlands and Hopkins and Kapany in the United Kingdom suggest a cladding will improve transmission characteristics.*
1960	*Maiman fabricates the first laser.*
1961	*Snitzer publishes the theory of single-mode fibers and also fabricates the first fiber laser (barium crown glass doped with Nd^{3+} ions).*
1966	*Kao and Hockham predict that if it were possible to produce optical fibers with attenuation of less than 20 dB km^{-1}, it could compete effectively with the conventional communication systems.*
1970	*Kapron, Keck, and Maurer (at Corning Glass in the United States) were successful in producing silica fibers with a loss of about 17 dB km^{-1}.*
1970	*Alferov in Leningrad and Panish and Hayashi at Bell Labs demonstrate room-temperature operation of Semiconductor Lasers.*
1975	*Continuous-wave semiconductor laser operating at room temperature commercially available.*
1975	*Payne and Gambling show very small pulse dispersion at 1.27 μm.*
1976	*Bell Labs tests parabolic index fiber-optic communication system transmitting 45 Mbits s^{-1}.*
1978	*NTT (Japan) transmits 32 Mbits s^{-1} through 53 km of graded index fibers at 1.3 μm.*
1987	*Payne, Mears, and Reekie (at University of Southampton) and Desurvire, Becker, and Simpson (at AT&T Bell Laboratories) develop EDFAs (erbium-doped fiber amplifiers) operating at 1.55 μm.*
1988	*First transatlantic fiber cable, using single-mode fibers, was made operative at 1.3 μm.*
1996	*Fujitsu, NTT Labs, and Bell Labs independently report sending over 1 Tbits s^{-1} through one single-mode fiber using WDM techniques.*

[1] A nice historical account of the development of the optical fiber has been given in Ref. 1. Some of the dates given above are as given in Refs. 1 and 2.

27.1 INTRODUCTION

The dramatic reduction of transmission loss in silica optical fibers coupled with equally important developments in the area of light sources and detectors has brought about a phenomenal growth of the fiber-optic industry during the past three decades. The birth of optical fiber communication coincided with the fabrication of low-loss silica fibers and room-temperature operation of semiconductor lasers in 1970. Since then, the scientific and technological progress in this field has been phenomenal. Recent developments in optical amplifiers and wavelength division multiplexing (WDM) are taking us to a communication system with extremely small loss and an unbelievably large bandwidth. Optical fiber communication systems are fulfilling the increased demand on communication links especially with the proliferation of the Internet. Major advantages of silica optical fibers are their insensitivity to electromagnetic interference, small size and weight, low cost, and capability of carrying information at extremely high bit rates. Although the most important application of optical fibers has been in the area of telecommunications, many new related areas such as fiber-optic sensors, nonlinear fiber optics, fiber-optic devices and components, and integrated optics have witnessed considerable growth.[2] Because of all this, light wave propagation through optical fibers has recently become an extremely important subject in both teaching and research.

This chapter is an introduction to the basics of the optical fiber, discussing especially the characteristics of optical fibers as regard to their application to fiber-optic communication systems and to fiber-optic sensors. Following a historical introduction, we will use ray optics to discuss the basic principle of light guidance in an optical fiber and its two important characteristics: attenuation and pulse dispersion. We will also briefly discuss plastic optical fibers and very simple fiber-optic sensors. Fiber amplifiers and fiber lasers were very briefly discussed in Chap. 26. In single-mode fibers, it is necessary to use the concept of modes which we will discuss in the following two chapters. The ray optics treatment used in this chapter is applicable to what are known as multimode fibers.

27.2 SOME HISTORICAL REMARKS

Communication implies transfer of information from one point to another. When it is required to transmit some information such as speech, images, data, etc. over a distance, one generally uses the concept of carrier wave communication. In such a system, the information to be sent modulates an electromagnetic wave such as a radio wave or microwave which acts as a carrier. This modulated wave is then transmitted to the receiver through a channel, and the receiver receives the modulated wave and demodulates it to retrieve the signal. For example, the amplitude-modulated (AM) broadcast band usually ranges from about 600 kHz to about 2 MHz. If we assume that the highest frequency associated with music is about 20 kHz (= 0.02 MHz), then at a carrier frequency of 1.5 MHz, the spectral range of the AM wave must vary from 1.48 to 1.52 MHz—a bandwidth of 40 kHz. Thus in the entire AM broadcast range from about 600 kHz to about 2 MHz we can have at most about 30 channels; indeed we will have fewer channels if we use greater bandwidth for each channel. On the other hand, in TV transmission since we have to scan pictures, more information needs to be sent, and we require much greater bandwidth (about 5 MHz), necessitating higher carrier frequency; the carrier frequencies associated with the TV broadcast range from about 500 to about 900 MHz.

Since optical beams have frequencies in the range of 10^{14} to 10^{15} Hz, the use of such beams as the carrier would imply a tremendously large increase in the information transmission capacity of the system compared to systems employing radio waves or microwaves. It is this large information-carrying capacity of a light beam that has generated interest among communication engineers to develop a communication system using light waves as carrier waves.

Now, in a conventional telephone hookup, voice signals are converted to equivalent electric signals by the microphone and are transmitted as electric currents through metallic (copper or aluminum) wires to the local telephone exchange. Thereafter, these signals continue to travel as electric currents through metallic wire cable (or for long-distance transmission as radio/microwaves to another telephone exchange) usually with several repeaters in between. From the local-area telephone exchange at the receiving end, these signals travel to the receiver telephone via metallic wire pairs where they are converted back to corresponding sound waves. Through such cabled wire-pair telecommunication systems, one can send at most 48 simultaneous telephone conversations intelligibly. On the other hand, in an optical communication system, which utilizes glass fibers as the transmission medium and light waves as carrier waves, it has been possible (in 2001) to send over 1 Tbit of information in 1 s (which is roughly equivalent to transmission of about 15 million simultaneous telephone conversations) through one hair-thin optical fiber. This is certainly one of the extremely important technological achievements of the twentieth century.

[2] Reference 3 is a comprehensive treatise on recent developments in guided wave optical components and devices.

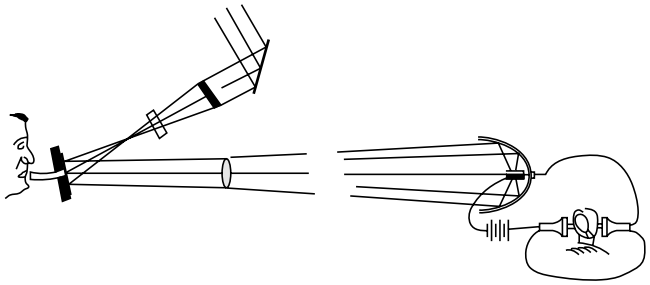


Fig. 27.1 The diagram of the Photophone; this has been taken from Alexander Graham Bell's 1880 paper "On the Production and Reproduction of Sound by Light," *American Journal of Sciences*, Third Series, Vol. XX, No. 118, pp. 305–324, October 1880. In this system, sunlight was modulated by a diaphragm and transmitted through a distance of about 200 m in the air to a receiver containing a selenium cell connected to the earphone.

The idea of using light waves for communication can be traced as far back as 1880 when Alexander Graham Bell invented the Photophone (see Fig. 27.1) shortly after he invented the telephone[3] in 1876. In this remarkable experiment, speech was transmitted by modulating a light beam, which traveled through air to the receiver. The transmitter consisted of a flexible reflecting diaphragm which could be activated by sound and which was illuminated by sunlight. The reflected light was collimated by a lens, and the reflected beam was received by a parabolic reflector placed at a distance. The parabolic reflector concentrated the light on a photoconducting selenium cell, which forms a part of a circuit with a battery and a receiving earphone. Sound waves present in the vicinity of the diaphragm vibrate the diaphragm which leads to a consequent variation of the light reflected by the diaphragm. The variation of the light falling on the selenium cell changes the electrical conductivity of the cell, which in turn changes the current in the electric circuit. This changing current reproduces the sound on the earphone. To quote from Ref. 4:

> The *Photophone* was invented jointly by Alexander Graham Bell and his assistant Charles Sumner Tainter on February 19, 1880. . . . The device allowed for the transmission of sound on a beam of light. On June 3, 1880, Bell transmitted the first wireless telephone message on his newly-invented Photophone. The Photophone used crystalline selenium cells as the receiver. This material's electrical resistance varies inversely with the illumination, i.e., its resistance is higher when it is in the dark, and lower when it is lighted. The idea of the Photophone was thus to modulate a light beam: the resulting varying illumination of the receiver would induce corresponding varying resistance in the selenium cells, which could be used by a telephone to regenerate the sounds captured at the receiver. The modulation of the light beam was done by a vibrating mirror: a thin mirror would alternate between concave and convex forms, thus focussing or dispersing the light from the light source. The Photophone functioned similarly to the telephone, except the Photophone used light as a means of projecting the information, while the telephone relied on electricity.

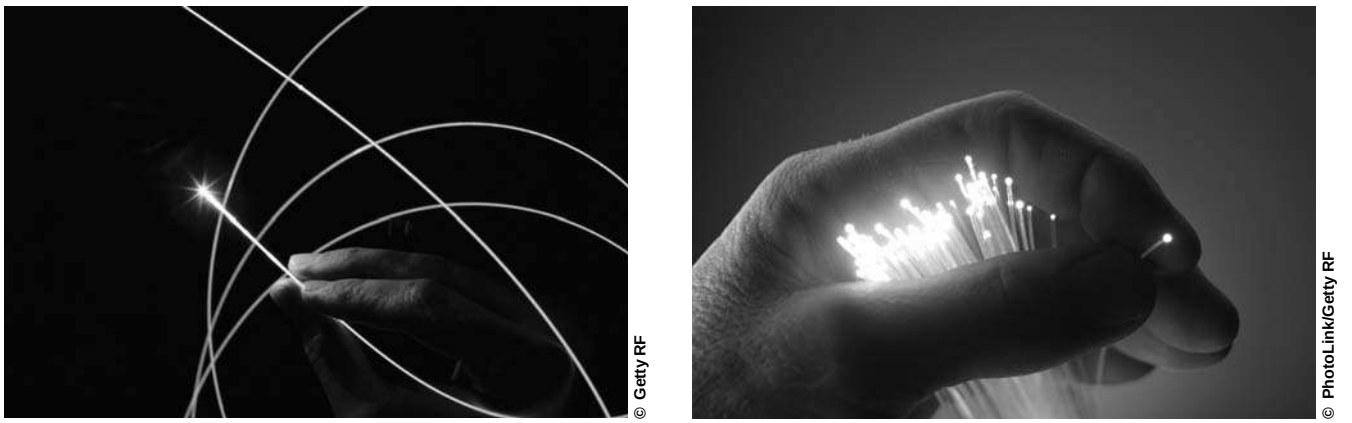
To quote from Ref. 5;

> In 1880 he (Graham Bell) produced his "Photophone" which to the end of his life, he insisted was *"... the greatest invention I have ever made, greater than the telephone."* Unlike the telephone it had no commercial value.

The modern impetus for telecommunications with carrier waves at optical frequencies owes its origin to the discovery of the laser in 1960. Earlier, there was no suitable light source available that could reliably be used as the information carrier.[4] On the other hand, around the same time telecommunications traffic was growing so rapidly that it was felt that conventional telecommunication systems based on, say, coaxial cables, radio and microwave links, and wire-pair cable could soon reach a saturation point. The advent of lasers thus immediately triggered a great deal of investigation aimed at examining the possibility of building optical analogs of conventional communication systems. The very first such modern optical communication experiments involved laser beam transmission through the atmosphere. However, it was soon realized that laser beams could not be sent in open atmosphere through reasonably long distances to carry signals, unlike, for example, microwave or radio systems operating at longer wavelengths. This is so because a light beam (of wavelength about 1 μm) is severely attenuated and distorted owing to scattering and absorption by the atmosphere. Thus

[3] Actually according to recent newspaper reports (published in June 2002), an Italian immigrant named Antonio Meucci was the inventor of the telephone. According to this report, Antonio Meucci demonstrated his "teletrfono" in New York in 1860. Alexander Graham Bell took out his patent 16 years later. This has apparently been recognized by the U.S. Congress.

[4] Although incoherent sources such as light-emitting diodes (LEDs) are often used in present-day optical communication systems, it was the discovery of the laser which triggered serious interest, for the first time, in the development of optical communication systems.



(a) (b)

Fig. 27.2 (a) Guidance of light beam through optical fibers; the light scattered out of the fiber is due to Rayleigh scattering. (b) Optical fibers held by a hand. Color photographs appear in the insert at the back of the book.

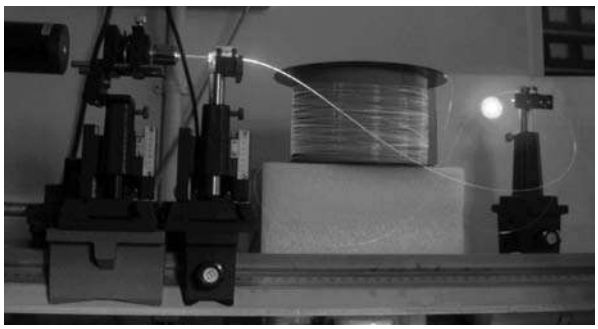


Fig. 27.3 A step index multimode fiber illuminated by He-Ne laser with bright output light spot. The light coming out of the optical fiber is primarily due to Rayleigh scattering. A color photograph appears in the insert at the back of the book. [The fiber was produced at the fiber drawing facility at CGCRI, Kolkata; Photograph courtesy Dr. Shyamal Bhadra and Ms. Atasi Pal.]

for reliable light wave communication, it would be necessary to provide a transmission medium that can protect the signal-carrying light beam from the vagaries of the terrestrial atmosphere. This guiding medium is the optical fiber (having core dimensions from a few micrometers to about 50 μm) which guides the light beam from one place to another (see Figs. 27.2 and 27.3); the guidance of the light beam through the optical fiber takes place because of the phenomenon of total internal reflection which we will discuss in the following section.

In addition to the capability of carrying a huge amount of information, optical fibers fabricated with recently developed technology are characterized by extremely low losses[5] (< 0.25 dB km^{-1}) as a consequence of which the distance between two consecutive repeaters (used for amplifying and reshaping the attenuated signals) could be as large as 250 km. It was the important paper of Kao and Hockham in 1966 (Ref. 7) that suggested that optical fibers based on silica glass could provide the necessary transmission medium if metallic and other impurities could be removed. To quote from the 1966 paper of Kao and Hockham:

> Theoretical and experimental studies indicate that a cladded glass fiber with a core diameter of about λ_0 and an overall diameter of about 1000 λ_0 represents a possible practical optical waveguide with important potential as a new form of communication medium. The refractive index of the core needs to be about 1% higher than that of cladding. However, the attenuation should be around 20 dB/km which is much higher than the lower limit of loss figure imposed by fundamental mechanisms.

Indeed this 1966 paper triggered the beginning of serious research in purifying silica and developing low-loss optical fibers. In 1970, Kapron, Keck, and Maurer (at Corning Glass in the United States) were successful in producing silica fibers with a loss of about 17 dB km^{-1} at a wavelength of 0.633 μm (Ref. 8). Since then, the technology has advanced with tremendous rapidity. By 1985 glass fibers were routinely produced with extremely low losses (< 0.25 dB km^{-1}). Figure 27.4 shows a typical optical fiber communication system. It consists of a transmitter which could be either an LED or a laser

[5] The attenuation is usually measured in decibels (dB)—we will define this in Sec. 27.8. A loss of 0.25 dB km^{-1} would imply that the power will decrease by a factor of 2 in traversing a distance of about 12 km.

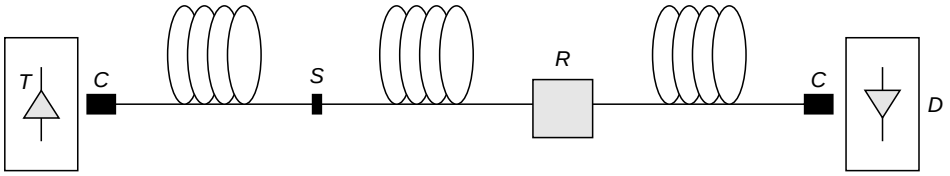


Fig. 27.4 Typical optical fiber communication system. It consists of a transmitter *T* which could be either a laser diode or an LED, the light from which is coupled into an optical fiber by means of a connector *C*. Along the path of the optical fiber, there are splices (denoted by *S*) which are permanent joints between sections of fibers and also repeaters (denoted by *R*) which boost the signal and correct any distortion that may have accumulated along the path of the fiber. At the end of the link, a coupler *C* is used to couple the light to a photodetector *D* and processed to retrieve the signal.

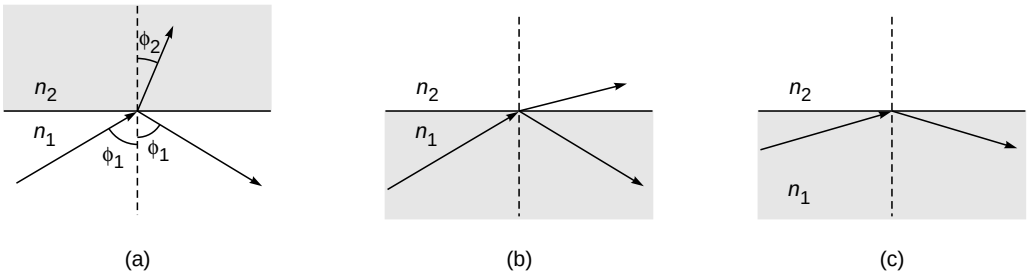


Fig. 27.5 (a) For a ray incident on a denser medium ($n_2 > n_1$), the angle of refraction is less than the angle of incidence. (b) For a ray incident on a rarer medium ($n_2 < n_1$), the angle of refraction is greater than the angle of incidence. (c) If the angle of incidence is greater than critical angle, it will undergo total internal reflection.

diode, the light from which is coupled into an optical fiber. Along the path of the optical fiber, there are splices which are permanent joints between sections of fibers and also repeaters which boost the signal and correct any distortion that may have accumulated along the path of the fiber. At the end of the link, the light is detected by a photodetector and electronically processed to retrieve the signal.

27.3 TOTAL INTERNAL REFLECTION

At the heart of an optical communication system is the optical fiber that acts as the transmission channel carrying the light beam from one place to the other; and as mentioned earlier, the guidance of the light beam (through the optical fiber) takes place because of the phenomenon of total internal reflection (often abbreviated TIR). Now, if a ray is incident at the interface of a rarer medium ($n_2 < n_1$), then the ray will bend away from the normal [see Fig. 27.5(b)]. The angle of incidence, for which the angle of refraction is 90^o, is known as the critical angle and is denoted by ϕ_c. Thus, when

$$\phi_1 = \phi_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \qquad \textbf{(1)}$$

the angle of refraction $\phi_2 = 90^o$. When the angle of incidence exceeds the critical angle (i.e., when $\phi_1 > \phi_c$), there is no refracted ray and we have what is known as total internal reflection.[6] See Fig. 27.5(b).

Example 27.1 For the glass-air interface, $n_1 = 1.5$ and $n_2 = 1.0$, and the critical angle is given by

$$\phi_c = \sin^{-1}\left(\frac{1.0}{1.5}\right) \approx 41.8°$$

On the other hand, for the glass-water interface, $n_1 = 1.5$, $n_2 = 4/3$, and

$$\phi_c = \sin^{-1}\left(\frac{4/3}{1.5}\right) \approx 62.7°$$

The phenomenon of total internal reflection can be very easily demonstrated through a simple experiment as shown in Fig. 27.6. A thick semicircular glass disc is immersed in a glass vessel filled with water. A laser beam from a He-Ne laser or simply a laser

[6] As shown in Example 24.4, energy does penetrate into the rarer medium, resulting in what is known as an evanescent wave; in any case, the reflection coefficient is unity—see also Sec. 28.2.

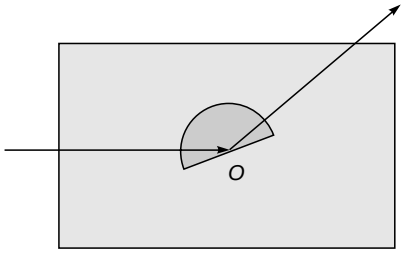


Fig. 27.6 A simple laboratory experiment to demonstrate the phenomenon of total internal reflection.

pointer is directed toward the center of the semicircular disc so that it is incident normally on the glass surface and goes undeviated as shown in the figure. The angle of incidence (at the glass-water interface) is increased by rotating the glass disc about point *O*; eventually when the angle of incidence exceeds the critical angle ($\approx 62.7^o$), the laser beam undergoes total internal reflection which can be clearly seen when viewed from the top. If one puts in a drop of ink in the water (to induce scattering of the beam), the light path becomes very beautiful to look at! The experiment is very simple, and we urge the reader to carry it out by using a laser pointer.

The first experimental demonstration of total internal reflection was carried out by sending a light beam in a water jet; this was first demonstrated by Daniel Colladon in 1841 and by Jacques Babinet in 1842. A schematic of this demonstration is shown in Fig. 27.7; light undergoes total internal reflection at

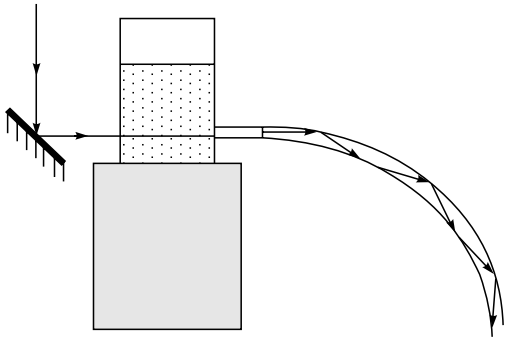


Fig. 27.7 Light guidance through a water jet demonstrating the phenomenon of total internal reflection; this was first demonstrated by Daniel Colladon in 1841.

the water-air interface and travels along the curved path of water emanating from an illuminated vessel. John Tyndall is usually credited with the first demonstration of light guidance in water jets; however, he demonstrated light guiding in water jets only in 1855, duplicating but not acknowledging Babinet; for a nice historical survey, we refer the reader to Ref. 1.

27.4 THE OPTICAL FIBER

Figure 27.8(a) shows an optical fiber, which consists of a (cylindrical) central dielectric core cladded by a material of slightly lower refractive index. The corresponding refractive index

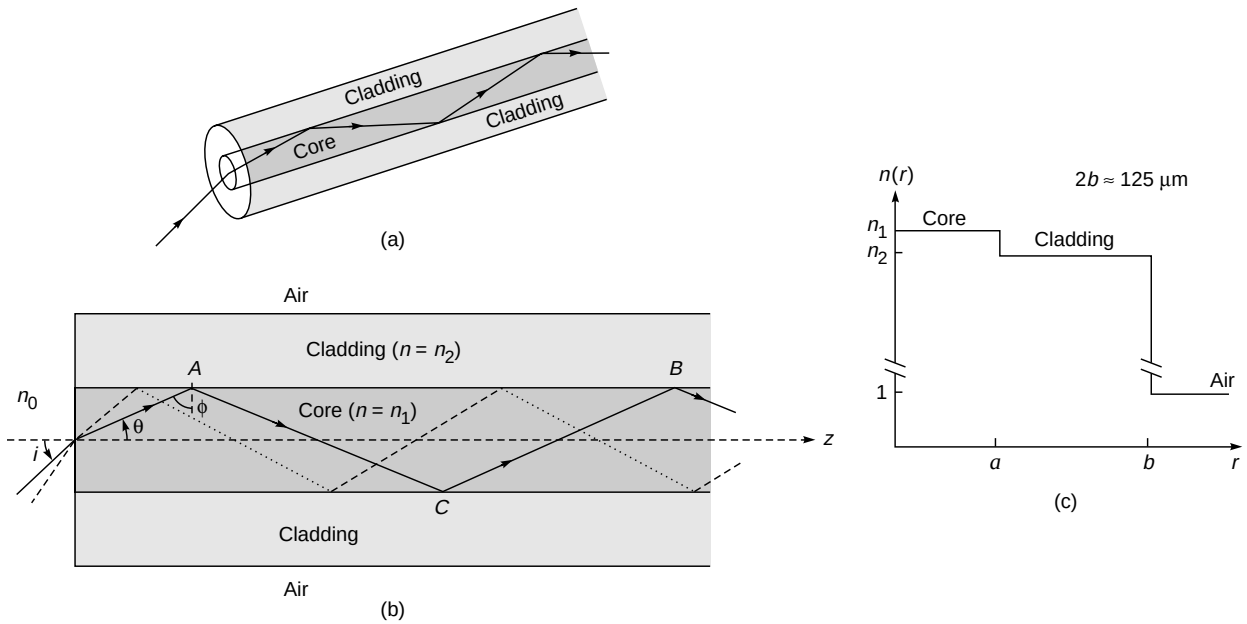


Fig. 27.8 (a) A glass fiber consists of a cylindrical central core cladded by a material of slightly lower refractive index. (b) Light rays incident on the core-cladding interface at an angle greater than the critical angle are trapped inside the core of the fiber. (c) Refractive index distribution for a step-index fiber. The diameter of the cladding is almost always 125 μm. For multimode fibers, the core diameters are usually in the range of 25 to 50 μm. For single-mode fibers, the core diameters are usually between 5 and 10 μm.

distribution (in the transverse direction) is given by

$$n = \begin{cases} n_1 & 0 < r < a \\ n_2 & r > a \end{cases} \quad \textbf{(2)}$$

where n_1 and n_2 ($< n_1$) represent, respectively, the refractive indices of core and cladding and a represents the radius of the core. We define a parameter Δ through the following equations.

$$\Delta \equiv \frac{n_1^2 - n_2^2}{2n_1^2} \quad \textbf{(3)}$$

When $n_1 \approx n_2$, i.e., when $\Delta << 1$ (as is true for most silica fibers),

$$\Delta = \frac{n_1 - n_2}{n_1} \frac{n_1 + n_2}{2n_1} \approx \frac{n_1 - n_2}{n_2} \approx \frac{n_1 - n_2}{n_1} \quad \textbf{(4)}$$

For a typical (multimoded) fiber, $a \approx 25$ μm, $n_2 \approx 1.45$ (pure silica), and $\Delta \approx 0.01$, giving a core index of $n_1 \approx 1.465$. The cladding is usually pure silica while the core is usually silica doped with germanium; doping by germanium results in an increase of refractive index.

Now, for a ray entering the fiber, if the angle of incidence (at the core-cladding interface) is greater than the critical angle ϕ_c, then the ray will undergo TIR at that interface. Thus, for TIR to occur at the core-cladding interface

$$\phi > \phi_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \quad \textbf{(5)}$$

or θ should be less than θ_c:

$$\theta < \theta_c = \cos^{-1}\left(\frac{n_2}{n_1}\right) \quad \textbf{(6)}$$

Further, because of the cylindrical symmetry in the fiber structure, the ray will suffer TIR at the lower interface also and therefore get guided through the core by repeated total internal reflections. Even for a bent fiber, light guidance can occur through multiple total internal reflections (see Figs. 27.2 and 27.7). Figures 27.2 and 27.3 show the actual guidance of a light beam as it propagates through a long optical fiber; in the photograph, the light emerging from the side of the fiber is mainly due to Rayleigh scattering, the same phenomenon that is responsible for the blue color of the sky and the red color of the rising or the setting Sun.

A cladded fiber (Fig. 27.8) rather than a bare fiber, i.e., without a cladding, was necessary because for transmission of light from one place to another, the fiber must be supported, and supporting structures may considerably distort the fiber, thereby affecting the guidance of the light wave. This can be avoided by choosing a sufficiently thick cladding. Further, in a fiber bundle, in the absence of the cladding, light can leak through from one fiber to another. The idea of adding a second layer of glass (namely, the cladding) came in 1955 from Hopkins and Kapany in the United Kingdom; however, during that time the use of optical fibers was mainly in image transmission rather than in communications. Indeed, the early pioneering works in fiber optics (in the 1950s) were by Hopkins and Kapany in the United Kingdom and by Van Heel in Holland; these works led to the use of the fiber in optical devices.

The retina of the human eye consists of a large number of rods and cones which have the same kind of structure as the optical fiber; i.e., they consist of dielectric cylindrical rods surrounded by another dielectric of slightly lower refractive index. The core diameters are in the range of a few micrometers. The light absorbed in these "light guides" generates electric signals, which are then transmitted to the brain through various nerves.

27.5 WHY GLASS FIBERS?

Why are optical fibers made of glass? To quote Prof. W. A. Gambling, who is one of the pioneers in the field of fiber optics (Ref. 2): "We note that glass is a remarkable material which has been in use in "pure" form for at least 9000 years. The compositions remained relatively unchanged for millennia and its uses have been widespread. The three most important properties of glass which makes it of unprecedented value are:

1. First, there is a wide range of accessible temperatures where its viscosity is variable and can be well controlled unlike most materials, like water and metals which remain liquid until they are cooled down to their freezing temperatures and then suddenly become solid. Glass, on the other hand, does not solidify at a discrete freezing temperature but gradually becomes stiffer and stiffer and eventually becoming hard. In the transition region it can be easily drawn into a thin fiber.
2. The second most important property is that highly pure silica is characterized with extremely low-loss; i.e., it is highly transparent. Today, in most commercially available silica fibers 96% of the power gets transmitted after propagating through 1 km of optical fiber. This indeed represents a truly remarkable achievement.
3. The third most remarkable property is the intrinsic strength of glass. Its strength is about 2,000,000 lb/in^2 so that a glass fiber of the type used in the telephone network and having a diameter (125 μm) of twice the thickness of a human hair can support a load of 40 lb."

27.6 THE COHERENT BUNDLE

If a large number of fibers are put together, it forms what is known as a *bundle*. If the fibers are not aligned, i.e., they are all jumbled up, the bundle is said to form an incoherent bundle. However, if the fibers are aligned properly, i.e., if the relative positions of the fibers in the input and output ends are the same, the bundle is said to form a coherent bundle. Now, if a particular fiber is illuminated at one of its ends, then there will be a bright spot at the other end of the same fiber; thus a coherent bundle will transmit the image from one end to another (see Fig. 27.9).

Perhaps the most important application of a coherent bundle is in a fiber-optic endoscope where it can be put inside a human body, and the interior of the body can be viewed from outside; for illuminating the portion that is to be seen, the bundle is enclosed in a sheath of fibers which carry light from outside to the interior of the body (see Fig. 27.10). A typical fiber scope can have about 10,000 fibers which would form a bundle of about 1 mm in diameter capable of resolving objects 70 μm across.

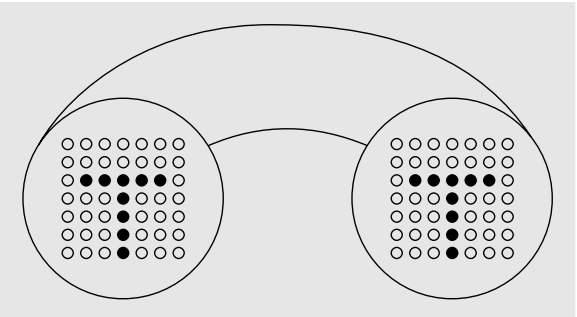
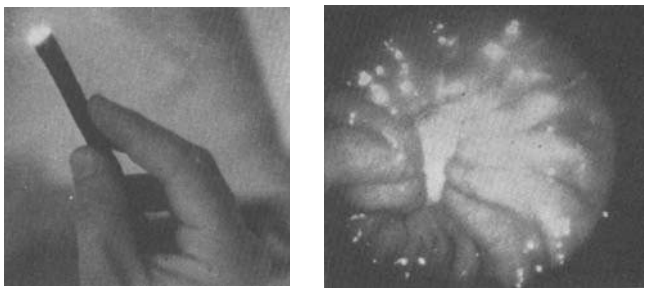


Fig. 27.9 A bundle of aligned fibers. A bright (or dark) spot at the input end of the fiber produces a bright (or dark) spot at the output end. Thus an image will be transmitted (in the form of bright and dark spots) through a bundle of aligned fibers.



(a) (b)

Fig. 27.10 (a) An optical fiber medical probe called an endoscope enables doctors to examine the inner parts of the human body, (b) A stomach ulcer as seen through an endoscope. A color photograph appears in the insert at the back of the book. [Photographs courtesy United States Information Service, New Delhi]

In an incoherent bundle the output image will be scrambled. Because of this property, an incoherent bundle can be used as a coder; the transmitted image can be decoded by using a similar bundle at the output end. In a bundle, since there can be hundreds of thousands of fibers, decoding without the original bundle configuration should be extremely difficult. Incoherent bundles are also used in illumination such as in traffic lights or road signs (see, e.g., Ref. 9). They can also be used as cold light sources (i.e., light sources giving only light and no heat) by cutting off the heat radiation, using a filter at the input to the fiber bundle. The light emerging from the bundle is also free from UV radiation and is suitable for illumination of paintings etc. in museums.

27.7 THE NUMERICAL APERTURE

We return to Fig. 27.8 and consider a ray which is incident on the entrance aperture of the fiber, making an angle i with the axis. Let the refracted ray make an angle θ with the axis. Assuming the outside medium to have a refractive index n_0 (which for most practical cases is unity), we get

$$\frac{\sin i}{\sin \theta} = \frac{n_1}{n_0} \tag{7}$$

Obviously if this ray has to suffer total internal reflection at the core-cladding interface,

$$\sin \phi \ (= \cos \theta) > \frac{n_2}{n_1} \tag{8}$$

or

$$\sin \theta < \sqrt{1 - \left(\frac{n_2}{n_1}\right)^2} \tag{9}$$

$$\sin i < \frac{n_1}{n_0}\sqrt{1 - \left(\frac{n_2}{n_1}\right)^2} = \sqrt{\frac{\left(n_1^2 - n_2^2\right)}{n_0^2}} \tag{10}$$

In most cases, the outside medium is air, i.e., $n_0 = 1$; and therefore the maximum value of sin i for a ray to be guided is given by

$$\sin i_m = \begin{cases} \sqrt{n_1^2 - n_2^2} & \text{if} \quad n_1^2 < n_2^2 + 1 \\ 1 & \text{if} \quad n_1^2 > n_2^2 + 1 \end{cases} \tag{11}$$

Thus, if a cone of light is incident on one end of the fiber, it will be guided through it provided the semiangle of the

cone is less than i_m. The quantity sin i_m is known as the *numerical aperture* (NA) of the fiber and is a measure of the *light-gathering power* of the fiber.

In almost all practical situations, $n_1^2 < n_2^2 + 1$, and therefore one defines the numerical aperture of the fiber by the following equation:

$$\text{NA} = \sqrt{n_1^2 - n_2^2} \tag{12}$$

Example 27.2 For a typical step index (multimode) fiber with $n_1 \approx 1.45$ and $\Delta \approx 0.01$, we get

$$\sin i_m \approx 0.205 \qquad \Rightarrow \qquad i_m \approx 12°$$

Now, in a short length of an optical fiber, if all rays between $i = 0$ and i_m are launched, then, the light coming out of the fiber will also appear as a cone of semiangle i_m emanating from the fiber end. If we now allow this beam to fall normally on a white paper (see Fig. 27.11) and measure its diameter, we can easily calculate the NA of the fiber. This allows us to estimate the NA of the optical fiber by a very simple experiment. The procedure is as follows:

Several concentric circles of increasing radii, say, starting from 0.5 to 1.5 cm, are drawn on a small paper screen, and the screen is positioned in the far field such that the axis of the fiber, at the output end, passes perpendicularly through the center of these circles on the screen. The fiber end, which is mounted on a *XYZ*-stack, is moved slightly towards or away from the screen so that one of the circles just circumscribes the far-field radiation spot. The distance z between the fiber end and the screen and the diameter D of the coinciding circle are measured accurately. The NA is calculated using the following equation:

$$\text{NA} = \sin i_m = \sin\left[\tan^{-1}\left(\frac{D}{2z}\right)\right] \tag{13}$$

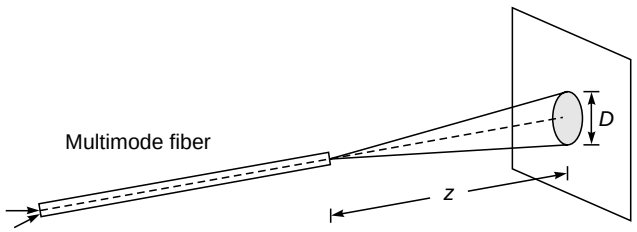


Fig. 27.11 Measurement of the diameter D of the spot on a screen placed at a far-field distance z from the output end of a multimode fiber can be used to measure the NA of the fiber.

27.8 ATTENUATION IN OPTICAL FIBERS

Attenuation and pulse dispersion represent the two most important characteristics of an optical fiber that determine the information-carrying capacity of a fiber-optic communication system. Obviously, the lower the attenuation (and similarly, the lower the dispersion), the greater the required repeater spacing and therefore the lower will be the cost of the communication system. Pulse dispersion will be discussed in the next section, while in this section, we will briefly discuss the various attenuation mechanisms in an optical fiber.

The attenuation of an optical beam is usually measured in decibels (dB). If an input power P_1 results in an output power P_2, then the loss in decibels is given by

$$\alpha = 10 \log\left(\frac{P_{\text{input}}}{P_{\text{output}}}\right) \tag{14}$$

Thus

- If the output power is the same as the input power, then the loss is = 0 dB.
- If the output power is only one-tenth of the input power, then the loss is = 10 dB.
- If the output power is only one-hundredth of the input power, then the loss is = 20 dB.
- If the output power is only one-thousandth of the input power, then the loss is = 30 dB.

Similarly, if the output power is only half of the input power, then the loss is 10 log 2 ≈ 3 dB. On the other hand, if 96% of the light is transmitted through the fiber, the loss is about 0.18 dB. In a typical fiber amplifier, a power amplification by a factor of 100 implies a power gain of 20 dB.

Figure 27.12(a) shows the variation of the loss coefficient (i.e., loss per unit length) as a function of wavelength of a typical silica optical fiber. One can notice two important low-loss windows at 1300 and 1550 nm. Typical losses at these wavelengths are 0.3 to 0.4 dB km^{-1} and about 0.25 dB km^{-1}, respectively. This is the reason why most fiber-optic systems operate in either the 1300 or 1550 nm window. The latter window has become extremely important in view of the availability of optical amplifiers (see Sec. 26.1.3).

The losses are caused by various mechanisms such as Rayleigh scattering, absorption due to metallic impurities, and water and by intrinsic absorption of silica molecule itself. Even 1 ppm (part per million) of iron can cause a loss of about 0.68 dB km^{-1} at 1100 nm. Similarly a concentration of 1 ppm of OH^- ion can cause a loss of 4 dB km^{-1} at 1380 nm. This shows the level of purity that is required to achieve very low-loss optical fibers. In Fig. 27.12(a) the two peaks are due

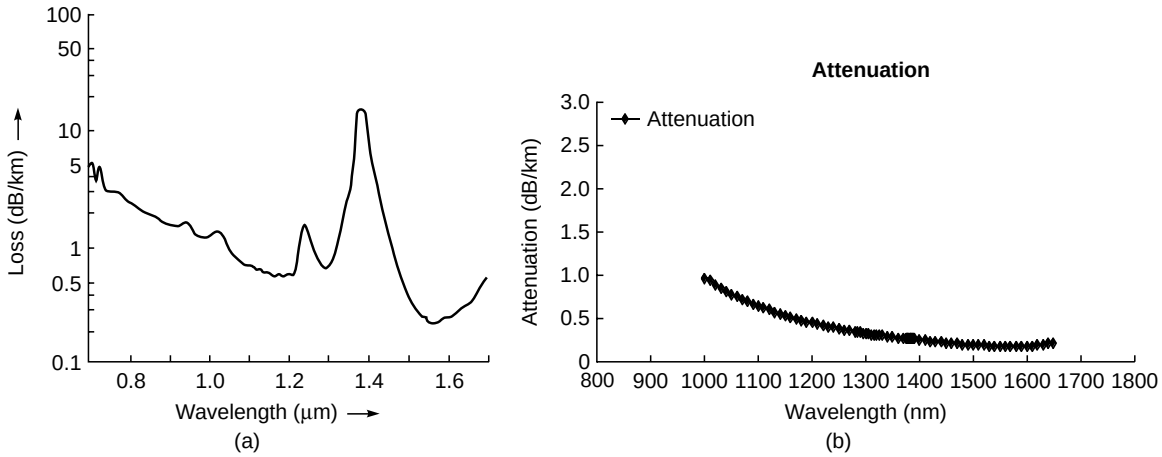


Fig. 27.12 (a) Typical wavelength dependence of loss for a silica fiber. The peaks in the attenuation curve in the wavelength regions of 1.25 and 1.40 μm are due to the presence of minute amounts of water and other impurities. Notice that the lowest loss occurs at 1550 nm (adapted from Ref. 12). (b) Using sophisticated techniques, it is possible to remove the trace amount of water and other impurities. The loss is less than 0.4 dB km^{-1} in the entire wavelength range from 1250 nm to 1650 nm. The diagram corresponds to the fiber fabricated by Sterlite Industries at Aurangabad and is courtesy S. Bhatia of Sterlite Industries.

to traces of water (and other impurities) present in the fiber. However, with sophisticated fabrication techniques it is possible to remove these impurities, and one can obtain very low loss in the entire wavelength region from 1200 to 1650 nm; see Fig. 27.12(b). For $\lambda_0 \gtrsim 1650$ nm, the loss increases because of the occurence of the infrared absorption band in silica.

It is possible to demonstrate the wavelength dependence of Rayleigh scattering by using a long optical fiber. White light (from a lamp such as a tungsten halogen lamp) is coupled into approximately a 1 km long multimode optical fiber, and we look into the output and notice the color of the light. Next, we cut the fiber, leaving about 1 m from the input end of the fiber, and repeat the experiment with this 1 m of the fiber. In the former case, the emerging light looks reddish while in the latter case it looks white. This difference is due to the decrease of loss with increase in wavelength due to Rayleigh scattering; light wavelengths toward the blue region have suffered greater scattering out of the fiber than those of the red region. Thus although at the input end all wavelengths are coupled, there is more power in the red part at the output giving it a reddish color.

Example 27.3 Calculation of losses using the decibel scale becomes very easy. For example, if we have a 40 km fiber link (with a loss of 0.4 dB km^{-1}), having three connectors in its path and if each connector has a loss of 1.8 dB, then the total loss will be 0.4 dB $km^{-1} \times 40$ km + 3×1.8 dB = 21.4 dB.

Example 27.4 Let us assume that the power of a 5 mW laser beam decreases to 30 μW after traversing through 40 km of an optical fiber. The attenuation of the fiber is therefore

$$\frac{1}{40}\left[10 \log\left(\frac{5\ \text{mW}}{0.03\ \text{mW}}\right)\right] = 0.56\ \text{dB km}^{-1}$$

It is often very convenient to measure the power level of a beam in dBm, which is defined as

$$P\ \ (\text{dBm}) = 10 \log P\ \ (\text{mW}) \qquad \textbf{(15)}$$

Thus

$$1\ \text{mW} \Leftrightarrow 0\ \text{dBm}$$
$$1\ \text{W} \Leftrightarrow 30\ \text{dBm}$$
$$1\ \mu\text{W} \Leftrightarrow -30\ \text{dBm}$$
$$1\ \text{nW} \Leftrightarrow -60\ \text{dBm}$$

Similarly,

$$0.2\ \text{W} = 200\ \text{mW} \Leftrightarrow \approx 2.3\ \text{dBm}$$

Using the dBm scale, Eq. (8) becomes

$$\alpha = P_{\text{input}}(\text{dBm}) - P_{\text{output}}(\text{dBm}) \qquad \textbf{(16)}$$

or,

$$P_{\text{output}}(\text{dBm}) = P_{\text{input}}(\text{dBm}) - \alpha(\text{dB}) \qquad \textbf{(17)}$$

Because of the above equation, calculation of power level losses using the dBm scale become very easy as shown in the examples below.

Example 27.5 Consider a 5 mW laser beam passing through a 40 km fiber link of loss 0.5 dB km^{-1}. The total loss is 20 dB. Since the input power is 6.99 dBm, the power at the output is –13.01 dBm which is equal to 0.05 mW.

Between a source and a detector, let N_s represent the number of splices and in each splice, the loss (in dB) is l_s; a splice represents the point where one fiber is joined to the other.

Similarly, let N_c represent the number of connectors, and in each connector the loss (in dB) is l_c. Thus the power received (in dBm) at the detector is given by

$$P_{\text{received}} = P_{\text{input}} - N_c l_c - N_s l_s - L\alpha$$

where α = fiber loss (dB km^{-1}) and L presents the fiber length (km).

Example 27.6 Let P_{input} = 1 mW $\Leftrightarrow$ 0 dBm; l_c = 1 dB/connector, N_c = 2; l_s = 0.5 dB/splice, N_s = 4; α = 0.5 dB km^{-1}, L = 40 km. Thus the loss in the fiber is 20 dB and

$$P_{\text{received}} = 0 - 2 - 2 - 20 = -24 \text{ dBm} \Leftrightarrow \approx 4\ \mu\text{W}$$

Example 27.7 In a typical optical communication system, let the available components be as given below:

Laser output	1.5 mW (1.76 dBm)
Laser wavelength	1300 nm
Fiber loss	1 dB km^{-1}
Required length of link	20 km
Loss in fiber 20 × 1 dB km^{-1}	20 dB
Splice (every 5 km) loss	0.5 dB/splice
Splices 3 × 0.5 dB	1.5 dB
Laser-to-fiber coupling loss	8 dB
Fiber-to-detector loss	2 dB
Total loss	31.5 dB

Since the laser power is 1.76 dBm, the power available at the detector is –29.74 dBm (≈ 1.06 μW) and if the detector margin is –40 dBm [i.e., the detector is able to detect –40 dBm of power (= 0.1 μW)], then there is an excess power margin of 10.26 dBm at the detector. The above represents a typical power budget calculation.

27.8.1 The Attenuation Limit

Let N_p represent the minimum number of photons (per bit of information) required for the pulse to be detected. The corresponding average optical power received by the detector is given by

$$P_{\text{min}} = \tfrac{1}{2} N_p BE \qquad \textbf{(18)}$$

where $E = h\nu$ = energy of each photon and B represents the bit rate (the number of bits per second) in the communication system. Typically $N_p \approx 1000$ and $B \approx 2.5$ Gbits s^{-1}.

Example 27.8 For $\lambda_0 \approx$ 1300 nm,

$$E = h\nu = \frac{h_c}{\lambda_0} \approx 1.53 \times 10^{-19} \text{ J}$$

where h is Planck's constant and we have assumed $h \approx 6.626 \times 10^{-34}$ J. S. Thus

$$P_{\text{min}} = \tfrac{1}{2} N_p BE \approx \tfrac{1}{2} \times 1000 \times (2.5 \times 10^9) \times (1.53 \times 10^{-19}) \text{ W} \approx 0.19\ \mu\text{W}\ (\approx -37.2 \text{ dBm})$$

Thus if $P_{\text{in}} \approx$ 1 mW (= 0 dBm), then the system can have a maximum loss of about 38 dB. If we neglect the splice and connector losses, then for a fiber loss of α = 0.5 dB km^{-1}, $L_{\text{max}} \approx$ 93 km.

Example 27.9 For $\lambda_0 \approx$ 1550 nm,

$$E = h\nu = \frac{h_c}{\lambda_0} \approx 1.28 \times 10^{-19} \text{ J}$$

$$\Rightarrow \quad P_{\text{min}} = \tfrac{1}{2} N_p BE \approx \tfrac{1}{2} \times 1000 \times (2.5 \times 10^9) \times (1.28 \times 10^{-19}) \text{ W} \approx 0.16\ \mu\text{W}\ (\approx -38 \text{ dBm})$$

Thus if $P_{\text{in}} \approx$ 1 mW (= 0 dBm), then the system can have a maximum loss of about 37 dB. If we neglect the splice and connector losses, then for a fiber loss of α = 0.2 dB km^{-1}, $L_{\text{max}} \approx$ 190 km.

27.9 MULTIMODE FIBERS

In the next section we will discuss broadening of an optical pulse as it passes through a multimode optical fiber. The obvious question is, What do we understand by a multimode optical fiber? The concept of modes will be discussed in Chaps. 28 and 29; it will suffice here to say that if we solve Maxwell's equations for an optical waveguide, then we obtain discrete modes that represent *transverse field distributions that suffer only a phase change as they propagate through the waveguide along z.* Each mode has a specific transverse field distribution and also a specific velocity (see Secs. 28.3 to 28.5). Now, while studying the propagation of rays in an optical fiber [see Fig. 27.8(b)], we assumed that all rays characterized by $\theta > \theta_c$ will be guided through the optical fiber. In Sec. 28.3, we will show by solving Maxwell's equations that each mode of the waveguide may be assumed to correspond to a "discrete" value of θ, which would imply discrete ray paths; thus, qualitatively speaking, we may say that only discrete values of θ are possible. When the number of such discrete ray paths becomes very large, we have what is known as a multimode fiber and may assume the validity of geometrical optics.

27.9.1 Power Law Profile

A broad class of *multimode* graded index fibers can be described by the following refractive index distribution

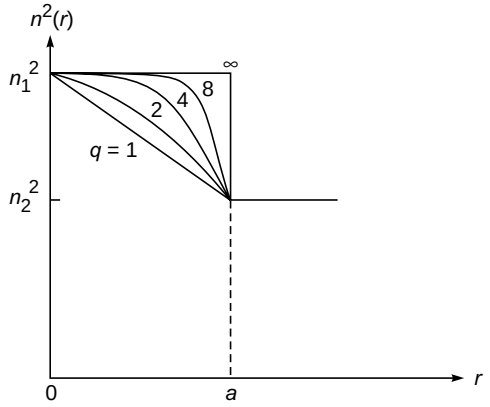


Fig. 27.13 Power law profiles for the refractive index distribution given by Eq. (19).

(see Fig. 27.13):

$$n^2(r) = \begin{cases} n_1^2\left[1-2\Delta\left(\dfrac{r}{a}\right)^q\right] & 0 < r < a \\ n_2^2 = n_1^2(1-2\Delta) & r > a \end{cases} \tag{19}$$

where r corresponds to a cylindrical radial coordinate, n_1 represents the value of the refractive index on the axis (i.e., at $r = 0$), n_2 represents the refractive index of the cladding, and a represents the radius of the core. Equation (19) describes what is usually referred to as a *power law profile* or a *q profile*; $q = 1$, $q = 2$, and $q = \infty$ correspond to the linear, parabolic, and step index profiles, respectively (see Fig. 27.13). One defines the normalized waveguide parameter as

$$V = \frac{2\pi}{\lambda_0} a\sqrt{n_1^2 - n_2^2} \tag{20}$$

where λ_0 is the free space wavelength of operation. The total number of modes in a highly multimode graded index optical fiber characterized by Eq. (19) is approximately given by [Ref. 13 (see also Ref. 14)]

$$N \approx \frac{q}{2(2+q)} V^2 \tag{21}$$

Thus, a parabolic index fiber ($q = 2$) with $V = 10$ will support approximately 25 modes. Similarly, a step index fiber ($q = \infty$) with $V = 10$ will support approximately 50 modes. When the fiber supports such a large number of modes, the fiber is said to be a multimode fiber. Each mode travels with a slightly different group velocity, leading to what is known as *intermodal dispersion*. In Ref. 13 (see also Ref. 14) it has been shown that for a highly multimode graded index optical fiber, the value of intermodal dispersion is very nearly the same as obtained from ray analysis. Thus in highly multimode fibers ($V \geq 10$), one is justified to use the ray optics result for *intermodal* (*or ray*) *dispersion*. For a given fiber (i.e., for given values of n_1, n_2, and a), the value of V depends on the operating wavelength λ_0. Thus, as the wavelength becomes smaller, the value of V (and hence the number of modes) increases; and in the limit of the operating wavelength becoming very small, we have the geometric optics limit. Also, as will be shown in Chap. 29, a step index fiber ($q = \infty$) has only one mode when $V < 2.4048$ and we have what is known as a single-mode fiber. For a given step index fiber, the wavelength at which V becomes equal to 2.4048 is known as the *cutoff wavelength*, and for all wavelengths greater than the *cutoff wavelength* the fiber is said to be single-mode (see Sec. 29.3.1). In all that follows we will assume the V number to be large (≥ 10), so that we may use ray optics to calculate pulse dispersion. Analysis of single-mode fibers will require solution of the wave equation which we will do in Chap. 29.

27.10 PULSE DISPERSION IN MULTIMODE OPTICAL FIBERS

In digital communication systems, first information to be sent is coded in the form of pulses, and then these pulses of light are sent from the transmitter to the receiver where the information is decoded. The larger the number of pulses that can be sent per unit time and still be resolvable at the receiver end, the larger the transmission capacity of the system. A pulse of light sent into a fiber broadens in time as it propagates through the fiber; this phenomenon is known as pulse dispersion and occurs primarily because of the following mechanisms:

1. In multimode fibers, different rays take different times to propagate through a given length of the fiber; we will discuss this for a step index fiber and for a parabolic index fiber in this and the following sections. In the language of wave optics, this is known as *intermodal dispersion* because it arises due to different modes traveling with different group velocities.
2. Any given light source emits over a range of wavelengths, and because of the intrinsic property of the material of the fiber, different wavelengths take different amounts of time to propagate along the same path. This is known as *material dispersion*, and obviously, it is present in both single-mode and multimode fibers.

3. In single-mode fibers since there is only mode, there is no intermodal dispersion; however, we have what is known as *waveguide dispersion* which is due to the geometry of the fiber. We will discuss single-mode fibers and waveguide dispersion in Chap. 29. Obviously, waveguide dispersion is present in multimode fibers also, but the effect is very small and can be neglected.

27.10.1 Ray Dispersion in Multimode Step Index Fibers

We first consider ray paths in a SIF (Step Index Fiber) as shown in Fig. 27.8. As can be seen, rays making larger angles with the axis (those shown as dotted rays) have to traverse a longer optical path length and therefore take a longer time to reach the output end.

We will now derive an expression for the intermodal dispersion for a step index fiber. Referring to Fig. 27.8, for a ray making an angle θ with the axis, the distance AB is traversed in time

$$t_{AB} = \frac{AC + CB}{c/n_1} = \frac{AB/\cos\theta}{c/n_1} \tag{22}$$

or

$$t_{AB} = \frac{n_1 AB}{c\cos\theta} \tag{23}$$

where c/n_1 represents the speed of light in a medium of refractive index n_1, with c being the speed of light in free space. Since the ray path will repeat itself, the time taken by a ray to traverse a length L of the fiber is

$$t_L = \frac{n_1 L}{c\cos\theta} \tag{24}$$

The above expression shows that the time taken by a ray is a function of the angle θ made by the ray with the z axis, which leads to pulse dispersion. If we assume that all rays lying between $\theta = 0$ and $\theta = \theta_c = \cos^{-1}(n_2/n_1)$ [see Eq. (6)] are present, then the time taken by these extreme rays for a fiber of length L is given by

$$t_{\min} = \frac{n_1 L}{c} \quad \text{corresponding to } \theta = 0 \tag{25}$$

$$t_{\max} = \frac{n_1^2 L}{c n_2} \quad \text{corresponding to } \theta = \theta_c = \cos^{-1}\left(\frac{n_2}{n_1}\right) \tag{26}$$

Hence if all the input rays were excited simultaneously, the rays would occupy a time interval at the output end of duration

$$\Delta\tau_i = t_{\max} - t_{\min} = \frac{n_1 L}{c}\left(\frac{n_1}{n_2} - 1\right) \tag{27}$$

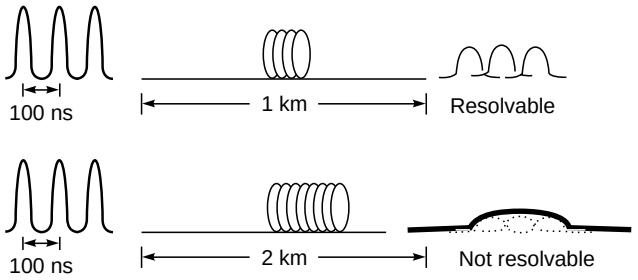


Fig. 27.14 Pulses separated by 100 ns at the input end would be resolvable at the output end of 1 km of the fiber. The same pulses would not be resolvable at the output end of 2 km of the same fiber [Figure adapted from Ref. 11].

or

$$\boxed{\Delta\tau_i \cong \frac{n_1 L}{c}\Delta \approx \frac{L}{2n_1 c}(\text{NA})^2} \quad \text{intermodal dispersion in multimode SIF} \tag{28}$$

where Δ has been defined earlier [see Eqs. (3) and (4)] and we used Eq. (12). Assuming the validity of ray optics, Eq. (27) is exact; however, in writing Eq. (28) we have assumed $\Delta << 1$, which is true for almost all commercially available fibers. The quantity $\Delta\tau_i$ represents the pulse dispersion due to different rays taking different times in propagating through the fiber which, in wave optics, is nothing but the intermodal dispersion and hence the subscript *i*. Note that the pulse dispersion is proportional to the square of NA. Thus to have a smaller dispersion, one must have a smaller NA, which of course reduces the acceptance angle and hence the light-gathering power. Now if at the input end of the fiber we have a pulse of width τ_1, then after propagating through a length L of the fiber the pulse will have a width τ_2 given approximately by

$$\tau_2^{\,2} = \tau_1^{\,2} + \Delta\tau_i^{\,2} \tag{29}$$

Consequently, the pulse broadens as it propagates through the fiber (see Fig. 27.14). Hence, even though two pulses may be well resolved at the input end, because of the broadening of the pulses they may not be so at the output end.

Example 27.10 For a typical (multimode) step index fiber, if we assume $n_1 = 1.5$, $\Delta = 0.01$, and $L = 1$ km, we get

$$\Delta\tau_i = \frac{1.5 \times 1000}{3 \times 10^8} \times 0.01 = 50 \text{ ns km}^{-1} \tag{30}$$

i.e., a pulse after traversing through the fiber of length 1 km will be broadened by 50 ns. Thus two pulses separated by, say, 500 ns at the input end would be quite resolvable at the end of 1 km of the fiber. However, if consecutive pulses are separated

by, say, 10 ns at the input end, they will be absolutely unresolvable at the output end. Hence in a 1 Mbit s^{-1} fiber-optic system, where we have one pulse every 10^{-6} s, a 50 ns km^{-1} dispersion will require repeaters to be placed every 3 to 4 km. On the other hand, in a 1 Gbit s^{-1} fiber-optic communication system, which requires the transmission of one pulse every 10^{-9} s, a dispersion of 50 ns km^{-1} will result in intolerable broadening even within 50 m or so which would be highly inefficient and uneconomical from a system point of view.

Where the output pulses are not resolvable, no information can be retrieved. Thus, the smaller the pulse dispersion, the greater the information-carrying capacity of the system.

From the discussion in Example 27.10 it follows that for a very high information-carrying system, it is necessary to reduce the pulse dispersion. Two alternative solutions exist: one involves the use of near parabolic index fibers, and the other involves single-mode fibers.

27.10.2 Parabolic index Fibers (PIFs)

In a step index fiber such as that pictured in Fig. 27.8, the refractive index of the core has a constant value. By contrast, in a PIF (Parabolic Index Fiber), the refractive index in the core decreases continuously (in a quadratic fashion) from a maximum value at the center of the core to a constant value at the core-cladding interface. The refractive index variation given by

$$n^2(r) = \begin{cases} n_1^2\left[1-2\Delta\left(\dfrac{r}{a}\right)^2\right] & 0<r<a \quad \text{core} \\ n_2^2 = n_1^2(1-2\Delta) & r>a \quad \text{cladding} \end{cases} \tag{31}$$

with Δ as defined in Eq. (4). In Sec. 3.4.1 we showed that the ray paths in a parabolic waveguide are sinusoidal (see Fig. 27.15). For a typical (multimode) parabolic index silica fiber $\Delta \approx 0.01$, $n_2 \approx 1.45$, and $a \approx 25$ μm.

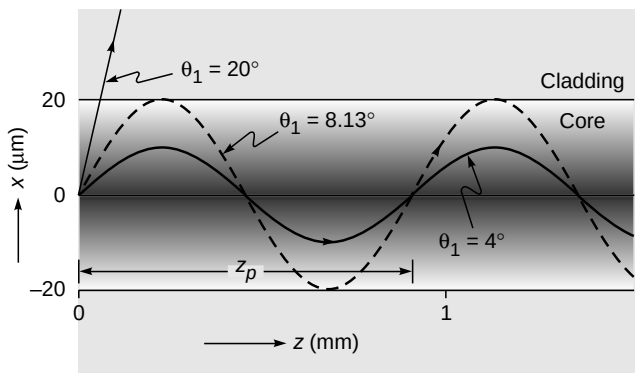


Fig. 27.15 Ray paths in a parabolic index fiber.

Now, even though rays making larger angles with the axis traverse a larger path length, they do so in a region of lower refractive index (and hence greater speed). The longer path length is almost compensated for by a greater average speed such that all rays take approximately the same amount of time in traversing the fiber. In Sec. 3.4.2 we made a detailed calculation of the time taken by a particular ray to propagate through a parabolic index waveguide; the final result for the intermodal dispersion is given by (see also Sec. 27.12)

$$\Delta\tau_i = \frac{n_2 L}{2c}\left(\frac{n_1-n_2}{n_2}\right)^2 \quad \text{pulse dispersion in multimode PIF} \tag{32}$$

When $\Delta << 1$, the above equation can be written as

$$\Delta\tau_i \approx \frac{n_2 L}{2c}\Delta^2 \approx \frac{L}{8cn_1^3}(\text{NA})^4 \tag{33}$$

Note that compared to a step index fiber, the pulse dispersion is proportional to the square of Δ. For a typical (multimode parabolic index) fiber with $n_2 \approx 1.45$ and $\Delta \approx 0.01$, we get

$$\Delta\tau_i \approx 0.25 \text{ ns km}^{-1} \tag{34}$$

Comparing it with Eq. (30), we find that for a parabolic index fiber, the pulse dispersion is reduced by a factor of about 200 in comparison to the step index fiber. For this reason first- and second-generation optical communication systems used near-parabolic index fibers. To further decrease the pulse dispersion, it is necessary to use single-mode fibers because there will be no intermodal dispersion. In almost all long-distance fiber-optic communication systems, one uses single-mode fibers; nevertheless, in many local-area communication systems (such as intra-office networks), one still uses parabolic index multimode fibers. In Sec. 27.12 we will give the general expression for pulse dispersion corresponding to the power law profile.

Now, in addition to the intermodal dispersion discussed above, in all fiber-optic systems we will have material dispersion which is a characteristic of the material itself and not of the waveguide.

27.10.3 Material Dispersion

Previously we considered the broadening of an optical pulse due to different rays taking different amounts of time to propagate through a certain length of the fiber. However, every source of light has a certain wavelength spread which is often referred to as the *spectral width of the source*. Thus a white light source (such as the Sun) has a spectral width of about 300 nm; on the other hand, an LED would have a spectral width of about 25 nm and a typical laser diode (LD) operating at 1300 nm has a spectral width of about 2 nm or

less. In Chap. 10 we said that the refractive index of the medium (and hence the group velocity v_g) depends on the wavelength. Thus, each wavelength component (of the pulse) will travel with a slightly different group velocity through the fiber, resulting in a broadening of a pulse. In Chap. 10 we showed that the pulse broadening (due to wavelength dependence of the refractive index) is given by

$$\Delta\tau_m = -\frac{L\,\Delta\lambda_0}{\lambda_0 c}\left(\lambda_0^2\,\frac{d^2n}{d\lambda_0^2}\right) \qquad \textbf{(35)}$$

where L is the length of the fiber, $\Delta\lambda_0$ is the spectral width of the source, and c is the speed of light in free space; the subscript m in Eq. (35) refers to the fact that we are considering material dispersion. We also defined the material dispersion coefficient (which is measured in ps km^{-1} nm^{-1}) as

$$D_m = \frac{\Delta\tau_m}{L\Delta\lambda_0} = -\frac{10^4}{3\lambda_0}\left(\lambda_0^2\,\frac{d^2n}{d\lambda_0^2}\right) \text{ ps km}^{-1}\text{ nm}^{-1} \qquad \textbf{(36)}$$

where we used $c \approx 3 \times 10^{-7}$ km ps^{-1} and λ_0 is measured in μm and the quantity inside the parentheses is dimensionless. Thus D_m represents the material dispersion in picoseconds per kilometer length of the fiber per nanometer spectral width of the source. At a particular wavelength, the value of D_m is a characteristic of the material and is (almost) the same for *all* silica fibers. The values of D_m for different wavelengths (for pure silica) are tabulated in Table 10.1. When D_m is negative, it implies that the longer wavelengths travel faster; similarly, a positive value of D_m implies that shorter wavelengths travel faster.

Example 27.11 The LEDs used in the earlier optical communication systems had a spectral width $\Delta\lambda_0$ of about 20 nm around $\lambda_0 = 0.85$ μm (= 850 nm); at this wavelength (see Table 10.1).

$$\frac{d^2n}{d\lambda_0^2} \approx 0.0297\ (\mu\text{m})^{-2} \quad \Rightarrow \quad \left[\lambda_0^2\,\frac{d^2n}{d\lambda_0^2}\right]$$

$$\approx 0.85 \times 0.85 \times 0.0297 \approx 0.02146$$

Thus

$$D_m \approx -\frac{10^4}{3\lambda_0}\left(\lambda_0^2\,\frac{d^2n}{d\lambda_0^2}\right)$$

$$\approx -\frac{10^4}{3\times 0.85} \times 0.02146 \approx -84.2 \text{ ps km}^{-1}\text{ nm}^{-1}$$

Thus a pulse will broaden by (disregarding the sign)

$\Delta\tau_m = D_m \times L \times \Delta\lambda$

= 84.2 ps km^{-1} nm^{-1} × 1 km × 20 nm ~ 1700 ps = 1.7 ns

in traversing 1 km length of the fiber. On the other hand, if we carry out a similar calculation around $\lambda_0 \approx 1.3$ μm (where $D_m \approx 2.4$ ps km^{-1} nm^{-1}), we will obtain a much smaller value of $\Delta\tau_m$; thus

$$\Delta\tau_m = D_m \times L \times \Delta\lambda = 2.4 \text{ ps km}^{-1}\text{ nm}^{-1} \times 1 \text{ km} \times 20 \text{ nm} \sim 0.05 \text{ ns}$$

in traversing 1 km length of the fiber. The very small value of $\Delta\tau_m$ is due to the fact that n_g is approximately constant around $\lambda_0 = 1300$ nm, as shown in Fig. 10.6. Indeed the wavelength $\lambda_0 \approx 1270$ nm is usually referred to as the zero material dispersion wavelength, and it is because of such low material dispersion that the optical communication systems shifted their operation to around $\lambda_0 \approx 1300$ nm.

Example 27.12 In the optical communication systems in operation today, one uses LDs (laser diodes) with $\lambda_0 \approx 1550$ nm having a spectral width of about 2 nm. At this wavelength, $D_m \approx 21.5$ ps km^{-1} nm^{-1} (see Table 10.1). Thus for a 1 km length of the fiber, the material dispersion $\Delta\tau_m$ becomes

$$\Delta\tau_m = D_m \times L \times \Delta\lambda = 21.5 \text{ ps km}^{-1}\text{ nm}^{-1} \times 1 \text{ km} \times 2 \text{ nm} \sim 43 \text{ ps}$$

the positive sign indicating that higher wavelengths travel slower than lower wavelengths. (Notice from Table 10.1 that for $\lambda_0 \geq 1300$ nm, n_g increases with λ_0).

27.11 DISPERSION AND MAXIMUM BIT RATES

In a digital communication system employing light pulses, pulse broadening results in an overlap of pulses, resulting in loss of resolution and leading to errors in detection. Thus pulse broadening is one of the mechanisms (other than attenuation) that limits the distance between two repeaters in a fiber-optic link. It is obvious that the larger the pulse broadening, smaller the number of pulses per second that can be sent down a link. Different criteria based on slightly different considerations are used to estimate the maximum permissible bit rate $B_{\max}$ for a given pulse dispersion. However, it is always of the order of $1/\tau$. In one type of extensively used coding [known as NRZ (Non Return to Zero)] we have

$$B_{\max} \approx \frac{0.7}{\Delta\tau} \qquad \textbf{(37)}$$

The above formula takes into account (approximately) only the limitation imposed by the pulse dispersion in the fiber. In an actual link, the source and detector characteristics are also taken into account while estimating the maximum bit rate. Note that in a fiber, the pulse dispersion is caused, in general, by intermodal dispersion, material dispersion, and waveguide dispersion. However, waveguide dispersion is important only in single-mode fibers and may be neglected in carrying out analysis for multimode fibers. Thus (considering multimode fibers), if $\Delta\tau_i$ and $\Delta\tau_m$ are the dispersion

due to intermodal and material dispersions respectively, then the total dispersion is given by

$$\Delta\tau = \sqrt{(\Delta\tau_i)^2 + (\Delta\tau_m)^2} \qquad \textbf{(38)}$$

Example 27.13 We consider a step index multimode fiber with $n_1 = 1.46$, and $\Delta = 0.01$ operating at 850 nm. For such a fiber, the intermodal dispersion (for a 1 km length of the fiber) is

$$\Delta\tau_i = \frac{n_1 L\Delta}{c} \approx \frac{1.46\times 1000\times 0.01}{3\times 10^8} \approx 49 \text{ ns}$$

which is usually written as

$$\Delta\tau_i \approx 49 \text{ ns km}^{-1}$$

If the source is an LED with $\Delta\lambda = 20$ nm, then using Table 10.1 the material dispersion $\Delta\tau_m$ is 1.7 ns km^{-1} (see Example 27.11). Thus in step index multimode fibers, the dominant pulse broadening mechanism is intermodal dispersion, and the total dispersion is given by

$$\Delta\tau = \sqrt{(\Delta\tau_i)^2 + (\Delta\tau_m)^2} = 49 \text{ ns km}^{-1} = 49\times 10^{-9} \text{ s km}^{-1}$$

Using Eq. (37), this gives a maximum bit rate of about

$$B_{\max} \approx \frac{0.7}{\Delta\tau} = \frac{0.7}{49\times 10^{-9}} \text{ bits km}^{-1}\text{ s}^{-1} \approx 14 \text{ Mbits km s}^{-1}$$

Thus a 10 km link can at most support only 1.4 Mbits s^{-1}.

Example 27.14 Let us now consider a parabolic index multimode fiber with $n_1 = 1.46$ and $\Delta = 0.01$ operating at 850 nm with an LED of spectral width 20 nm. For such a fiber, the intermodal dispersion, using Eq. (28), is

$$\Delta\tau_i = \frac{n_1}{2c}\Delta^2 L \approx 0.24 \text{ ns km}^{-1}$$

The material dispersion is again 1.7 ns km^{-1}. Thus in this case the dominant mechanism is material dispersion rather than intermodal dispersion. The total dispersion is

$$\Delta\tau = \sqrt{0.24^2 + 1.7^2} = 1.72 \text{ ns km}^{-1}$$

This gives a maximum bit rate of about

$$B_{\max} \approx \frac{0.7}{1.72\times 10^{-9}} \text{ bits km s}^{-1} \approx 400 \text{ Mbits km s}^{-1}$$

giving a maximum permissible bit rate of 20 Mbits s^{-1} for a 20 km link.

Example 27.15 If we now shift the wavelength of operation to 1300 nm and use the parabolic index fiber of Example 27.14, we see that the intermodal dispersion remains the same as 0.24 ns km^{-1} while the material dispersion (for an LED of $\Delta\lambda_0 = 20$ nm) becomes 0.05 ns km^{-1} (see Example 27.11). The material dispersion is now negligible in comparison to intermodal dispersion. Thus the total dispersion and maximum bit rate are, respectively, given by

$$\Delta\tau = \sqrt{0.24^2 + 0.05^2} = 0.25 \text{ ns km}^{-1}$$

$$\Rightarrow \quad B_{\max} = 2.8 \text{ Gbits km s}^{-1}$$

We reiterate that in the examples discussed above the maximum bit rate has been estimated by considering the fiber only. In an actual link, the temporal response of the source and detector must also be taken into account.

We end this section by mentioning that around 1977, we had the first-generation optical communication systems which used graded index multimode fibers, and the source used was the LED operating at 850 nm wavelength; the loss was ≈3 dB km^{-1}, the repeater spacing was ≈10 km, and the bit rate was ≈45 Mbits s^{-1}. Around 1981, we had the second-generation optical communication systems which again used graded index multimode fibers but now operating at 1300 nm wavelength (so that the material dispersion is very small); the bit rate was almost the same (≈45 Mbits s^{-1}) but since the loss was ≈1 dB km^{-1} and the dispersion was also less, the repeater spacing increased to ≈30 km. The third- and fourth-generation optical communication systems used single-mode fibers operating at 1300 and 1550 nm wavelengths, respectively.

27.12 GENERAL EXPRESSION FOR RAY DISPERSION CORRESPONDING TO A POWER LAW PROFILE

The time taken to propagate through a length L of a multimode fiber described by a q-profile [see Eq. (19)] is given by

$$\tau(\tilde{\beta}) = \left(A\tilde{\beta} + \frac{B}{\tilde{\beta}} \right) L \qquad \textbf{(39)}$$

where

$$A = \frac{2}{c(2+q)} \qquad B = \frac{q n_1^2}{c(2+q)} \qquad \textbf{(40)}$$

and for rays guided by the fiber $n_2 < \tilde{\beta} < n_1$. In the ray optics approximation Eq. (39) is rigorously correct [see Refs. 14 and 15 for derivation of Eq. (39)]. Using the above equations, we can calculate the ray dispersion in fibers with different q values. For the step profile, $q = \infty$ and

$$A = 0 \qquad B = \frac{n_1^2}{c} \quad \Rightarrow \quad \tau(\tilde{\beta}) = \frac{n_1^2}{c\tilde{\beta}} L \qquad \textbf{(41)}$$

Thus,

$$\tau_{\max} = \tau(\tilde{\beta} = n_2) = \frac{n_1^2}{c n_2} L$$

and

$$\tau_{\min} = \tau(\tilde{\beta} = n_1) = \frac{n_1}{c} L \qquad \textbf{(42)}$$

giving

$$\Delta\tau = \tau_{max} - \tau_{min} = \frac{n_1}{c}\frac{n_1 - n_2}{n_2}L \quad \textbf{(43)}$$

which is the same expression as given by Eq. (27). For the parabolic profile, $q = 2$ and

$$A = \frac{1}{2c} \quad \text{and} \quad B = \frac{n_1^2}{2c} \quad \Rightarrow \quad \tau(\tilde{\beta}) = \frac{1}{2c}\left[\tilde{\beta} + \frac{n_1^2}{\tilde{\beta}}\right]L \quad \textbf{(44)}$$

Thus,

$$\tau_{max} = \tau(\tilde{\beta} = n_2) = \frac{1}{2c}\left(n_2 + \frac{n_1^2}{n_2}\right)L$$

and

$$\tau_{min} = \tau(\tilde{\beta} = n_1) = \frac{n_1}{c}L \quad \textbf{(45)}$$

giving

$$\Delta\tau = \tau_{max} - \tau_{min} = \frac{n_2}{2c}\left(\frac{n_1 - n_2}{n_2}\right)^2 L \quad \textbf{(46)}$$

which is the same expression as given by Eq. (32). The calculation of the optimum value of q (which gives minimum ray dispersion) requires a plot of $\tau(\tilde{\beta})$ as a function of $\tilde{\beta}$ for different values of q. The details are given in Refs. 14 and 15, and the minimum dispersion occurs for $q \approx 2 - 2\Delta$ where the pulse dispersion is given by

$$\Delta\tau\,(\text{optimum profile}) = \frac{n_1}{8c}\left(\frac{n_1 - n_2}{n_2}\right)^2 L \quad \textbf{(47)}$$

However, because in a given fiber the profile itself depends on wavelength (because the refractive changes slightly with wavelength), most graded index fibers used in optical communication systems correspond to $q \approx 2$.

27.13 PLASTIC OPTICAL FIBERS

Plastic optical fibers (usually abbreviated as POFs) are fibers made from plastic materials such as PMMA (poly methyl methacrylate) ($n = 1.49$), polystyrene ($n = 1.59$), polycarbonates ($n = 1.5$–1.57), fluorinated polymers, etc. These fibers share the same advantages as glass optical fibers in terms of insensitivity to electromagnetic interference, small size and weight, low cost, and potential capacity to carry information at high rates. The most important attribute of POFs is their large core diameters of around 1 mm compared to glass fibers with core diameters around 50 μm. Such a large diameter (in POFs) results in easy alignments at joints. They are also more durable and flexible than glass fibers. In addition, they usually have a large numerical aperture and therefore much larger light-gathering power. Thus coupling to a POF is much easier than for a normal silica-based optical fiber. One of the major disadvantages of the POFs is their having much higher losses compared to silica-based fibers. The low-loss windows of POFs are around 570, 650, and 780 nm. For example, a graded index PMMA fiber will have a loss of about 110 dB km^{-1} around the wavelength of 650 nm. This value is much much larger than for silica fibers. Because of such high losses, POFs are never used in long-distance communication systems but are being used in intra-office communication systems where one requires only a few hundred meters of the fiber. Thus, although silica-based optical fibers dominate the long-distance optical communication systems, POFs are providing low-cost solutions to short-distance applications such as local-area networks (LANs), high-speed Internet access, etc.

27.14 FIBER-OPTIC SENSORS[7]

Although the most important application of optical fibers is in the field of transmission of information, optical fibers capable of sensing various physical parameters and generating information are finding widespread use as fiber-optic sensors. The use of optical fibers for such applications offers the same advantages as in the field of communication, i.e., lower cost, smaller size, greater accurately, greater flexibility and reliability. Compared to conventional electrical sensors, such fiber-optic sensors are immune to external electromagnetic interference and can also be used in hazardous and explosive environments. A very important attribute of fiber-optic sensors is the possibility of having distributed or quasi-distributed sensing geometries which would otherwise be too expensive or complicated using conventional sensors. Using fiber-optic sensors it is possible to measure pressure, temperature, electric current, rotation, strain, chemical and biological parameters, etc., with greater precision and speed. These advantages are leading to increased integration of such sensors into civil structures such as bridges and tunnels, process industries, medical instruments, aircrafts, missiles, and even cars.

Fiber-optic sensors can be broadly classified into two categories: extrinsic and intrinsic. In the case of extrinsic sensors, the optical fiber simply acts as a device to transmit

[7] Adapted from the unpublished lecture notes of Prof. K. Thyagarajan.

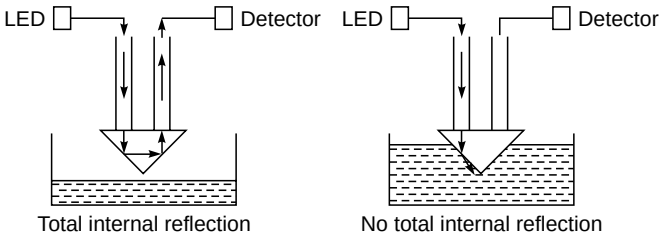


Fig. 27.16 A liquid level sensor based on changes in the critical angle due to liquid level moving up.

and collect light from a sensing element, which is external to the fiber. The sensing element responds to the external perturbation, and the change in the characteristics of the sensing element is transmitted by the return fiber for analysis. The optical fiber here plays no role other than that of transmitting the light beam. On the other hand, in the case of intrinsic sensors, the physical parameter to be sensed directly alters the properties of the optical fiber, which in turn leads to changes in a characteristic such as intensity, polarization, phase, etc., of the light beam propagating in the fiber.

A large variety of fiber-optic sensors have been demonstrated in the laboratory, and some are already being installed in real systems. In Sec. 15.6.1 we discussed temperature sensors (using fiber Bragg gratings) used on a power conductor at an electric power substation. Fiber Bragg gratings are also used in strain measurements in bridges and tunnels. In Sec. 22.15 we discussed fiber-based sensors for measuring large currents.

An interesting sensor is the liquid level sensor shown in Fig. 27.16. Light propagating down an optical fiber is total internally reflected from a small glass prism and couples back to the return fiber. As long as the external medium is air, the angle of incidence inside the prism is greater than the critical angle, and hence light suffers total internal reflection. As soon as the prism comes in contact with a liquid, the critical angle at the prism-liquid interface reduces and the light gets transmitted into the liquid, resulting in a loss of signal. By a proper choice of prism material, such a sensor can be used for sensing levels of various liquids such as water, gasoline, acids, oils, etc. More details about fiber-based sensors can be found in Refs. 14 and 16–17.

Problems

27.1 Consider a step index fiber with $n_1 = 1.5$, $\Delta = 0.015$, and $a = 25$ μm placed in air. Calculate n_2 and the maximum acceptance angle i_m. If the fiber tip is immersed in water ($n = 1.33$), calculate the maximum acceptance angle i_m.

[**Ans.:** 1.477; 0.26; 15°; 11.3°]

27.2 A step index optical fiber with $n_1 = 1.46$, $n_2 = 1.44$, and core radius $a = 50$ μm is placed in air. Calculate the maximum acceptance angle. If the fiber is now immersed in water ($n = 1.33$), calculate the maximum acceptance angle.

[**Ans.:** 13.9°; 10.4°]

27.3 A step index fiber with $n_1 = 2$, and $n_2 = \sqrt{3}$ is placed in air. What is the maximum angle an incident ray can make with the axis of the fiber at the input end in air, so that it is guided after entering the fiber?

[**Ans.:** 90°]

27.4 Consider a bare fiber with: $n_1 = 1.46$ (pure silica), $n_2 = 1.0$ (air), and core radius $a = 30$ μm.

(a) Show that all rays (inside the core) making an angle $\theta < 46.77°$ with the z axis will be guided through the fiber.

(b) Assume $\theta = 30°$ and calculate the number of reflections that will occur in propagating through 1 km length of the fiber. Assume only a 0.01% decrease in power at each reflection; calculate the power loss at each reflection and in propagating through 1 km length of the fiber.

[**Ans.:** 9.6×10^6, 4.34×10^{-4} dB, 4179 dB km^{-1}]

27.5 The power of a 2 mW laser beam decreases to 15 μW after traversing through 25 km of a single-mode optical fiber. Calculate the attenuation of the fiber.

[**Ans.:** 0.85 dB km^{-1}]

27.6 A 5 mW laser beam passes through a 26 km fiber of loss 0.2 dB km^{-1}. Calculate the power at the output end.

[**Ans.:** 1.5 mW]

27.7 Consider a 15 mW laser beam passing through a 40 km fiber link of loss 0.5 dB km^{-1}. Calculate the output power in dBm and then in mW.

[**Ans.:** 0.15 mW]

27.8 The power of a 10 mW laser beam decreases to 40 μW after traversing through 40 km of an optical fiber. Calculate the attenuation of the fiber in dB km^{-1}.

[**Ans.:** 0.6 dB km^{-1}]

27.9 Consider a 50 km fiber link (with a loss of 0.25 dB km^{-1}) having four connectors in its path. If each connector has a loss of 1.8 dB, then calculate the total loss. The loss at the source to the fiber is 2 dB, and the loss from the fiber to the detector is 2.5 dB. The input laser power is 10 mW; calculate the output power in dBm and in mW.

27.10 **(a)** Consider a step index fiber with $n_1 = 1.46$, $n_2 = 1.44$, and $a = 50$ μm. Assume that the operating wavelength $\lambda_0 = 0.85$ μm, calculate the V value and show that it is a multimode fiber. Calculate the ray dispersion in ns km^{-1}.

(b) Next consider a bare step index fiber with $n_1 = 1.46$, $n_2 = 1.0$, and $a = 50$ μm. Assume that the operating wavelength $\lambda_0 = 0.85$ μm, calculate the V value and

show that it is a multimode fiber. Calculate the ray dispersion.

[**Ans.: (a)** 67.6 ns km^{-1}; **(b)** 2239 ns km^{-1}]

27.11 Assume that the material dispersion coefficient D_m is given by

$$D_m = \frac{\Delta\tau_m}{L\,\Delta\lambda_0} = -\frac{10^4}{3\lambda_0}\left(\lambda_0^2\,\frac{d^2n}{d\lambda_0^2}\right) \text{ ps km}^{-1}\text{ nm}^{-1}$$

where λ_0 is measured in μm. For silica fibers $d^2n/d\lambda_0^2 \approx 0.0297\ (\mu m)^{-2}$ at $\lambda_0 = 0.85$ μm; $\approx 0.0120\ (\mu m)^{-2}$ at $\lambda_0 = 1.0$ μm; $\approx -0.00055\ (\mu m)^{-2}$ at $\lambda_0 = 1.30$ μm; and $\approx -0.00416\ (\mu m)^{-2}$ at $\lambda_0 = 1.55$ μm.

(a) At $\lambda_0 = 0.85$, 1.0, 1.30, and 1.55 μm, calculate the material dispersion (in ns km^{-1}) when $\Delta\lambda_0$ (the spectral width of source) is 50 nm (LED) and 2.5 nm (LD), respectively.

(b) Consider a SIF with $n_1 = 1.5$, $a = 40$ μm, and $\Delta = 0.015$ operating at 850 nm with a spectral width of 50 nm. Is this a single-mode fiber or a multimode fiber? Calculate the material dispersion, ray dispersion, total pulse dispersion, and hence the maximum bit rate.

(c) Next, consider a parabolic index fiber with $n_1 = 1.5$, $a = 40$ μm, and $\Delta = 0.015$ operating at 850 nm with a spectral width of 50 nm. Is this a single-mode fiber or a multimode fiber? Calculate the material dispersion, ray dispersion, total pulse dispersion, and hence the maximum bit rate.

(d) Finally, consider a parabolic index fiber with $n_1 = 1.5$, $a = 40$ μm, and $\Delta = 0.015$ operating at 1300 nm with a spectral width of 50 nm. Calculate the material dispersion, ray dispersion, total pulse dispersion, and hence the maximum bit rate.

[**Ans.: (b)** 4.2, 75, 75.1 ns km^{-1};
(c) 4.2, 0.6, 4.2 ns km^{-1}]

REFERENCES AND SUGGESTED READINGS

1. J. Hecht, *City of Light*, Oxford, 1999.
2. W. A. Gambling, "Glass, Light, and the Information Revolution," Ninth W. E. S. Turner Memorial Lecture, *Glass Technology*, Vol. 27, No. 6, p. 179, 1986.
3. B. P. Pal (Ed.), *Guided Wave Optical Components and Devices: Basics, Technology and Applications*, Academic Press, 2006.
4. http://en.wikipedia.org/wiki/Photophone.
5. D. J. H. Maclean, *Optical Line Systems*, Wiley, Chichester, 1996.
6. A. G. Chynoweth, "Lightwave Communications: The Fiber Lightguide," *Phys. Today*, Vol. 29, No. 5, p. 28, 1976.
7. C. K. Kao and G. A. Hockham, "Dielectric-Fibre Surface Waveguides for Optical Frequencies," *Proc. IEE*, Vol. 113, No. 7; p. 1151, 1966.
8. F. P. Kapron, D. B. Keck, and R. D. Maurer, "Radiation Losses in Glass Optical Waveguides," *Appl. Phys. Lett.*, Vol. 17, p. 423, 1970.
9. "Schott is Lighting the Way Home," *Fiberoptic Product News*, February 1997, p. 13.
10. "Fiber Optic Technology Put to Work—Big Time," *Photonics Spectra*, August 1994, p. 114.
11. A. Ghatak and K. Thyagarajan, "Optical Waveguides and Fibers," in *Fundamentals of Photonics* (Eds. A. Guenther, L. Pedrotti, and C. Roychoudhuri), Materials developed under project STEP (Scientific and Technology Education in Photonics) by University of Connecticut and CORD, National Science Foundation, United States.
12. T. Miya, Y. Terunama, T. Hosaka, and T. Miyashita, "An Ultimate Low Loss Single Mode Fiber at 1.55 mm," *Electron. Letts.* Vol. 15, p. 106, 1979.
13. D. Gloge and E. A. J. Marcatili, "Multimode Theory of Graded-Core Fibers," *Bell. Syst. Tech. J.*, Vol. 52, p. 1563, 1973.
14. A. Ghatak and K. Thyagarajan, *Introduction to Fiber Optics*, Cambridge University Press, Cambridge, 1998.
15. A. Ankiewicz and C. Pask, "Geometric Optics Approach to Light Acceptance and Propagation in Graded Index Fibers," *Opt. Quant. Electr*. Vol. 9, p. 87, 1977.
16. B. Culshaw, "Principles of Fiber Optic Sensors," in *Guided Wave Optical Components and Devices* (Ed. B. P. Pal), Academic Press, Amsterdam, 2006.
17. B. D. Gupta, *Fiber Optic Sensors: Principles and Applications*, New India Publishing Agency, New Delhi, 2006.

Chapter Twenty-Eight

OPTICAL WAVEGUIDES II: BASIC WAVEGUIDE THEORY AND CONCEPT OF MODES

28.1 INTRODUCTION

In the design of an optical communication system, it is necessary to have a good understanding of the propagation characteristics of the optical fiber. In Chap. 27 we used ray optics to understand the propagation characteristics of the optical fiber. Such an analysis is valid when the fiber supports a large number of modes. However, today single-mode fibers are extensively used in optical communication systems. And in single-mode fibers, ray optics is not applicable and one has to solve Maxwell's equations to determine the modes of the waveguide. Thus the first thing to do is to understand the concept of modes, which we plan to do in this chapter. And to understand the concept of modes, it is probably best to consider the simplest planar optical waveguide that consists of a thin dielectric film sandwiched between materials of slightly lower refractive indices which is characterized by the following refractive index variation (see Fig. 28.1):

$$n(x) = \begin{cases} n_1 & |x| < \dfrac{d}{2} \\ n_2 & |x| > \dfrac{d}{2} \end{cases} \qquad \textbf{(1)}$$

with $n_1 > n_2$. Equation (1) describes what is usually referred to as a step index profile. The waveguide is assumed to extend to infinity in the y and z directions. To start, we consider a more general case of the refractive index depending only on the x coordinate:

$$n^2 = n^2(x) \qquad \textbf{(2)}$$

When the refractive index variation depends only on the x coordinate, we can always choose the z axis along the direction of propagation of the wave, and we may, *without any loss of generality*, write the solutions of Maxwell's equations in the form

$$\mathcal{E} = \mathbf{E}(x)e^{i(\omega t - \beta z)} \qquad \textbf{(3)}$$

$$\mathcal{H} = \mathbf{H}(x)e^{i(\omega t - \beta z)} \qquad \textbf{(4)}$$

The above equations *define* the modes of the system. Thus

> Modes represent transverse field distributions that suffer a phase change only as they propagate through the waveguide along z.

The transverse field distributions described by $\mathbf{E}(x)$ and $\mathbf{H}(x)$ do not change as the field propagates through the waveguide. The quantity β represents the propagation constant of the mode. If we substitute the above solutions in Maxwell's equations, we obtain two independent sets of equations (see App. E). The first set of equations corresponds to nonvanishing values of E_y, H_x, and H_z with E_x, E_z, and H_y vanishing, giving rise to what are known as TE modes because the electric field has only a transverse component. The second set of equations corresponds to nonvanishing values of E_x, E_z, and H_y with E_y, H_x, and H_z vanishing, giving rise to what are known as TM modes because the magnetic field now has only a transverse component.

For TE modes, we show in App. E that $E_y(x)$ satisfies the following differential equation

$$\frac{d^2E_y}{dx^2} + [k_0^2 n^2(x) - \beta^2]E_y = 0 \qquad \textbf{(5)}$$

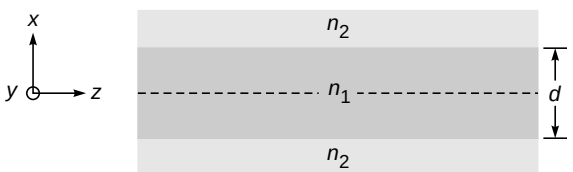


Fig. 28.1 A planar dielectric waveguide of thickness d (along x direction) but infinitely extended along the y direction. Light propagates along the z direction.

where

$$k_0 = \omega\sqrt{\varepsilon_0\mu_0} = \frac{\omega}{c} \qquad (6)$$

is the free space wave number and $c\ \left(=1/\sqrt{\varepsilon_0\mu_0}\right)$ is the speed of light in free space. Once $E_y(x)$ is known, we can determine H_x and H_z from the following equations (see App. E):

$$H_x = -\frac{\beta}{\omega\mu_0}E_y(x) \quad \text{and} \quad H_z = \frac{i}{\omega\mu_0}\frac{dE_y}{dx} \qquad (7)$$

Equations (3) to (7) are rigorously correct as long as the refractive index distribution depends only on the x coordinate. Equation (5) is an eigenvalue equation with β^2 representing the eigenvalue. By applying the appropriate boundary conditions, we will show that β^2 can have a set of discrete values (corresponding to guided modes of the waveguide) and also a continuum of values corresponding to the radiation modes of the waveguide.

28.2 TE MODES OF A SYMMETRIC STEP INDEX PLANAR WAVEGUIDE[1]

Until now our analysis has been valid for an arbitrary x-dependent profile. We now assume that the refractive index variation is given by Eq. (1) (see Fig. 28.1). Substituting for $n(x)$ in Eq. (5), we obtain

$$\frac{d^2E_y}{dx^2} + (k_0^2n_1^2 - \beta^2)E_y = 0 \qquad |x| < \frac{d}{2} \quad \text{film} \qquad (8)$$

$$\frac{d^2E_y}{dx^2} + (k_0^2n_2^2 - \beta^2)E_y = 0 \qquad |x| > \frac{d}{2} \quad \text{cover} \qquad (9)$$

We will solve Eqs. (8) and (9) subject to the appropriate boundary and continuity conditions. Since E_y and H_z represent tangential components on the planes $x = \pm d/2$, they must be continuous at $x = \pm d/2$; and since H_z is proportional to dE_y/dx [see Eq. (7)], we must have

$$E_y \text{ and } \frac{dE_y}{dx} \text{ continuous at } x = \pm\frac{d}{2} \qquad (10)$$

The above represents the continuity conditions that have to be satisfied.[2] Now, guided modes are those modes that are mainly confined to the film, and hence their field should decay in the cover, i.e., the field should decay in the region $|x| > d/2$, so that most of the energy associated with the modes lies inside the film. Thus, we must have

$$\beta^2 > k_0^2n_2^2 \qquad (11)$$

When $\beta^2 < k_0^2n_2^2$, the solutions are oscillatory in the region $|x| > d/2$ and they correspond to what are known as *radiation modes* of the waveguide. These radiation modes correspond to rays that undergo refraction (rather than total internal reflection) at the film-cover interface, and when these are excited, they quickly leak away from the core of the waveguide. Furthermore, we must also have $\beta^2 < k_0^2n_1^2$; otherwise, the boundary conditions cannot be satisfied[3] at $x = \pm d/2$. Thus, for guided modes we must have

$$n_2^2 < \frac{\beta^2}{k_0^2} < n_1^2 \qquad \text{guided modes} \qquad (12)$$

One often defines the effective index of the mode as

$$n_{\text{eff}} \equiv \frac{\beta}{k_0} \qquad (13)$$

giving

$$n_2 < n_{\text{eff}} < n_1 \qquad (14)$$

Recall the discussion in Sec. 3.4 where we said that for an optical waveguide, guided rays correspond to

$$n_2 < \tilde{\beta} < n_1 \qquad \text{guided rays} \qquad (15)$$

and refracting rays correspond to $\tilde{\beta} < n_2$; further, there cannot be any ray with $\tilde{\beta} > n_1$. Thus, $\tilde{\beta}$ (in ray optics) can be said to correspond to β/k_0 in wave optics:

$$\tilde{\beta} \Leftrightarrow \frac{\beta}{k_0} = n_{\text{eff}} \qquad (16)$$

In Sec. 3.4, we defined the parameter $\tilde{\beta}$ as equal to $n(x)\cos\theta(x)$, where $\theta(x)$ was the angle that the ray makes with the z axis. For the step index waveguide $\tilde{\beta} = n_1\cos\theta$ where the angle θ (that the ray makes with the z axis) remains constant within the core of the waveguide. In Sec. 28.3 we will show that in a step index waveguide, a mode can be represented as a superposition of two plane waves propagating at angles $\pm\cos^{-1}(\beta/k_0n_1)[=\pm\cos^{-1}(n_{\text{eff}}/n_1)]$ with the z axis. Since n_{eff} will be shown to take a set of discrete values, each discrete value of n_{eff} will therefore correspond to a discrete value of θ (see also Sec. 28.3).

---

[1] More details about waveguide modes can be found in Refs. 1 to 4.

[2] The very fact that E_y satisfies Eq. (5) also implies that E_y and dE_y/dx are continuous unless $n^2(x)$ has an infinite discontinuity. The follows from the fact that if dE_y/dx is discontinuous, then d^2E_y/dx^2 will be a delta function (see Prob. 9.2) and Eq. (5) will lead to an inconsistent equation.

[3] It is left as an exercise for the reader to show that if we assume $\beta^2 > k_0^2n_1^2$ and also assume decaying fields in the region $|x| > d/2$, then the boundary conditions at $x = +d/2$ and at $x = -d/2$ can never be satisfied simultaneously.

We use Eq. (12) to write Eqs. (8) and (9) in the form

$$\frac{d^2E_y}{dx^2} + \kappa^2 E_y = 0 \qquad |x| < \frac{d}{2} \quad \text{film} \tag{17}$$

$$\frac{d^2E_y}{dx^2} - \gamma^2 E_y = 0 \qquad |x| > \frac{d}{2} \quad \text{cover} \tag{18}$$

where

$$\kappa^2 = k_0^2 n_1^2 - \beta^2 \tag{19}$$

and

$$\gamma^2 = \beta^2 - k_0^2 n_2^2 \tag{20}$$

Now, when the refractive index distribution is symmetric about $x = 0$, that is, when

$$n^2(-x) = n^2(x) \tag{21}$$

the solutions are either symmetric or antisymmetric functions of x (see Prob. 28.8; see also pp. 126–127 of Ref. 2)[4]; thus we must have

$$E_y(-x) = E_y(x) \qquad \text{symmetric modes} \tag{22}$$

$$E_y(-x) = -E_y(x) \qquad \text{antisymmetric modes} \tag{23}$$

For the symmetric mode, we must have

$$E_y(x) = \begin{cases} A\cos\kappa x & |x| < \dfrac{d}{2} \\ Ce^{-\gamma|x|} & |x| > \dfrac{d}{2} \end{cases} \tag{24}$$

where we have neglected the exponentially amplifying solution in the region $|x| > d/2$. Continuity of $E_y(x)$ and dE_y/dx at $x = \pm d/2$ gives

$$A\cos\frac{\kappa d}{2} = Ce^{-\gamma d/2} \tag{25}$$

and

$$-\kappa A\sin\frac{\kappa d}{2} = -\gamma Ce^{-\gamma d/2} \tag{26}$$

respectively. Dividing Eq. (26) by Eq. (25), we get

$$\xi\tan\xi = \frac{\gamma d}{2} \tag{27}$$

where

$$\xi \equiv \frac{\kappa d}{2} \tag{28}$$

Now, if we add Eqs. (19) and (20), we get

$$\left(\kappa^2 + \gamma^2\right)\frac{d^2}{4} = \frac{1}{4}\left[k_0^2 d^2\left(n_1^2 - n_2^2\right)\right] = \frac{1}{4}V^2 \tag{29}$$

where

$$V = k_0 d\sqrt{n_1^2 - n_2^2} \tag{30}$$

is known as the dimensionless waveguide parameter, which is an extremely important parameter in waveguide theory. Using Eqs. (28) and (29), we can write

$$\frac{\gamma d}{2} = \sqrt{\tfrac{1}{4}V^2 - \xi^2} \tag{31}$$

and Eq. (27) can be put in the form

$$\xi\tan\xi = \sqrt{\tfrac{1}{4}V^2 - \xi^2} \tag{32}$$

Similarly, for the antisymmetric mode we have

$$E_y(x) = \begin{cases} B\sin\kappa x & |x| < -\dfrac{d}{2} \\ De^{-\gamma x} & x > \dfrac{d}{2} \\ -De^{\gamma x} & x < -\dfrac{d}{2} \end{cases} \tag{33}$$

and following an exactly similar procedure, we get

$$-\xi\cot\xi = \sqrt{\tfrac{1}{4}V^2 - \xi^2} \tag{34}$$

Thus, we have

$$\xi\tan\xi = \sqrt{\left(\frac{V}{2}\right)^2 - \xi^2} \qquad \text{for symmetric modes} \tag{35}$$

and

$$-\xi\cot\xi = \sqrt{\left(\frac{V}{2}\right)^2 - \xi^2} \qquad \text{for antisymmetric modes} \tag{36}$$

Since the equation

$$\eta = \sqrt{\left(\frac{V}{2}\right)^2 - \xi^2} \tag{37}$$

(for positive values of ξ) represents a circle (of radius $V/2$) in the first quadrant of the $\xi\eta$ plane,[5] the numerical evaluation of the allowed values of ξ (and hence of the propagation constants) is quite simple. In Fig. 28.2 we have plotted the functions $\xi\tan\xi$ (solid curve) and $-\xi\cot\xi$ (dashed curve) as a function of ξ. For a given value of V, the points of intersection of these curves with the quadrant of the circle determine the allowed (discrete) values of ξ. The two circles in Fig. 28.2 correspond to $V/2 = 2$ and $V/2 = 5$. Obviously, as can be seen from the figure, for $V = 4$ we will have one symmetric and one antisymmetric mode, and for $V = 10$ we will have two symmetric and two antisymmetric modes.[6]

[4] The same situation arises in quantum mechanics—see, e.g., pp. 155–157 of Ref. 5.

[5] This follows from the fact that if we square Eq. (37), we get $\eta^2 + \xi^2 = (V/2)^2$, which represents a circle of radius $V/2$.

[6] Those who are familiar with basic quantum mechanics will notice that the procedure for determining the discrete TE modes in a planar waveguide is almost identical to the one used in obtaining the discrete energy eigenvalues of the one-dimensional Schrödinger equation. Similarly, the modal analysis of the parabolic index planar waveguide is almost identical to the linear harmonic oscillator problem in quantum mechanics (see Sec. 28.6).

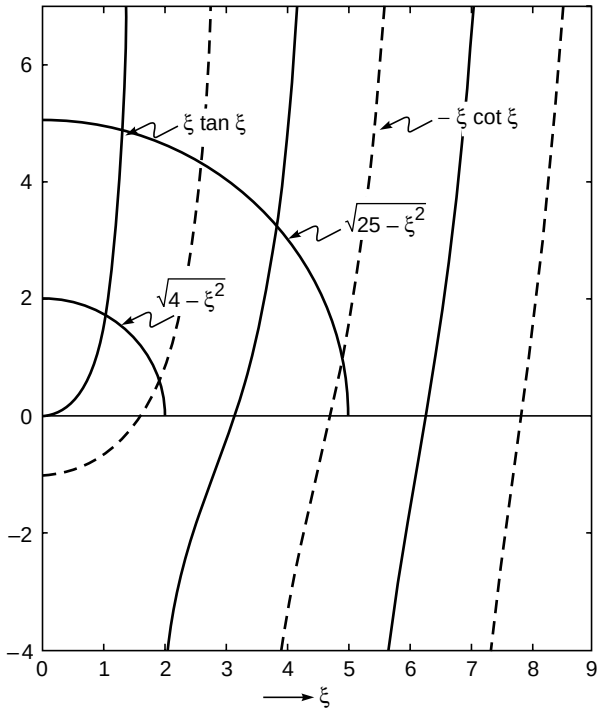


Fig. 28.2 Variation of $\xi \tan \xi$ (solid curve) and $-\xi \cot \xi$ (dashed curve) as a function of ξ. The points of intersection of these curves with the quadrant of a circle of radius $V/2$ determine the discrete propagation constants of the waveguide.

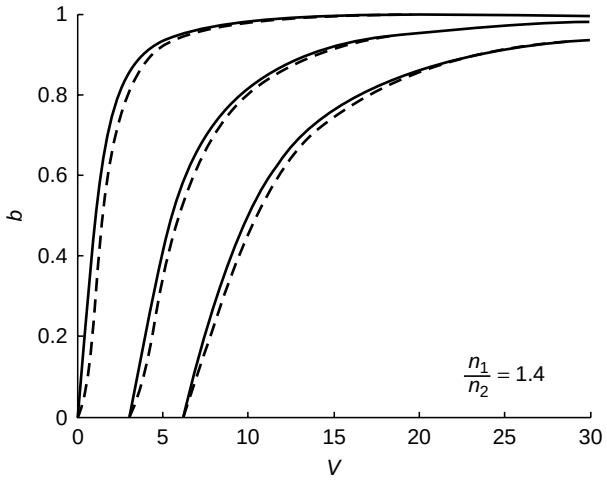


Fig. 28.3 Dependence of b on V for a step index planar waveguide. For the TE modes, the b–V curves are universal; however, for the TM modes, the b–V curves require the value of n_1/n_2. For the curves shown in the figure $n_1/n_2 = 1.4$; calculations courtesy of Triranjita Srivastava.

It is often very convenient to define the dimensionless propagation constant

$$b \equiv \frac{\beta^2/k_0^2 - n_2^2}{n_1^2 - n_2^2} = \frac{n_{\text{eff}}^2 - n_2^2}{n_1^2 - n_2^2} \qquad \textbf{(38)}$$

Thus

$$b = \frac{\beta^2 - k_0^2 n_2^2}{k_0^2\left(n_1^2 - n_2^2\right)} = \frac{\gamma^2 d^2}{V^2}$$

giving

$$\frac{\gamma d}{2} = \frac{1}{2} V\sqrt{b} \qquad \textbf{(39)}$$

Further, using Eqs. (29) and (39), we can write

$$\xi = \frac{\kappa d}{2} = \sqrt{\frac{1}{4}V^2 - \frac{\gamma^2 d^2}{4}}$$

$$= \frac{1}{2}V\sqrt{1-b} \qquad \textbf{(40)}$$

Thus Eqs. (35) and (36) can be written in the form

$$\left(\tfrac{1}{2}V\sqrt{1-b}\right)\tan\left(\tfrac{1}{2}V\sqrt{1-b}\right) = \tfrac{1}{2}V\sqrt{b}$$

for symmetric modes **(41)**

$$-\left(\tfrac{1}{2}V\sqrt{1-b}\right)\cot\left(\tfrac{1}{2}V\sqrt{1-b}\right) = \tfrac{1}{2}V\sqrt{b}$$

for antisymmetric modes **(42)**

Obviously, because of Eq. (12), for guided modes we will have

$$0 < b < 1 \qquad \textbf{(43)}$$

For a given value of V, solutions of Eqs. (41) and (42) will give us discrete values of b; the mth solution ($m = 0, 1, 2, 3, \ldots$) is referred to as the TE_m mode. In Table 28.1, we have tabulated the discrete values of b for various values of V; these discrete values have been obtained by using the software in Ref. 7. The universal curves describing the dependence of b on V are shown in Fig. 28.3. For any given (step index) waveguide we just have to calculate V and then obtain the corresponding value of b either by solving Eqs. (41) and (42) or by using Table 28.1. From the values of b, one can obtain the propagation constants by using the following equation [see Eq. (36)]:

$$\frac{\beta}{k_0} = \sqrt{n_2^2 + b\left(n_1^2 - n_2^2\right)} \qquad \textbf{(44)}$$

Figure 28.4 shows typical field patterns of some of the low-order TE_m modes of a step index waveguide.

Example 28.1 We consider a step index planar waveguide with $d = 3$ μm, $n_1 = 1.5$, and $n_2 = 1.49153$. The value of n_2 is chosen

Table 28.1 Values of the normalized propagation constant (corresponding to TE modes) for a symmetric planar waveguide; the values are generated by using the software in Ref. 7. Notice that for $V < \pi$ we will have only one TE mode which will be symmetric in x, and for $\pi < V < 2\pi$ we will have two TE modes; one of them will be symmetric in x and the other antisymmetric in x.

V	$b(\mathrm{TE}_0)$	$b(\mathrm{TE}_1)$
1.000	0.189339	
1.125	0.225643	
1.250	0.261714	
1.375	0.297049	
1.500	0.331290	
1.625	0.364196	
1.750	0.395618	
1.875	0.425479	
2.000	0.453753	
2.125	0.480453	
2.250	0.505616	
2.375	0.529300	
2.500	0.551571	
2.625	0.572502	
2.750	0.592169	
2.875	0.610649	
3.000	0.628017	
3.125	0.644344	
3.250	0.659701	0.002702
3.375	0.674151	0.011415
3.500	0.687758	0.024612
3.625	0.700579	0.041077
3.750	0.712667	0.059875
3.875	0.724073	0.080292
4.000	0.734844	0.101775

V	$b(\mathrm{TE}_0)$	$b(\mathrm{TE}_1)$	$b(\mathrm{TE}_2)$
4.000	0.734844	0.101775	
4.125	0.745021	0.123903	
4.250	0.754647	0.146349	
4.375	0.763756	0.168864	
4.500	0.772384	0.191259	
4.625	0.780563	0.213390	
4.750	0.788321	0.235151	
4.875	0.795686	0.256461	
5.000	0.802683	0.277265	
5.125	0.809335	0.297523	
5.250	0.815663	0.317210	
5.375	0.821689	0.336310	
5.500	0.827429	0.354817	
5.625	0.832902	0.372731	
5.750	0.838123	0.390056	
5.875	0.843107	0.406800	
6.000	0.847869	0.422976	
6.125	0.852420	0.438596	
6.250	0.856772	0.453676	
6.375	0.860938	0.468231	0.001845
6.500	0.864926	0.482278	0.008819
6.625	0.868748	0.495834	0.019189
6.750	0.872412	0.508916	0.031806
6.875	0.875926	0.521541	0.045942
7.000	0.879298	0.533727	0.061106

such that $\sqrt{n_1^2 - n_2^2} = 1/2\pi$ so that

$$V = \frac{2\pi}{\lambda_0} d\sqrt{n_1^2 - n_2^2} = \frac{d}{\lambda_0} = \frac{3}{\lambda_0}$$

where λ_0 is measured in μm

and

$$\frac{\beta}{k_0} = \sqrt{n_2^2 + \frac{b}{4\pi^2}}$$

For $\lambda_0 = 1.5$ μm, V is equal to 2.0 and from Table 28.1 we see that there will be only one TE mode with $b = 0.453753$; the corresponding value of $\beta/k_0 \approx 1.49538$. The same waveguide operating at $\lambda_0 = 1.0$ μm will have $V = 3.0$, and from Table 28.1 we see that there will be again only one TE mode with $b = 0.628017$; the corresponding value of $\beta/k_0 \approx 1.49686$. However, for $\lambda_0 = 0.6$ μm, $V = 5.0$ and there will be two TE modes with $b = 0.802683$ (the TE_0 mode) and the other with $b = 0.277265$ (the TE_1 mode). The corresponding values of $\beta/k_0 \approx 1.49833$ and 1.49389. Finally, for $\lambda_0 = 0.4286$ μm, $V = 7.0$ and there will be three TE modes with $b = 0.879298$ (TE_0), 0.533727 (TE_1), and 0.061106 (TE_2). The corresponding values of β/k_0 are ≈ 1.4990, 1.49606, and 1.49205, respectively. Notice that all the values of β/k_0 lie between n_1 and n_2. Note that in each case, the waveguide will support an equal number of TM modes (see Sec. 28.4). Further, as the wavelength is made smaller, the waveguide will support a larger number of modes, and in the limit of the wavelength tending to zero, we will have a continuum of modes which is nothing but the ray optics limit.

Example 28.2 We next consider a step index planar waveguide with $d = 2.5$ μm, $n_1 = 1.5$, and $n_2 = 1.47$. Assuming the operating wavelength $\lambda_0 = 1.0$ μm, we get $V = 4.6888$. If we carry out linear interpolation, we obtain for the TE_0 mode

$$b = 0.780563 + \frac{0.788321 - 0.780563}{0.125} \times 0.0638 \approx 0.78452$$

We therefore get $\beta/k_0 \approx 1.49359$. Similarly for the TE_1 mode,

$$b = 0.213390 + \frac{0.235151 - 0.213390}{0.125} \times 0.0638 \approx 0.22450$$

and the corresponding value of β/k_0 will be ≈ 1.47679.

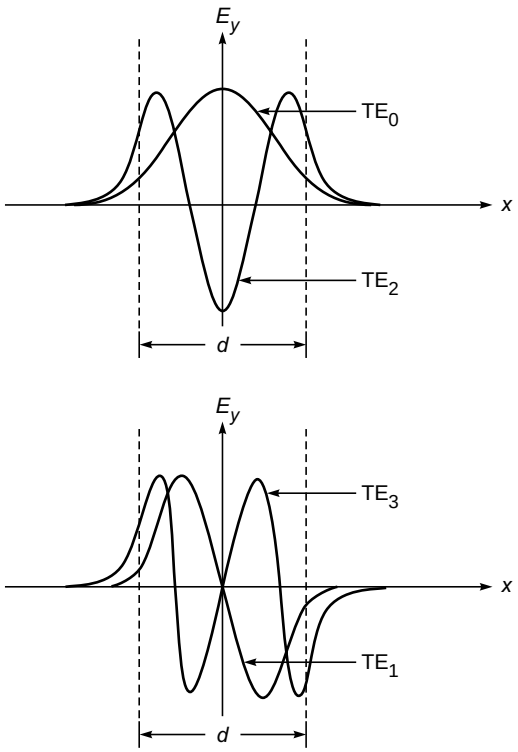


Fig. 28.4 Typical mode field distributions for TE modes in a step index planar waveguide with $n_1 = 1.49$, $n_2 = 1.56$, $a = 4$ μm, and $\lambda = 0.6328$ μm. TE_0 and TE_2 are symmetric in x and are known as even modes, while TE_1 and TE_3 are antisymmetric in x and are known as odd modes.

28.3 PHYSICAL UNDERSTANDING OF MODES

To have a physical understanding of modes, we consider the electric field pattern inside the film ($-d/2 < x < d/2$). For example, for a symmetric TE mode, this is given by [see Eq. (22)] $E_y(x) = A \cos \kappa x$. Thus the complete field inside the film is given by

$$E_y(x) = A \cos \kappa x e^{i(\omega t - \beta z)}$$
$$= \tfrac{1}{2} A e^{i(\omega t - \beta z - \kappa x)} + \tfrac{1}{2} A e^{i(\omega t - \beta z + \kappa x)} \quad \textbf{(45)}$$

Now

$$e^{i(\omega t - \mathbf{k}\cdot\mathbf{r})} = e^{i(\omega t - k_x x - k_y y - k_z z)}$$

represents a wave propagating along the direction of **k** whose x, y, and z components are k_x, k_y, and k_z, respectively. Thus, for the two terms on the RHS of Eq. (45) we will have

$$k_x = \kappa \qquad k_y = 0 \qquad k_z = \beta \quad \textbf{(46)}$$

and

$$k_x = -\kappa \qquad k_y = 0 \qquad k_z = \beta \quad \textbf{(47)}$$

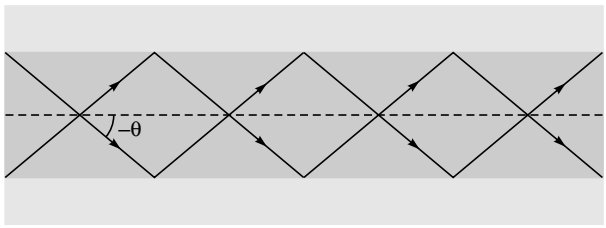


Fig. 28.5 A guided mode in a step index waveguide corresponds to the superposition of two plane waves propagating at particular angles $\pm\theta$ with the z axis.

which represent plane waves with propagation vectors parallel to the xz plane making angles $+\theta$ and $-\theta$ with the z axis (see Fig. 28.5) where

$$\tan \theta = \frac{k_x}{k_z} = \frac{\kappa}{\beta}$$

or

$$\cos \theta = \frac{\beta}{\sqrt{\beta^2 + \kappa^2}} = \frac{\beta}{k_0 n_1} \quad \textbf{(48)}$$

Thus, a guided mode can be considered to be

> A superposition of two plane waves propagating at angles $\pm \cos^{-1} \dfrac{\beta}{k_0 n_1}$ with the z axis

(see Fig. 28.5). Referring to the waveguide discussed in Example 28.1, at $\lambda_0 = 0.6$ μm, V will be 5.0 and we will have two TE modes with $\beta/k_0 \approx 1.49833$ and 1.49389. Since $n_1 = 1.5$, the values of $\cos\theta$ will be 0.99889 and 0.99593 and therefore

$$\theta \approx 2.70^\circ \quad \text{and} \quad 5.17^\circ$$

corresponding to the symmetric TE_0 mode and the antisymmetric TE_1 mode, respectively. Each mode is therefore characterized by a discrete angle of propagation θ_m. According to ray optics, the angle θ could take *all* possible values from 0 (corresponding to a ray propagating parallel to the z axis) to $\cos^{-1}(n_2/n_1)$ (corresponding to a ray incident at the critical angle on the core-cladding interface). However, we now find that according to wave optics, only discrete values of θ are allowed, and each "discrete" ray path corresponds to a mode of the waveguide. This is the basic principle of the prism film-coupling technique for determining the (discrete) propagation constants of an optical waveguide (see Fig. 28.6). The method consists of placing a prism (whose refractive index is greater than that of the film) close to the waveguiding film. In the presence of the prism, the rays undergo refraction and leak away from the waveguide. The direction at which the light beam emerges from the prism is directly related to θ_m. From the measured values of θ_m one

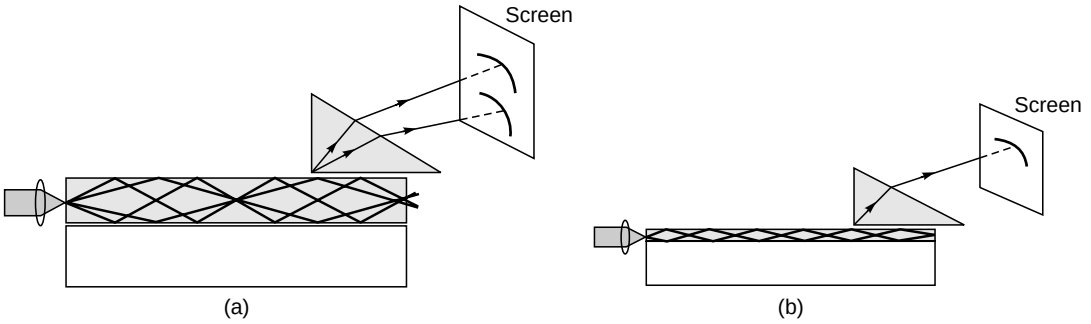


Fig. 28.6 The prism film-coupling technique for determining the (discrete) propagation constants of an optical waveguide.

can obtain the discrete values of the propagation constant β by using the formula

$$\beta = k_0 n_1 \cos\theta \qquad \textbf{(49)}$$

As mentioned in Sec. 27.9, for a given waveguide, if λ_0 is made to go to 0, the value of V becomes very large and the waveguide will support a very large number of modes. In this limit, we can assume all values of θ to be allowed, and it will be quite appropriate to use ray optics in studying the propagation characteristics of the waveguide.

From Fig. 28.3 we can derive the following conclusions about TE modes (similar discussion can be had for TM modes, which are discussed in the next section):

1. If $0 < V/2 < \pi/2$, that is, when
$$0 < V < \pi \qquad \textbf{(50)}$$
we have only one discrete (TE) mode of the waveguide and this mode is symmetric in x. When this happens, we refer to the waveguide as a *single-mode waveguide*. In Example 28.1, the waveguide will be single-mode for $\lambda_0 > 0.955\ \mu m$; this wavelength (for which V becomes equal to π) is referred to as the cutoff wavelength.[7]
2. From Fig. 28.2, it is easy to see that if $\pi/2 < V/2 < \pi$ (or $\pi < V < 2\pi$), we will have one symmetric and one antisymmetric TE mode. In general, if
$$2m\pi < V < (2m+1)\pi \qquad \textbf{(51)}$$
we will have $m + 1$ symmetric modes and m antisymmetric modes, and if
$$(2m+1)\pi < V < (2m+2)\pi \qquad \textbf{(52)}$$
we will have $m + 1$ symmetric modes and m + 1 antisymmetric modes where $m = 0, 1, 2, \ldots$. Thus, the total number of modes will be the integer closest to (and greater than) V/π.
3. When the waveguide supports many modes (i.e., when $V >> 1$), the points of intersection (in Fig. 28.2) will be very close to $\xi = \pi/2, 3\pi/2, \ldots$; thus, the propagation constants corresponding to the first few modes can be approximately determined by the following equation:
$$\xi = \xi_m = \sqrt{k_0^2 n_1^2 - \beta_m^2}\ \frac{d}{2} \approx (m+1)\frac{\pi}{2} \qquad V >> 1 \qquad \textbf{(53)}$$
where
$m = 0, 2, 4, \ldots$ correspond to symmetric modes
and
$m = 1, 3, 5, \ldots$ correspond to antisymmetric modes

28.4 TM MODES OF A SYMMETRIC STEP INDEX PLANAR WAVEGUIDE

In the above discussion we considered the TE modes of the waveguide. A very similar analysis can also be performed for the TM modes. In App. D we show that for TM modes $H_y(x)$ satisfies the following equation:

$$n^2(x)\frac{d}{dx}\left[\frac{1}{n^2(x)}\frac{dH_y}{dx}\right] + \left[k_0^2 n^2(x) - \beta^2\right]H_y(x) = 0 \qquad \textbf{(54)}$$

For a step index waveguide [see Eq. (1)], $n^2(x)$ will be constant in each region, and therefore $H_y(x)$ will also satisfy Eqs. (15) and (16) in the regions $|x| < d/2$ and $|x| > d/2$, respectively. Now $H_y(x)$ is a tangential component, and hence it will be continuous at the core-cladding interface. Further, since

$$E_z = \frac{1}{i\omega\varepsilon_0 n^2(x)}\frac{dH_y}{dx} \qquad \textbf{(55)}$$

[7] Actually for $V < \pi$, the waveguide will support one TE and one TM mode (see Sec. 28.4), and when n_1 has a value very close to n_2, the two modes will have very nearly the same propagation constants.

(see App. E) and since $E_z(x)$ is a tangential component, the continuity conditions are now

$$H_y \text{ and } \frac{1}{n^2}\frac{dH_y}{dx} \text{ continuous at } x = \pm d/2 \qquad \textbf{(56)}$$

If we incorporate these continuity conditions and use the same procedure as in Sec. 28.2, we get the following transcendental equations:

$$\xi \tan\xi = \left(\frac{n_1}{n_2}\right)^2 \sqrt{\left(\frac{V}{2}\right)^2 - \xi^2}$$

for symmetric TM modes **(57)**

A similar derivation gives us

$$-\xi \cot\xi = \left(\frac{n_1}{n_2}\right)^2 \sqrt{\left(\frac{V}{2}\right)^2 - \xi^2}$$

for antisymmetric TM modes **(58)**

where ξ and V have been defined earlier. One can again use a graphical method to determine the discrete propagation constants for TM modes. In terms of the parameters b and V, we have

$$\left(\frac{1}{2}V\sqrt{1-b}\right)\tan\left(\frac{1}{2}V\sqrt{1-b}\right) = \left(\frac{n_1}{n_2}\right)^2 \frac{1}{2}V\sqrt{b}$$

for symmetric TM modes **(59)**

$$-\left(\frac{1}{2}V\sqrt{1-b}\right)\cot\left(\frac{1}{2}V\sqrt{1-b}\right) = \left(\frac{n_1}{n_2}\right)^2 \frac{1}{2}V\sqrt{b}$$

for antisymmetric TM modes **(60)**

One now requires the value of $(n_1/n_2)^2$ to obtain the b–V curves (see Fig. 28.3). Obviously if n_1 has a value very close to n_2, then $(n_1/n_2)^2$ is very close to 1 and the propagation constants for TM modes will be very close to the propagation constants for TE modes—this is the weakly guiding approximation.

28.5 TE MODES OF A PARABOLIC INDEX PLANAR WAVEGUIDE

As another example, we consider parabolic variation of refractive index (see Sec. 3.4.1)

$$n^2(x) = n_1^2 - \gamma^2 x^2 \qquad \textbf{(61)}$$

Thus Eq. (5) takes the form

$$\frac{d^2E_y}{dx^2} + \left[\left(k_0^2 n_1^2 - \beta^2\right) - k_0^2\gamma^2 x^2\right] E_y = 0 \qquad \textbf{(62)}$$

which can be written in the form

$$\frac{d^2E_y}{d\xi^2} + \left(\Lambda - \xi^2\right) E_y = 0 \qquad \textbf{(63)}$$

where $\xi = \alpha x$ and we have chosen $\alpha = \sqrt{k_0\gamma}$. Further,

$$\Lambda = \frac{k_0^2 n_1^2 - \beta^2}{\alpha^2} = \frac{k_0^2 n_1^2 - \beta^2}{k_0\gamma} \qquad \textbf{(64)}$$

For the wave function not to blow up at $x = \pm\infty$ (which represents the boundary condition), Λ must be equal to an odd integer (see App. F); i.e.,

$$\Lambda = \frac{k_0^2 n_1^2 - \beta^2}{k_0\gamma} = 2m + 1 \qquad m = 0, 1, 2, 3, \ldots \qquad \textbf{(65)}$$

For those who are familiar with quantum mechanics, Eq. (63) is identical to the one obtained while solving the one-dimensional Schrödinger equation for the linear harmonic oscillator problem (see, e.g., Refs. 5 and 6). Equation (65) gives the following expression for the discrete propagation constants:

$$\beta = \beta_m = k_0 n_1 \left[1 - \frac{(2m+1)\gamma}{k_0 n_1^2}\right]^{1/2} \qquad m = 0, 1, 2, 3, \ldots \qquad \textbf{(66)}$$

The corresponding modal patterns are Hermite Gauss functions (see Appendix F):

$$E_y(x) = NH_m(\xi)\exp\left(-\frac{1}{2}\xi^2\right) \qquad m = 0, 1, 2, 3, \ldots \qquad \textbf{(67)}$$

where N is a constant and $H_m(\xi)$ are the Hermite polynomials:

$$H_0(\xi) = 1 \quad H_1(\xi) = 2\xi \quad H_2(\xi) = 4\xi^2 - 2$$
$$H_3(\xi) = 8\xi - 12\xi^3 \ldots \qquad \textbf{(68)}$$

Notice that the modes corresponding to even values of m are symmetric in x, and modes corresponding to odd values of m are antisymmetric in x. This is so because the refractive index variation $n^2(x)$ is symmetric in x. Equations (66) and (67) represent rigorously correct propagation constants and field profiles (corresponding to TE modes) in an infinitely extended parabolic index medium; of course the refractive index distribution is itself unrealistic. A more realistic distribution is given by (see Sec. 3.4.1)

$$\begin{aligned} n^2(x) &= n_1^2\left[1 - 2\Delta\left(\frac{x}{a}\right)^2\right] && |x| < a \quad \text{core} \\ &= n_2^2 = n_1^2(1 - 2\Delta) && |x| > a \quad \text{cladding} \end{aligned} \qquad \textbf{(69)}$$

The region $|x| < a$ is known as the core of the waveguide, and the region $|x| > a$ is referred to as the cladding. Thus

$$\gamma = \frac{n_1\sqrt{2\Delta}}{a} \qquad \textbf{(70)}$$

The waveguide parameter is given by

$$V = k_0 a\sqrt{n_1^2 - n_2^2} = k_0 a n_1\sqrt{2\Delta} \qquad \textbf{(71)}$$

In a typical parabolic index medium,

$$n_1 \approx 1.5 \qquad \Delta \approx 0.01 \qquad a \approx 20\,\mu\text{m} \qquad \textbf{(72)}$$

giving $n_2 \approx 1.485$ and $\gamma \approx 1.0607 \times 10^4$ m^{-1}. For discrete guided modes we must have

$$n_2^2 < \frac{\beta^2}{k_0^2} < n_1^2 \tag{73}$$

and therefore the maximum value of m will correspond to $\beta = \beta_m = \beta_{\min} = k_0 n_2$. Indeed, when the waveguide supports a very large number of modes, the low-order modes are accurately given by Eq. (66). Now when $\gamma/k_0 n_1^2 \ll 1$ and for not too large values of m, we may carry out a binomial expansion in Eq. (66) to obtain

$$\beta = \beta_m \approx k_0 n_1 - \left(m + \frac{1}{2}\right)\frac{\gamma}{n_1}$$

$$\approx \frac{\omega}{c} n_1 - \left(m + \frac{1}{2}\right)\frac{\gamma}{n_1} \qquad m = 0, 1, 2, 3, \ldots \tag{74}$$

Thus the group velocity v_g of the mode will be given by [see Eqs. (6) and (41) of Chap. 10]:

$$\frac{1}{v_g} = \frac{d\beta}{d\omega} \approx \frac{n_1}{c} \tag{75}$$

independent of the mode number! Thus, in this approximation, all modes travel with the same group velocity. Indeed, using ray optics, we showed in Sec. 3.4 that *all* rays take approximately the same time to propagate through a certain distance of a parabolic index waveguide. It is for this reason that parabolic index waveguides are often used in fiber-optic communication systems.

For a cladded waveguide, if we assume the validity of Eq. (66), we can easily calculate the total number of modes. Since the minimum value of β is $k_0 n_2$, we will have

$$\frac{k_0^2\left(n_1^2 - n_2^2\right)}{k_0 \gamma} = (2m_{\max} + 1) \tag{76}$$

where $m_{\max}$ represents the maximum value of m. Thus the total number of modes is given by

$$N \approx 2m_{\max} \approx V \tag{77}$$

where we have used Eqs. (70) and (71) and the fact that there would be an equal number of TM modes. For the parameters given by Eq. (72) we obtain $N \approx 27$. Some exact solutions for TM modes in graded index slab waveguides are given in Ref. 8.

28.6 WAVEGUIDE THEORY AND QUANTUM MECHANICS

In Sec. 28.3, we showed that for a given waveguide, if λ_0 is made to go to 0, the value of V becomes very large and the waveguide will support a very large number of modes. In this limit, we can assume all values of θ to be allowed, and it will be quite appropriate to use ray optics to study the propagation characteristics of the waveguide.

In this section, we will show that for a given quantum well structure, if $\hbar$ is made to go to 0, the quantum well structure will have a very large number of bound states, and in this limit, we can assume all values of energy to be allowed and it will be quite appropriate to use classical mechanics. Further, the one-dimensional Schrödinger equation is very similar to the wave equation for TE modes; the former leads to the bound states of a quantum mechanical problem, and the latter leads to guided modes of a waveguide problem. Obviously, the methodology of solving either of the equations is the same. Indeed the modal analysis of the step index planar waveguide is almost identical to the procedure used for solving the one-dimensional Schrödinger equation corresponding to the symmetric potential well. Similarly, the modal analysis of the parabolic index planar waveguide is almost identical to the linear harmonic oscillator problem in quantum mechanics (see, e.g., Refs. 5 and 6). Thus, it is often easier to understand a concept in quantum mechanics through fiber optics and vice versa. Further, we can say that

> The relationship between geometric and wave optics is very similar to the relation between classical and quantum mechanics. In the limit of $\lambda_0 \to 0$, wave optics goes over to ray optics and in the limit of $\hbar \to 0$, quantum mechanics goes over to classical mechanics.

Now, for a particle of mass μ the one-dimensional Schrödinger equation is given by

$$\frac{d^2\psi}{dx^2} + \frac{2\mu}{\hbar^2}[E - V(x)]\psi(x) = 0 \tag{78}$$

We consider a potential energy function given by [cf. Eq. (1)]

$$V(x) = \begin{cases} 0 & |x| < \dfrac{d}{2} \\ V_0 & |x| > \dfrac{d}{2} \end{cases} \tag{79}$$

Thus the Schrödinger equation can be written in the form

$$\frac{d^2\psi}{dx^2} + \kappa^2 \psi(x) = 0 \qquad |x| < \frac{d}{2} \tag{80}$$

$$\frac{d^2\psi}{dx^2} - \gamma^2 \psi(x) = 0 \qquad |x| > \frac{d}{2} \tag{81}$$

where

$$\kappa^2 = \frac{2\mu E}{\hbar^2} \tag{82}$$

and

$$\gamma^2 = \frac{2\mu}{\hbar^2}(V_0 - E) \tag{83}$$

As in the waveguide problem, we will solve Eqs. (80) and (81) subject to the appropriate boundary and continuity conditions. The continuity conditions are

$$\psi \text{ and } \frac{d\psi}{dx} \text{ continuous at } x = \pm\frac{d}{2} \tag{84}$$

Now, for a bound state, the wave function is mainly confined to the region $|x| < \frac{d}{2}$ and hence its field should decay in the region $|x| < \frac{d}{2}$, so that there is a large probability of finding the particle inside the well. Thus, we must have

$$E < V_0$$

When $E > V_0$, the solutions are oscillatory in the region $|x| > d/2$ and they correspond to what are known as scattering states. Furthermore, E cannot be less than the minimum value of $V(x)$ (in this case the minimum value is zero); otherwise, the boundary conditions cannot be satisfied at $x = \pm d/2$. Thus, for bound states we must have

$$0 < E < V_0 \qquad \text{bound states} \tag{85}$$

Now, when the potential energy variation is symmetric about $x = 0$, that is, when

$$V(-x) = V(x) \tag{86}$$

the solutions are either symmetric or antisymmetric functions of x (see Prob. 28.8; see also pp. 126–127 of Ref. 2); thus we must have

$$\psi(-x) = \psi(x) \qquad \text{symmetric states} \tag{87}$$

$$\psi(-x) = -\psi(x) \qquad \text{antisymmetric states} \tag{88}$$

Carrying out an analysis identical to that in Sec. 28.2, we find that the wave function for symmetric states is given by Eq. (24) and the wave function for antisymmetric states is given by Eq. (33). Continuity of ψ and $d\psi/dx$ at $x = \pm d/2$ gives the following equations:

$$\xi \tan \xi = \sqrt{\alpha^2 - \xi^2} \qquad \text{for symmetric states} \tag{89}$$

$$-\xi \cot \xi = \sqrt{\alpha^2 - \xi^2} \qquad \text{for antisymmetric states} \tag{90}$$

where

$$\alpha \equiv \sqrt{\frac{2\mu V_0 d^2}{\hbar^2}} \tag{91}$$

For a given value of α, the solutions of Eqs. (89) and (90) will give the bound states for the potential well problem given by Eq. (79). Obviously, for $\alpha < \pi/2$ we will have only bound state—similar to the condition we had for a single-mode waveguide. For given values of V_0, μ, and d, as $\hbar \to 0$, the value of α will become large and we will have a continuum of states, implying that all energy levels are possible. Thus in the limit of $\hbar \to 0$, we have the results of classical mechanics.

Now, when $E < V_0$, there is a finite probability of finding the particle in the region $x > d/2$; this region is forbidden in classical mechanics because the total energy E is less than the potential energy ($= V_0$) and therefore the kinetic energy will be negative. Similarly, in the waveguide problem the ray undergoes total internal reflection at the core-cladding interface, and a geometrical ray is not possible in the rarer medium; on the other hand, while solving Eq. (18), we had the evanescent wave in the region $|x| > d/2$. Indeed when a light beam is incident on a layer of lower refractive index at an angle of incidence greater than the critical angle, then a part of the beam "tunnels through" the rarer medium and appears in the third medium, as shown in Fig. 28.7(a); this phenomenon is known as frustrated total internal reflection (usually abbreviated as FTIR) and is a consequence of the evanescent wave present in the rarer medium. Such tunneling is not allowed in geometrical optics because the beam will undergo total internal reflection at the first interface. An almost identical situation arises in quantum mechanics when a particle of energy E ($< V_0$), incident on a potential barrier (of height V_0), has a finite probability of tunneling through as shown in Fig. 28.7(b). Such tunneling is not possible in classical mechanics, and as shown in almost all textbooks in quantum mechanics, the tunneling probability will go to zero when $\hbar \to 0$.

Finally we consider the linear harmonic oscillator problem in quantum mechanics where the potential energy function is given by

$$V(x) = \tfrac{1}{2}\mu\omega^2 x^2 \tag{92}$$

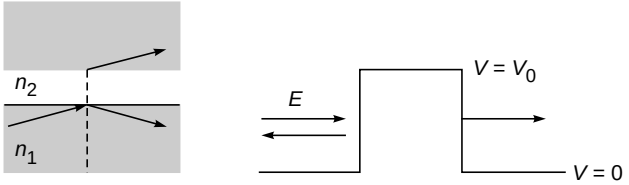


Fig. 28.7 (a) When a light beam is incident on a layer of lower refractive index at an angle of incidence greater than the critical angle, a part of the beam tunnels through to the third medium; this phenomenon is known as frustrated total internal reflection (usually abbreviated as FTIR) and is a consequence of the evanescent wave present in the rarer medium. (b) A particle of energy E ($< V_0$), incident on a potential barrier (of height V_0), has a finite probability of tunneling through the potential barrier.

and the Schrödinger equation [Eq. (78)] becomes

$$\frac{d^2\psi}{d\xi^2} + (\Lambda - \xi^2)\psi = 0 \qquad \textbf{(93)}$$

where $\xi = \alpha x$ and we have chosen

$$\alpha = \sqrt{\frac{\mu\omega}{\hbar}} \qquad \text{so that} \qquad \Lambda \equiv \frac{2E}{\hbar\omega} \qquad \textbf{(94)}$$

For the wave function not to blow up at $x = \pm\infty$ (which represents the boundary condition), Λ must be equal to an odd integer (see App. F); i.e.,

$$\Lambda = \frac{2E}{\hbar\omega} = 2m + 1 \qquad m = 0, 1, 2, 3, \ldots \qquad \textbf{(95)}$$

The above equation gives the following expression for the discrete energy eigenvalues:

$$E = E_m = \left(m + \frac{1}{2}\right)\hbar\omega \qquad m = 0, 1, 2, 3, \ldots \qquad \textbf{(96)}$$

The relationship of the quantum mechanical oscillator with classical oscillator is discussed in detail in Ref. 6, and the relationship of ray optics in a parabolic index waveguide with the results obtained in modal theory is discussed in Ref. 2.

Problems

28.1 Consider a symmetric step index waveguide [see Eq. (1)] with $n_1 = 1.50$, $n_2 = 1.46$, and $d = 4$ μm operating at $\lambda_0 = 0.6328$ μm. Calculate the number of TE and TM modes.

28.2 Consider TE modes in a step index planar waveguide with $d = 2.0$ μm, $n_1 = 1.5$, and the value of n_2 chosen such that $\sqrt{n_1^2 - n_2^2} = 1/\pi$. For $\lambda_0 = 1$, 0.8, and 0.66667 μm, calculate (using Table 28.1) the values of b and the corresponding value of β/k_0. Show that the values of β/k_0 lie between n_1 and n_2.

28.3 Consider now a parabolic index waveguide [see Eq. (60)] with $n_1 = 1.50$, $n_2 = 1.46$, and $a = 2$ μm operating again at $\lambda_0 = 0.6328$ μm. Assuming the validity of Eq. (57) and that for discrete guided modes we must have $n_2^2 < \beta^2/k_0^2 < n_1^2$, calculate the maximum value of m and the total number of TE modes.

28.4 Consider a step index symmetric waveguide with $n_1 = 1.50$ and $n_2 = 1.48$ operating at $\lambda_0 = 0.6328$ μm. Calculate the value of d so that $V = 6$. Using Table 28.1, calculate the values of b, the corresponding propagation constants β/k_0, and the angles that the component waves make with the z axis.

[**Ans:** $d = 2.4752$ μm]

28.5 We consider the same waveguide as in Prob. 28.4. At what wavelength will the value of V be equal to 3. Using Table 28.1, calculate the value of b and the corresponding propagation constant β/k_0.

28.6 **(a)** Consider a symmetric step index waveguide [see Eq. (1)] with $n_1 = 1.49$, $n_2 = 1.46$, and $d = 4$ μm operating at $\lambda_0 = 0.6328$ μm. Use Table 28.1 and linear interpolation to calculate the values of β/k_0.

(b) Calculate the corresponding values of θ_m.

[**Ans: (a)** The values of β/k_0 are 1.4885, 1.4839, 1.4765, and 1.4668;
(b) $\theta_1 \approx 2.6^\circ$, $\theta_2 \approx 5.2^\circ$, $\theta_3 \approx 7.7^\circ$, $\theta_4 \approx 10.1^\circ$]

28.7 **(a)** Consider a step index symmetric waveguide with $n_1 = 1.503$, $n_2 = 1.500$, and $d = 4$ μm. For $\lambda_0 = 1$ μm, calculate the value of V and use linear interpolation of the numbers given in Table 28.1 to calculate the value of β/k_0.

(b) If the operating wavelength is changed to 0.5 μm, show that $V = 4.771$, and by linear interpolation of the numbers given in Table 28.1, calculate the discrete values of β/k_0 and the corresponding angles that the waves make with the z axis.

[**Ans: (a)** $\frac{\beta}{k_0} \approx 1.5016$; **(b)** $\frac{\beta}{k_0} \approx 1.5024$ and 1.5007]

28.8 In Eq. (5), make the transformation $x \to -x$, and assuming $n^2(-x) = n^2(x)$, show that $E_y(-x)$ satisfies the same equation as $E_y(x)$; hence we must have $E_y(-x) = \lambda E_y(x)$. Make the transformation $x \to -x$ again to prove that the solutions are either symmetric or antisymmetric functions of x [i.e., prove Eqs. (20) and (21)].

REFERENCES AND SUGGESTED READINGS

1. A. W. Snyder and J. D. Love, *Optical Waveguide Theory*, Chapman and Hall, London, 1983.
2. A. Ghatak and K. Thyagarajan, *Introduction to Fiber Optics*, Cambridge University Press, Cambridge, United Kingdom, 1998; reprinted in India by Foundation Books, New Delhi.
3. D. K. Mynbaev and L. L. Scheiner, *Fiber-Optic Communications Technology*, Prentice-Hall, Englewood Cliffs, N. J., 2001.
4. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, Wiley, New York, 1991.
5. B. H. Bransen and C. J. Joachain, *Introduction to Quantum Mechanics*, Longman Group, United Kingdom, 1989.
6. A. Ghatak and S. Lokanathan, *Quantum Mechanics: Theory and Applications*, 5th ed., Macmillan India, New Delhi, 2004. reprinted by Kluwer Academic Publishers, Dordrecht, 2004.
7. A. Ghatak, I. C. Goyal, and R. Varshney, *FIBER OPTICA : A Software for Characterizing Fiber and Integrated-Optic Waveguides*, Viva Books, New Delhi, 1999.
8. J. D. Love, and A. Ghatak, *Exact Solutions for TM modes in Graded-Index Slab Waveguides*, IEEE J. Quant. Electr. **QE15**, pp. 14–16, 1979.

Chapter Twenty-Nine

OPTICAL WAVEGUIDES III: SINGLE-MODE FIBERS

A new era is dawning in the West, the era of light. Under city streets and beneath oceans, in commercial skyscrapers . . . , a host of new technologies based on lasers, ultrapure glass fibers and exotic new materials are challenging the wonders of conventional electronic gadgetry . . . . With growing speed, the new technology promises to turn the electronic age into the age of optics, in which gadgetry built around beams of light becomes virtually indispensable.

—*Time Magazine*, October 6, 1986

29.1 INTRODUCTION

At the heart of an optical communication system is the optical fiber that acts as the transmission channel carrying the light beam loaded with information. According to ray optics, the light beam gets guided through the optical fiber due to the phenomenon of total internal reflection (often abbreviated as TIR); we discussed this in Chap. 27. However, for a single-mode fiber (which is now extensively used in optical communication systems), the core diameter is very small (few micrometers) and ray optics does not remain valid, and one has to use Maxwell's electromagnetic theory to study the propagation characteristics of the (single-mode) fiber. In Chap. 28 we carried out modal analysis of planar waveguides which enabled us to understand the concept of modes. In this chapter we carry out modal analysis of the step index fiber to help us in the design of a fiber-optic communication system. In a single-mode fiber, there is no intermodal dispersion and we will show that by appropriately tailoring the transverse refractive index profile, the total dispersion can be made extremely small. This would lead to very large information carrying capacity systems.

29.2 BASIC EQUATIONS

The simplest refractive index variation is that of a step index fiber which is characterized by the following refractive index distribution (see Fig. 29.1):

$$n(r) = \begin{matrix} n_1 & 0 < r < a & \text{core} \\ n_2 & r > a & \text{cladding} \end{matrix} \qquad \textbf{(1)}$$

where we are using the cylindrical system of coordinates (r, ϕ, z). In actual fibers

$$\frac{n_1 - n_2}{n_2} \leq 0.01 \qquad \textbf{(2)}$$

and this allows use of the so-called scalar wave approximation (also known as the weakly guiding approximation[1]). In

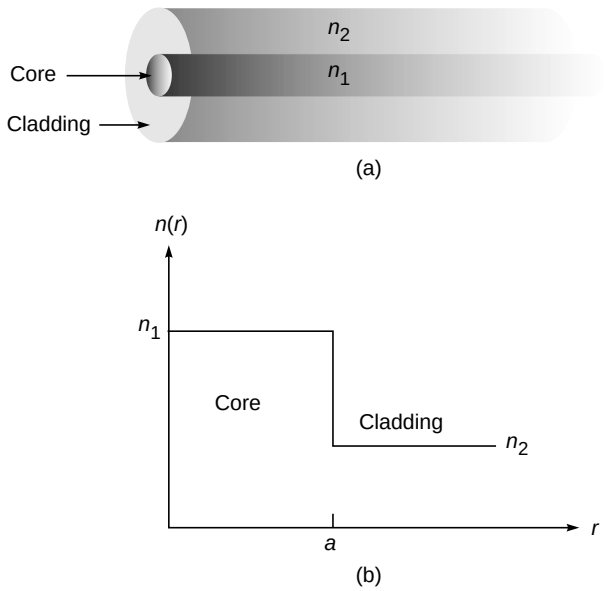


Fig. 29.1 (a) A step index fiber is a cylindrical structure in which the refractive index is n_1 for $0 < r < a$ and n_2 for $r > a$. (b) The refractive index variation of a step index fiber.

[1] For more details about the weakly guiding approximation see, e.g., Refs. 1 and 2.

this approximation, the modes are assumed to be nearly transverse and can have an arbitrary state of polarization. Thus, the two independent sets of modes can be assumed to be *x*-polarized and *y*-polarized, and in the weakly guiding approximation they have the same propagation constants. These are usually referred to as LP modes; LP stands for linearly polarized. We may compare this with the discussion in Sec. 28.5 where we mentioned that when $n_1 \approx n_2$, the modes are nearly transverse and the propagation constants of the TE and TM modes are almost equal. In the weakly guiding approximation, the transverse component of the electric field (E_x or E_y) satisfies the scalar wave equation [see Eq. (51) of Chap. 23]:

$$\nabla^2\Psi = \varepsilon_0\mu_0 n^2 \frac{\partial^2\Psi}{\partial t^2} = \frac{n^2}{c^2}\frac{\partial^2\Psi}{\partial t^2} \tag{3}$$

where $c\,(= 1/\sqrt{\varepsilon_0\mu_0}) \approx 3 \times 10^8$ m s^{-1} is the speed of light in free space. In most practical fibers n^2 depends only on the cylindrical coordinate r, and therefore it is convenient to use the cylindrical system of coordinates (r, ϕ, z,) and write the solution of Eq. (3) in the form

$$\Psi(r, \phi, z, t) = \psi(r, \phi)e^{i(\omega t - \beta z)} \tag{4}$$

where ω is the angular frequency and β is known as the propagation constant. The above equation defines the modes of the system. Since $\psi(r, \phi)$ depends only on the transverse coordinates r and ϕ,

> The modes represent transverse field configurations that do not change as they propagate through the optical fiber except for a phase change.

In the cylindrical system of coordinates (r, ϕ, z) we have

$$\nabla^2\Psi = \frac{\partial^2\Psi}{\partial r^2} + \frac{1}{r}\frac{\partial\Psi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\Psi}{\partial\phi^2} + \frac{\partial^2\Psi}{\partial z^2} \tag{5}$$

Now, from Eq. (4) it readily follows that

$$\frac{\partial^2\Psi}{\partial t^2} = -\omega^2\Psi = -\omega^2\psi(r, \phi)e^{i(\omega t - \beta z)} \tag{6}$$

and

$$\frac{\partial^2\Psi}{\partial z^2} = -\beta^2\Psi = -\beta^2\psi(r, \phi)e^{i(\omega t - \beta z)} \tag{7}$$

Substituting Eq. (4) in Eq. (3) and using Eqs. (5) to (7), we obtain

$$\frac{\partial^2\psi}{\partial r^2} + \frac{1}{r}\frac{\partial\psi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\psi}{\partial\phi^2} + [k_0^2 n^2(r) - \beta^2]\psi = 0 \tag{8}$$

where

$$k_0 = \frac{\omega}{c} = \frac{2\pi}{\lambda_0}$$

is the free space wave number. Because the medium has cylindrical symmetry, i.e., n^2 depends only on the cylindrical coordinate r, we can solve Eq. (8) by the method of separation of variables:

$$\psi(r, \phi) = R(r)\,\Phi(\phi)$$

On substituting and dividing by $\psi(r, \phi)/r^2$, we obtain

$$\frac{r^2}{R}\left(\frac{d^2R}{dr^2} + \frac{1}{r}\frac{dR}{dr}\right) + r^2\left[n^2(r)k_0^2 - \beta^2\right] = -\frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2} = +l^2 \tag{9}$$

Thus the variables have separated out, and we have set each side equal to a constant ($= l^2$). Solving the equation depending only on ϕ, we find that the ϕ dependence will be of the form cos $l\phi$ or sin $l\phi$, and for the function to be single-valued [i.e., for $\Phi(\phi + 2\pi) = \Phi(\phi)$] we must have

$$l = 0, 1, 2, \ldots$$

Negative values of l correspond to the same field distribution. Thus the complete transverse field is given by

$$\Psi(r, \phi, z, t) = R(r)e^{i(\omega t - \beta z)}\begin{Bmatrix}\cos l\phi \\ \sin l\phi\end{Bmatrix} \quad l = 0, 1, 2, \ldots \tag{10}$$

where $R(r)$ satisfied the radial part of the equation

$$r^2\frac{d^2R}{dr^2} + r\frac{dR}{dr} + \left\{\left[k_0^2 n^2(r) - \beta^2\right]r^2 - l^2\right\}R = 0 \tag{11}$$

Equation (11) is an eigenvalue equation with β^2 representing the eigenvalue. By applying the appropriate boundary conditions, we will show that β^2 can have a set of discrete values (corresponding to guided modes of the waveguide) and also a continuum of values corresponding to radiation modes of the waveguide.

Since for each value of l there can be two independent states of polarization, modes with $l \geq 1$ are fourfold degenerate (corresponding to two orthogonal polarization states and to the ϕ dependence being cos $l\phi$ or sin $l\phi$). Modes with $l = 0$ are ϕ independent and have twofold degeneracy.[2] We cannot set the right-hand side of Eq. (9) equal to a negative constant, because then the ϕ dependence of the field will not be single-valued. In the next section we give the solution of

[2] The word *degeneracy* means that for the same value of the propagation constant there is more than one field profile. For $l = 0$ we will have two independent states of polarization; thus the mode is said to be twofold degenerate. On the other hand, for $l = 1, 2, 3, \ldots$ the mode will be fourfold degenerate because (for the same value of β^2) we will have two field profiles, one proportional to cos $l\phi$ and the other to sin $l\phi$, and for each field profile, we will again have two independent states of polarization.

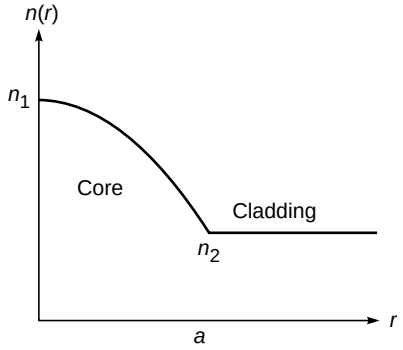


Fig. 29.2 A cylindrically symmetric refractive index profile having a refractive index that decreases monotonically from a value n_1 on the axis to a constant value n_2 beyond the core-cladding interface $r = a$.

Eq. (11) for a step index profile. However, for an arbitrary cylindrically symmetric profile having a refractive index that decreases monotonically from a value n_1 on the axis to a constant value n_2 beyond the core-cladding interface $r = a$ (see Fig. 29.2), we can make the general observation that the solutions of Eq. (11) can be divided into two distinct classes (compare with the discussions in Sec. 28.2). The first class of solutions corresponds to

$$n_2^2 < \frac{\beta^2}{k_0^2} < n_1^2 \qquad \text{guided modes} \qquad \textbf{(12)}$$

For β^2 lying in the above range, the fields $R(r)$ are oscillatory in the core and decay in the cladding and β^2 assumes only discrete values; these are known as the *guided modes* of the waveguide. For a given value of l, there will be a finite number of guided modes, these are designated as LP_{lm} modes ($m = 1, 2, 3, \ldots$). The second class of solutions corresponds to

$$\beta^2 < k_0^2 n_2^2 \qquad \text{radiation modes} \qquad \textbf{(13)}$$

For such β values, the fields are oscillatory even in the cladding and β can assume a continuum of values. These are known as the *radiation modes*. For more details about radiation modes and also excitation of leaky modes, see, e.g., Refs. 1 and 3.

29.3 GUIDED MODES OF A STEP INDEX FIBER

In this section, we obtain the modal fields and the corresponding propagation constants for guided modes in a step index fiber for which the refractive index variation is given by Eq. (1).

For such a fiber, for guided modes (for which $n_2^2 < \beta^2/k_0^2 < n_1^2$), Eq. (11) can be written in the form

$$r^2 \frac{d^2R}{dr^2} + r \frac{dR}{dr} + \left(U^2 \frac{r^2}{a^2} - l^2 \right) R = 0 \qquad 0 < r < a \qquad \textbf{(14)}$$

and

$$r^2 \frac{d^2R}{dr^2} + r \frac{dR}{dr} - \left(W^2 \frac{r^2}{a^2} + l^2 \right) R = 0 \qquad r > a \qquad \textbf{(15)}$$

where

$$U \equiv a\sqrt{k_0^2 n_1^2 - \beta^2} \qquad \textbf{(16)}$$

and

$$W \equiv a\sqrt{\beta^2 - k_0^2 n_2^2} \qquad \textbf{(17)}$$

Because of Eq. (12), both U and W are real. The normalized waveguide parameter V is defined by

$$V = \sqrt{U^2 + W^2} = k_0 a \sqrt{n_1^2 - n_2^2} \qquad \textbf{(18)}$$

In terms of the wavelength

$$V = \frac{2\pi}{\lambda_0} a \sqrt{n_1^2 - n_2^2} \qquad \textbf{(19)}$$

The waveguide parameter V is an extremely important quantity characterizing an optical fiber. It is convenient to define the normalized propagation constant

$$b = \frac{\beta^2/k_0^2 - n_2^2}{n_1^2 - n_2^2} = \frac{W^2}{V^2} \qquad \textbf{(20)}$$

Thus

$$W = V\sqrt{b} \qquad \textbf{(21)}$$

and

$$U = V\sqrt{1 - b} \qquad \textbf{(22)}$$

From Eq. (12) we find that for guided modes $0 < b < 1$. The two independent solutions of Eq. (14) are $J_l(Ur/a)$ and $Y_l(Ur/a)$ (see, e.g., Refs. 4 to 6]; however, the solution $Y_l(Ur/a)$ has to be rejected since it diverges as $r \to 0$. The solutions of Eq. (15) are the modified Bessel functions $K_l(Wr/a)$ and $I_l(Wr/a)$; the solution $I_l(Wr/a)$ has to be rejected since it diverges as $r \to \infty$. Thus, for guided modes, the transverse dependence of the modal field is given by

$$\psi(r, \phi) = \begin{cases} \dfrac{A}{J_l(U)} J_l\left(\dfrac{Ur}{a}\right) \begin{bmatrix} \cos l\phi \\ \sin l\phi \end{bmatrix} & r < a \\ \dfrac{A}{K_l(W)} K_l\left(\dfrac{Wr}{a}\right) \begin{bmatrix} \cos l\phi \\ \sin l\phi \end{bmatrix} & r > a \end{cases} \qquad \textbf{(23)}$$

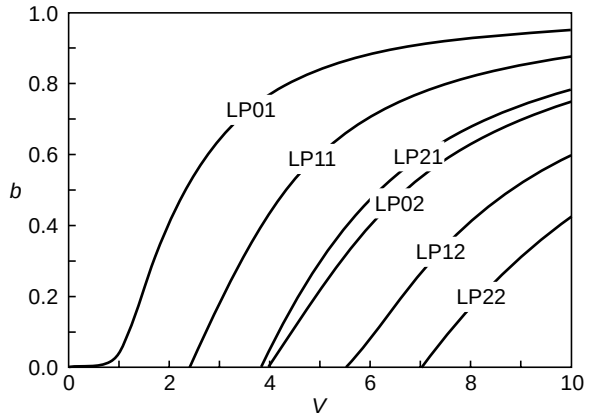


Fig. 29.3 Variation of the normalized propagation constant b with normalized waveguide parameter V corresponding to a few lower-order modes; calculations courtesy of Dr. Sunil Khijwania.

Table 29.1 Values of b, $(bV)'$, and $V(bV)''$ versus V for a step index fiber; the values in the second, fourth, and fifth columns are generated by solving Eq. (25) for a step index fiber, using the software given in Refs. 7 and 8.

V	b	b [using Eq. (30)]	$\frac{d}{dV}(bV)$	$V(bV)''$
1.5	0.229248	0.229249	0.849	1.063
1.6	0.270063	0.270712	0.913	0.919
1.7	0.309467	0.310157	0.965	0.785
1.8	0.347068	0.347471	1.006	0.664
1.9	0.382660	0.382653	1.039	0.556
2.0	0.416163	0.415767	1.065	0.462
2.1	0.447581	0.446911	1.086	0.380
2.2	0.476969	0.476200	1.102	0.309
2.3	0.504416	0.503754	1.114	0.248
2.4	0.530026	0.529693	1.124	0.195
2.5	0.553915	0.554131		

where A is a constant and we have assumed the continuity of ψ at the core-cladding interface ($r = a$). Continuity of $\partial\psi/\partial r$ at $r = a$ and use of identities involving Bessel functions [see, e.g., Ref. 3] give the following transcendental equations which determine the allowed discrete values of the normalized propagation constant b of the guided LP_{lm} modes:

$$V\sqrt{1-b}\,\frac{J_{l-1}\left[V\sqrt{1-b}\right]}{J_l\left[V\sqrt{1-b}\right]} = -V\sqrt{b}\,\frac{K_{l-1}\left[V\sqrt{b}\right]}{K_l\left[V\sqrt{b}\right]};\ l \geq 1 \qquad \textbf{(24)}$$

and

$$V\sqrt{1-b}\,\frac{J_1\left[V\sqrt{1-b}\right]}{J_0\left[V\sqrt{1-b}\right]} = V\sqrt{b}\,\frac{K_1\left[V\sqrt{b}\right]}{K_0\left[V\sqrt{b}\right]};\ l = 0 \qquad \textbf{(25)}$$

The solution of the above transcendental equations will give universal curves describing the dependence of b (and therefore of U and W) on V. For a given value of l, there will be a finite number of solutions, and the mth solution ($m = 1, 2, 3, \ldots$) is referred to as the LP_{lm} mode. The variation of b with V forms a set of universal curves, which are plotted in Fig. 29.3. Table 29.1 gives the numerical values of b (corresponding to the LP_{0m} mode) for values of V lying between 1.0 and 2.5.

29.3.1 Cutoff Frequencies

From Fig. 29.3 we see that the value of b decreases as we decrease the value of V. For every mode, there is a value of V when b becomes zero (i.e., when β/k_0 becomes equal to n_2), and the mode ceases to be a guided mode. The value of V for which b becomes zero is known as the *cutoff frequency* of the mode. Now, for a given step index fiber, the value of V decreases as we increase the wavelength [see Eq. (19)], and the value of the wavelength at which b becomes zero is known as the *cutoff wavelength* for that mode.

We can see from Eq. (25) that the cutoff frequencies of the LP_{0m} modes will occur at the zeros of $J_1(V)$, i.e., when $V = 0$ (LP_{01}), 3.8317 (LP_{02}), 7.0156 (LP_{03}), 10.1735 (LP_{04}), ...

Similarly, we can see from Eq. (24) that

- Cutoff frequencies of the LP_{1m} modes will occur at the zeros of $J_0(V)$, i.e., when $V = 2.4048$ (LP_{11}), 5.5201 (LP_{12}), 8.6537 (LP_{13}), 11.7915 (LP_{14}), ...
- Cutoff frequencies of the LP_{2m} modes occur at the zeros of $J_1(V)$ (excluding the value $V = 0$), i.e., when $V = 3.8317$ (LP_{21}), 7.0156 (LP_{22}), 10.1735 (LP_{23}), ...

For $l \geq 1$, cutoff frequencies of the LP_{lm} modes will occur at the zeros of $J_{l-1}(V)$ (excluding the value $V = 0$); thus[3]

- Cutoff frequencies of the LP_{3m} modes occur when $V = 5.1356$ (LP_{31}), 8.4172 (LP_{32}), 11.6198 (LP_{33}), ...
- Cutoff frequencies of the LP_{4m} modes occur when $V = 6.3802$ (LP_{41}), 9.7610 (LP_{42}), 13.015 (LP_{43}), ...
- Cutoff frequencies of the LP_{5m} modes occur when $V = 7.5883$ (LP_{51}), 11.0647 (LP_{52}), ...
- Cutoff frequencies of the LP_{6m} modes occur when $V = 8.7715$ (LP_{61}), 12.3386 (LP_{62}), ...

[3] The values of the zeros of the Bessel functions are taken from Ref. 9.

Thus, as can also be seen from the figure:

- For $0 < V < 2.4048$ we will only have the LP_{01} mode (which is referred to as the fundamental mode); $V = 2.4048$ represents the cutoff of the LP_{11} mode where (for the LP_{11} mode) b becomes 0, i.e., β/k_0 becomes equal to n_2.
- For $2.4048 < V < 3.8317$ we will only have LP_{01} and LP_{11} modes; $V = 3.8317$ represents the cutoff of the LP_{02} and the LP_{21} modes where (for the LP_{02} and the LP_{21} modes) b becomes 0, i.e., β/k_0 becomes equal to n_2.
- For $3.8317 < V < 5.1356$ we will only have LP_{01}, LP_{02}, LP_{11}, and LP_{21} modes; $V = 5.1356$ represents the cutoff of the LP_{31} mode.

Thus at a particular V value the fiber can support only a finite number of modes. Note that each LP_{0m} mode is twofold degenerate; i.e., there are two independent modes with the same value of b, corresponding to two independent states of polarization. Further each LP_{lm} mode ($l > 1$) is fourfold degenerate; i.e., there are four independent modes with the same value of b, corresponding to ϕ dependence of $\cos l\phi$ and $\sin l\phi$ with each mode having two independent states of polarization.

Example 29.1 We consider a step index fiber with $n_1 = 1.5$, $n_2 = 1.49$, and the core radius $a = 3.0$ μm. Thus

$$V = \frac{2\pi}{\lambda_0} a\sqrt{n_1^2 - n_2^2} = \frac{3.2594}{\lambda_0}$$

where λ_0 is measured in μm. Thus

- Cutoff wavelength of the LP_{11} mode will be 1.355 μm.
- Cutoff wavelengths of the LP_{21} and LP_{02} modes will be 0.8506 μm.
- Cutoff wavelength of the LP_{31} mode will be 0.6347 μm.

The LP_{01} mode has no cutoff. Thus for $\lambda_0 > 1.355$ μm, we will only have the LP_{01} mode; and for $0.8506\ \mu m < \lambda_0 < 1.355\ \mu m$, we will have LP_{01} and LP_{11} modes. For $0.6347\ \mu m < \lambda_0 < 0.8506\ \mu m$, we will have LP_{01}, LP_{11}, LP_{21}, and LP_{02} modes.

In the data sheet of a single-mode fiber, the manufacturer always specifies the cutoff wavelength of the fiber—that cutoff wavelength corresponds to that of the LP_{11} mode. In the above example, the cutoff wavelength is 1.355 μm because for all wavelengths *greater* than this, the fiber will be single-mode, supporting only the LP_{01} mode. Thus,

> The minimum wavelength for which we will have only the LP_{01} mode (which, for a step index fiber, will correspond to $V = 2.4045$) is known as the *cutoff wavelength* and is denoted by λ_c.

It is λ_c that is almost always mentioned in the data sheet of a silica fiber (see, e.g., Ref. 11). For a parabolic index fiber [see Eq. (31) of Chap. 27], the cutoff of the LP_{11} mode occurs at $V = 3.518$, and therefore for the same values of n_1, n_2, and a, the cutoff wavelength (for a parabolic index fiber) is smaller than that for a step index fiber.

Example 29.2 We consider a step index fiber with $n_1 = 1.5$, $n_2 = 1.48$, and core radius $a = 6.0$ μm. Assuming the operating wavelength $\lambda_0 = 1.3$ μm, we get $V = 7.0796$. Thus we will have two each of LP_{01}, LP_{02}, and LP_{03} modes; four each of LP_{11}, LP_{12}, LP_{21}, LP_{22}, LP_{31}, and LP_{41} modes; and we will have a total of 30 modes. Now the total number of modes in a highly multimode ($V \geq 10$) step index fiber is approximately given by (see Sec. 27.9.1)

$$N \approx \frac{1}{2}V^2 \qquad \textbf{(26)}$$

For $V = 7.0796$, we get $N \approx 25$. For higher values of V the values given by Eq. (26) will become closer to the exact value (see Prob. 29.2).

29.4 SINGLE-MODE FIBER

The LP_{01} mode (for which $l = 0$ and $m = 1$) is known as the fundamental mode. As mentioned earlier, for a step index fiber when $0 < V < 2.4048$, we will only have the fundamental mode. When this happens, the fiber is referred to as a single-mode fiber which is extensively used in optical fiber communication systems. For the fundamental mode, the actual numerical values of b for various values of V are tabulated in Table 29.1. Thus for a given step index fiber operating at a particular wavelength, we just have to calculate the value of V and then use simple interpolation to calculate the value of b from Table 29.1. From the value of b, we can obtain the corresponding propagation constant by using the following equation [see Eq. (20)]:

$$\frac{\beta}{k_0} = \sqrt{n_2^2 + b\left(n_1^2 - n_2^2\right)} \approx n_2\sqrt{1 + (2\Delta)b} \qquad \textbf{(27)}$$

where in the last step we have assumed $n_1 \approx n_2$.

Example 29.3 We consider a step index fiber with $n_2 = 1.447$, $\Delta = 0.003$, and $a = 4.2$ μm, giving $V = 2.958/\lambda_0$, where λ_0 is measured in μm. Thus for $\lambda_0 > 1.23$ μm, the fiber will be single-mode. The cutoff wavelength λ_c (for which $V = 2.4045$) is 1.23 μm. We assume the operating wavelength $\lambda_0 = 1.479$ μm so that $V = 2.0$ and therefore (from Table 29.1)

$$b \approx 0.4162 \quad \Rightarrow \quad \frac{\beta}{k_0} \approx n_2\sqrt{1 + (2\Delta)b} \approx 1.4488 \qquad \textbf{(28)}$$

$$\Rightarrow \quad \beta \approx 6.1549 \times 10^6\ \text{m}^{-1}$$

Example 29.4 In continuation of Example 29.3, we consider the same step index fiber ($n_2 = 1.447$, $\Delta = 0.003$ and $a = 4.2$ μm) now operating at $\lambda_0 = 1.55$ μm. Thus $V \approx 1.908$ and we again have a single-mode fiber. Using Table 29.1 and linear interpolation, we get

$$b \approx 0.382660 + \frac{0.416163 - 0.382660}{0.1} \times 0.008 \approx 0.38534$$

$$\Rightarrow \quad \frac{\beta}{k_0} \approx n_2\sqrt{1 + (2\Delta)b} \approx 1.4487$$

$$\Rightarrow \quad \beta \approx 5.8725 \times 10^6\ \text{m}^{-1}$$

Example 29.5 For reasons that will be discussed later, the fibers used in fourth-generation optical communication systems (operating at 1.55 μm) have a small value of core radius and a large value of Δ. A typical fiber (operating at $\lambda_0 \approx 1.55$ μm) would have $n_2 = 1.444$, $\Delta = 0.0075$, and $a = 2.3$ μm. Thus at $\lambda_0 = 1.55$ μm

$$V = \frac{2\pi}{1.55} \times 2.3 \times 1.444 \times \sqrt{0.015} \approx 1.649$$

The fiber will be single-mode at 1.55 μm, and

$$b \approx 0.270063 + \frac{0.309467 - 0.270063}{0.1} \times 0.049 \approx 0.28937$$

$$\Rightarrow \qquad \frac{\beta}{k_0} \approx n_2\sqrt{1+(2\Delta)b} \approx 1.44713$$

Further, for the given fiber we may write

$$V = \frac{2.556}{\lambda_0} \qquad \textbf{(29)}$$

and therefore the cutoff wavelength will be

$$\lambda_c = 2.556/2.4045 \approx 1.06 \text{ μm}.$$

29.4.1 Empirical Formula for the Normalized Propagation Constant

For a single-mode step index fiber, a convenient empirical formula for $b(V)$ is given by

$$b(V) = \left(A - \frac{B}{V}\right)^2 \qquad 1.5 \lesssim V \lesssim 2.5 \qquad \textbf{(30)}$$

with $A \approx 1.1428$ and $B \approx 0.996$. The above formula gives values of b which are within about 0.2% of the exact values (see Table 29.1).

29.4.2 Spot Size of the Fundamental Mode

As mentioned earlier, a single-mode fiber supports only one mode that propagates through the fiber; this is also referred to as the fundamental mode of the fiber. The transverse field distribution associated with the fundamental mode of a single-mode fiber is an extremely important quantity, and it determines various important parameters such as splice loss at joints, launching efficiencies, and bending loss. For a step index fiber one has an analytical expression for the fundamental field distribution in terms of Bessel functions (see Sec. 29.3). For most single-mode fibers, the fundamental mode field distributions can be well approximated by a Gaussian function, which may be written in the form

$$\psi(x, y) = A\, e^{-\frac{x^2+y^2}{w^2}} = A\, e^{-\frac{r^2}{w^2}} \qquad \textbf{(31)}$$

where w is referred to as the spot size of the mode field pattern and $2w$ is called the mode field diameter (MFD). MFD is a very important characteristic of a single-mode optical fiber. For a step index fiber one has the following empirical expression for w (see Ref. 13):

$$\frac{w}{a} \approx 0.65 + \frac{1.619}{V^{3/2}} + \frac{2.879}{V^6}; \qquad 0.8 \le V \le 2.5 \qquad \textbf{(32)}$$

where a is the core radius. Many single-mode fibers used in optical communication systems do not have a step variation of refractive index—in fact, they often have very special refractive index distribution. Nevertheless, the modal field is very nearly Gaussian, and one usually describes the fiber though the mode field diameter (MFD). We may mention here that the light coming out of a He-Ne laser (or of a laser pointer) has a transverse intensity distribution very similar to that coming out of a single-mode fiber except that the spot size is much larger.

Example 29.6 Consider a step index fiber (operating at 1300 nm) with $n_2 = 1.447$, $\Delta = 0.003$, and $a = 4.2$ μ*m* (see Example 29.2). Thus $V \approx 2.28$, giving $w \approx 4.8$ μ*m*. The same fiber will have a V value of 1.908 at $\lambda_0 = 1550$ nm, giving a value of the spot size ≈ 5.5 μm. *Thus the spot size increases with wavelength*.

Example 29.7 For a step index fiber (operating at 1550 nm) with $n_2 = 1.444$, $\Delta = 0.0075$, and $a = 2.3$ μm (see Example 29.5), $V \approx 1.65$, giving $w \approx 3.6$ μm. The same fiber will have a V value of 1.97 at $\lambda_0 = 1300$ nm, giving a value of the spot size ≈ 3.0 μm.

29.4.3 Splice Loss Due to Transverse Misalignment

The most common misalignment at a joint between two similar fibers is the transverse misalignment similar to that shown in Fig. 29.4. Corresponding to a transverse misalignment of u, the loss in decibels is given by (see Prob. 29.7)

$$\alpha\,(\text{dB}) \approx 4.34\left(\frac{u}{w}\right)^2 \qquad \textbf{(33)}$$

Thus a larger value of w will lead to a greater tolerance to transverse misalignment. For $w \approx 5$ μm and a transverse offset of 1 μm, the loss at the joint will be approximately 0.17 dB; on the other hand, for $w \approx 3$ μm, a transverse offset of 1 μm will result in a loss of about 0.5 dB.

Fig. 29.4 A transverse alignment between two fibers would result in a loss of the optical beam.

Example 29.8 For a single-mode fiber operating at 1300 nm, $w = 5$ μm, and if the splice loss is to be below 0.1 dB, then from Eq. (18) we obtain $u < 0.76$ μm. Thus, for a low-loss joint, the transverse alignment is very critical, and connectors for single-mode fibers require precision matching and positioning for achieving low loss.

Many data sheets describing a commercially available single-mode fiber do not always give the actual refractive index profile. They instead give the MFD, maybe at more than one wavelength. They also give the cutoff wavelength (see, for example, Ref. 11). For example, the standard single-mode fiber designated as G.652 fiber when operating at 1.3 μm has a MFD of 9.2 ± 0.4 μm; the same fiber when operating at 1.55 μm has a MFD of 10.4 ± 0.8 μm.

29.5 PULSE DISPERSION IN SINGLE-MODE FIBERS

In single-mode fibers since there is only one mode, there is no intermodal dispersion. However, we have (in addition to material dispersion) waveguide dispersion which is characteristic of the transverse refractive index variation.[4] In Sections 10.2 and 27.10.3 we already discussed material dispersion. In this section we will show that even if n_1 and n_2 are independent of wavelength (i.e., even if there is no material dispersion), the group velocity of a particular mode will depend on the wavelength. This leads to what is known as the *waveguide dispersion.*

Since β represents the propagation constant, the group velocity of a particular mode is given by (see the analysis in Secs. 10.2 and 10.3)

$$\frac{1}{v_g} = \frac{d\beta}{d\omega} \tag{34}$$

Now from Eq. (20)

$$b = \frac{\beta/k_0 - n_2}{n_1 - n_2}\,\frac{\beta/k_0 + n_2}{n_1 + n_2} \tag{35}$$

Since for a guided mode β/k_0 lies between n_1 and n_2, and since for all practical single-mode fibers n_1 is very close to n_2 (see Examples 29.6 and 29.7), we may write the above equation as

$$b \approx \frac{\beta/k_0 - n_2}{n_1 - n_2} \tag{36}$$

Thus

$$\beta = \frac{\omega}{c}\left[n_2 + (n_1 - n_2)\,b(V)\right] \tag{37}$$

We will assume that n_1 and n_2 are independent of ω and calculate the group velocity:

$$\frac{1}{v_g} = \frac{d\beta}{d\omega} = \frac{1}{c}\left[n_2 + (n_1 - n_2)\,b(V)\right] + \frac{\omega}{c}(n_1 - n_2)\,\frac{db}{dV}\,\frac{dV}{d\omega} \tag{38}$$

Now

$$V = \frac{2\pi}{\lambda_0}\,a\sqrt{n_1^2 - n_2^2} = \frac{\omega}{c}\,a\sqrt{n_1^2 - n_2^2} \tag{39}$$

Thus

$$\frac{dV}{d\omega} = \frac{V}{\omega} \tag{40}$$

implying

$$\frac{1}{v_g} = \frac{1}{c}\left[n_2 + (n_1 - n_2)\,b(V)\right] + \frac{1}{c}(n_1 - n_2)\,V\,\frac{db}{dV} \tag{41}$$

or,

$$\frac{1}{v_g} = \frac{n_2}{c} + \frac{n_1 - n_2}{c}\left[\frac{d}{dV}(bV)\right] \tag{42}$$

Thus, the time taken by a pulse to traverse length L of the fiber is given by

$$\tau = \frac{L}{v_g} = \frac{L}{c}\,n_2\left[1 + \Delta\,\frac{d}{dV}(bV)\right] \tag{43}$$

where

$$\Delta \equiv \frac{n_1^2 - n_2^2}{2n_1^2} \approx \frac{n_1 - n_2}{n_2} \tag{44}$$

and we have assumed $n_1 \approx n_2$. From Eq. (43) we see that even if n_1 and n_2 are independent of wavelength (i.e., if there is no material dispersion), the group velocity (and hence τ) will depend on ω because, as is obvious from Fig. 29.3 [and Eq. (30)], b depends on V. This leads to what is known as the *waveguide dispersion.* Physically this arises due to the fact that the spot size depends on the wavelength (see Examples 29.6 and 29.7). For a source having a spectral width $\Delta\lambda_0$, the corresponding waveguide dispersion is given by [see Eqs. (16) and (17) of Chap. 10]

$$\Delta\tau_w = \frac{d\tau}{d\lambda_0}\,\Delta\lambda_0 \cong \frac{L}{c}\,n_2\,\Delta\,\frac{d^2}{dV^2}(bV)\,\frac{dV}{d\lambda_0}\,\Delta\lambda_0 \tag{45}$$

From Eq. (37) we find

$$\frac{dV}{d\lambda_0} = -\frac{V}{\lambda_0} \tag{46}$$

[4] At very high bit rates, we also have what is known as *polarization mode dispersion* (abbreviated as PMD). This may arise due to many factors; for example, if there is slight ellipticity in the core of the fiber, then the two states of polarization travel with slightly different group velocities, leading to what is known as PMD. However, this phenomenon becomes important at very high bit rates—above 40 Gbits s^{-1}. For a nice overview of PMD, see Ref. 14; for more details see references therein.

Thus

$$\Delta\tau_w = -\frac{L n_2 \Delta}{c} f(V)\Delta\lambda_0 \qquad \textbf{(47)}$$

where

$$f(V) \equiv V\frac{d^2}{dV^2}(bV) \qquad \textbf{(48)}$$

For a step index fiber, b as a function of V is a universal curve [in fact this is true for a fiber with a power law profile given by Eq. (27)]; therefore the variation of $f(V)$ with V will also be universal (see Table 29.1). A convenient empirical formula for a step index fiber is given by (Ref. 15)

$$f(V) \approx 0.080 + 0.549\left(2.834 - V\right)^2 \qquad 1.3 < V < 2.4 \qquad \textbf{(49)}$$

A comparison between the above empirical values and the exact values has been made in Ref. 3. Thus

$$\Delta\tau_w = -\frac{L}{c} n_2 \Delta\left[0.080 + 0.549\left(2.834 - V\right)^2\right]\frac{\Delta\lambda_0}{\lambda_0}$$

$$\text{for} \quad 1.3 < V < 2.4 \qquad \textbf{(50)}$$

As in Sec. 10.2, we assume $\Delta\lambda_0 = 1$ nm $= 10^{-9}$ m and $L = 1$ km $= 1000$ m, and we define the dispersion coefficient as

$$D_w \equiv \frac{\Delta\tau_w}{L\,\Delta\lambda_0} \approx -\frac{n_2\Delta}{3\lambda_0}\times 10^7\left[0.080 + 0.549\left(2.834 - V\right)^2\right]$$

$$\text{ps km}^{-1}\,\text{nm}^{-1} \qquad \textbf{(51)}$$

where λ_0 is measured in nanometers and we have assumed $c = 3\times 10^{-4}$ m ps^{-1}. The quantity D_w is referred as the waveguide dispersion coefficient (because it is due to the waveguiding properties of the fiber), hence the subscript w on D. In the single-mode regime, the quantity within the brackets in Eq. (51) is usually positive; hence the waveguide dispersion is negative, indicating that longer wavelengths travel faster. Since the sign of material dispersion depends on the operating wavelength region, it is possible that the two effects, namely, material and waveguide dispersions, cancel each other at a certain wavelength. Such a wavelength, which is a very important parameter of single-mode fibers, is referred to as the zero dispersion wavelength λ_{ZD}.

The total dispersion is given by the sum of material and waveguide dispersions[5]:

$$D_{\text{tot}} = D_m + D_w \qquad \textbf{(52)}$$

Let us consider the two single-mode fibers discussed in Examples 29.6 and 29.7.

29.5.1 Conventional Single-Mode (G 652) Fibers

We consider the fiber discussed in Example 29.6 for which $n_2 = 1.447$, $\Delta = 0.003$, and $a = 4.2$ μm so that $V = 2958/\lambda_0$, where λ_0 is measured in nanometers. Substituting in Eq. (51), we get

$$D_w = -\frac{1.447\times 10^4}{\lambda_0}\left[0.080 + 0.549\left(2.834 - \frac{2958}{\lambda_0}\right)^2\right]$$

$$\text{ps km}^{-1}\,\text{nm}^{-1}$$

Elementary calculations show that at $\lambda_0 \approx 1300$ nm, $D_w = -2.8$ ps km^{-1} nm^{-1}. The variations of D_m, D_w, and D_{tot} with λ_0 are shown in Fig. 29.5; the variation of D_m is calculated by using Eq. (36) of Chap. 27 and Table 10.1. The total dispersion passes through zero around $\lambda_0 \approx 1300$ nm which is the *zero total dispersion wavelength* and represents an extremely important parameter. Such fibers that have zero dispersion around $\lambda_0 \approx 1300$ nm are known as conventional single-mode (or G 652) fibers and are extensively used in optical communication systems.

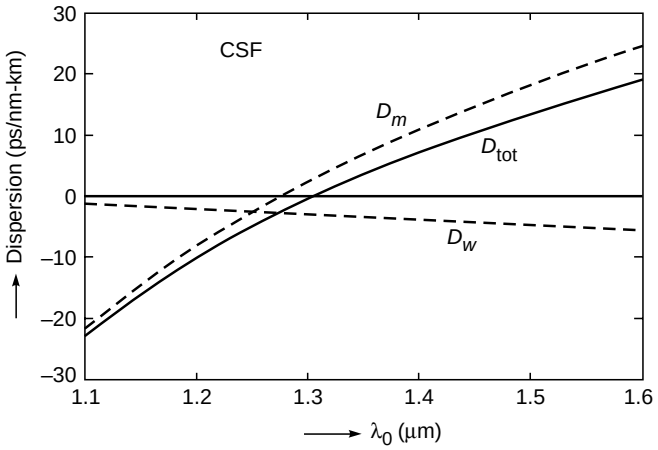


Fig. 29.5 The wavelength dependence of D_m, D_w and D_{tot} for a typical conventional single-mode fiber (CSF) with parameters as given in Example 29.3. The total dispersion passes through zero around $\lambda_0 \approx 1300$ nm which is known as zero dispersion wavelength.

[5] Strictly speaking, material and waveguide dispersions are not additive. For a given variation of $n^2(r)$, one really should solve Eq. (11) at different wavelengths, taking into account the wavelength dependence of the refractive index, and determine β as a function of λ_0. This is indeed done in the software developed in Ref. 8.

29.5.2 Dispersion Shifted (G 653) Fibers

We next consider the fiber discussed in Example 24.7 for which $n_2 = 1.444$, $\Delta = 0.0075$, and $a = 2.3$ μm, so that $V = 2556/\lambda_0$, where, once again, λ_0 is measured in nanometers. Substituting in Eq. (49) we get

$$D_w = -\frac{3.61\times10^4}{\lambda_0}\left[0.080 + 0.549\left(2.834 - \frac{2556}{\lambda_0}\right)^2\right] \text{ ps km}^{-1}\,\text{nm}^{-1}$$

Thus at $\lambda_0 \approx 1550$ nm,

$$D_w = -20 \text{ ps km}^{-1}\,\text{nm}^{-1}$$

On the other hand, the material dispersion at this wavelength is given by (see Table 10.1)

$$D_m = +20 \text{ ps km}^{-1}\,\text{nm}^{-1}$$

We therefore see that the two expressions are of opposite sign and almost cancel each other. Physically, because of waveguide dispersion, longer wavelengths travel slower than shorter wavelengths; and because of material dispersion, longer wavelengths travel faster than shorter wavelengths. The two effects compensate each other, resulting in zero total dispersion around 1550 nm. The corresponding variation of D_m, D_w, and D_{tot} with wavelength is shown in Fig. 29.6. As can be seen from the figure, we have been able to shift the zero dispersion wavelength by changing the fiber parameters; these are known as the *dispersion shifted fibers.* Thus dispersion shifted fibers are those fibers whose total dispersion becomes zero at a shifted wavelength. Commercially available dispersion shifted fibers (which are abbreviated as DSF and referred to as G 653 fibers) do not usually have a step variation of refractive index; the refractive index variation is bit complicated and is such that the total dispersion passes through zero around 1550 nm wavelength.

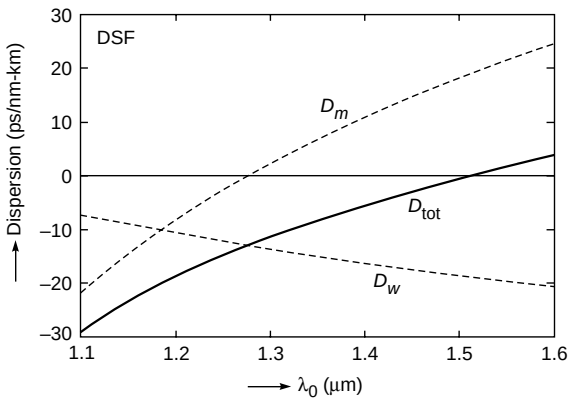


Fig. 29.6 The wavelength dependence of D_m, D_w and D_{tot} for a typical dispersion shifted fiber (DSF) with parameters as given in Example 29.4. The zero dispersion wavelength is around 1550 nm.

29.6 DISPERSION COMPENSATING FIBERS

In many countries there already exist millions of kilometers of conventional single-mode fibers (of the type discussed in Example 29.6) in the underground ducts operating at 1310 nm; and as mentioned in Sec. 29.5.1, these fibers have very low dispersions around 1310 nm. One could significantly increase the transmission capacity of these systems by operating these fibers at 1550 nm (where the loss is extremely small), and we can have the added advantage of using EDFAs (erbium-doped fiber amplifiers) for optical amplification in this wavelength range (see Sec. 26.1.3). However, if we operate the conventional single-mode fibers at 1550 nm, we will have a significant residual dispersion; and as discussed in Sec. 29.5.1, this residual dispersion is about 20 ps km^{-1} nm^{-1}. Such a large dispersion will result in a significant decrease in the information-carrying capacity of the communication system. On the other hand, replacing the existing conventional single-mode fibers by dispersion shifted fibers (DSFs) involves huge costs. As such, in recent years there has been a considerable amount of work in upgrading of the installed 1310 nm optimized optical fiber links for operation at 1550 nm. This is achieved by developing fibers with very large negative dispersion coefficients, a few hundred meters to a kilometer of which can be used to compensate for dispersion over tens of kilometers of the fiber in the link.

In Secs. 29.5.1 and 29.5.2 we have seen that by changing the refractive index profile, we can alter the waveguide dispersion and hence the total dispersion. Indeed, it is possible to have specially designed fibers whose dispersion coefficient D_{tot} is large and negative at 1550 nm. A typical refractive index profile, which is characterized by $D_{\text{tot}} \approx -1800$ ps km^{-1} nm^{-1} at 1550 nm, is shown in Fig. 29.7 (Ref. 16)[6]. These types of fibers are known as dispersion compensating fibers (DCFs). A short length of DCF can be used in conjunction with the 1310 nm optimized fiber link so as to have a small total dispersion value at the end of the link (see Fig. 29.8).

To understand this phenomenon, we have plotted in Fig. 29.9 (as a solid curve) a typical variation of the group velocity v_g with wavelength for a conventional single-mode fiber (CSF) with zero dispersion around 1300 nm wavelength. As can be seen from the figure, v_g attains a maximum value at the zero

[6] The refractive index variation used in Ref. 16 is based on the refractive index profile suggested in Ref. 17.

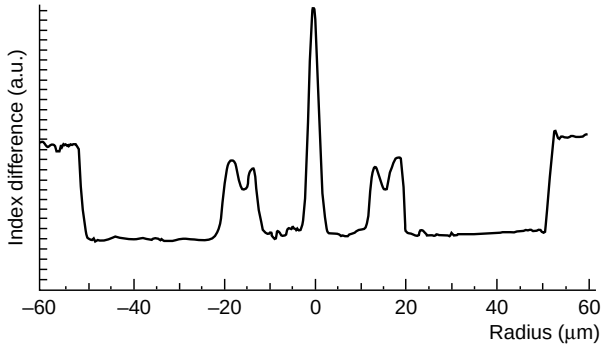


Fig. 29.7 The refractive index profile of a typical dispersion compensating fiber (DCF) characterized by $D_{tot} \approx -200$ ps km^{-1} nm^{-1} at 1550 nm [Adapted from Ref. 12].

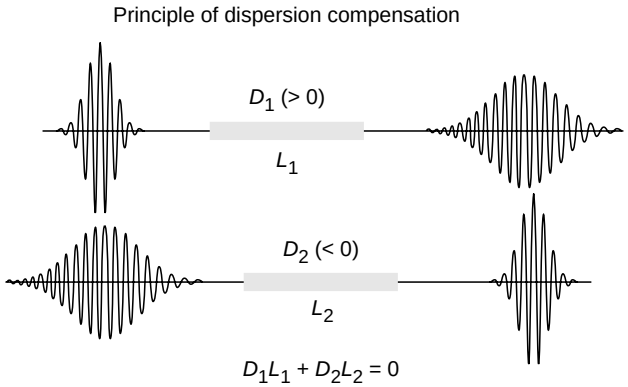


Fig. 29.8 A short length of a DCF can be used in conjunction with the conventional single mode fiber so as to have small dispersion value at the end of the link.

dispersion wavelength, and on either side it monotonically decreases with wavelength. Thus, if the central wavelength of the pulse is around 1550 nm, then the red components of the pulse (i.e., longer wavelengths) will travel slower than the blue components (i.e., smaller wavelengths) of the pulse. Because of this the pulse will get broadened. Now, after propagating through a CSF for a certain length L_1, the pulse is allowed to propagate through a length L_2 of the DCF in which the group velocity v_g varies as shown by the dashed curve in Fig. 29.9. The red components (i.e., longer wavelengths) will now travel faster than the blue components, and the pulse will tend to reshape itself into its original form. Indeed, if the lengths of the two fibers L_1 and L_2 are such that

$$D_1L_1 + D_2L_2 = 0 \qquad \textbf{(53)}$$

then the pulse emanating from the second fiber will be almost identical to the pulse entering the first fiber as shown in Fig. 29.9.

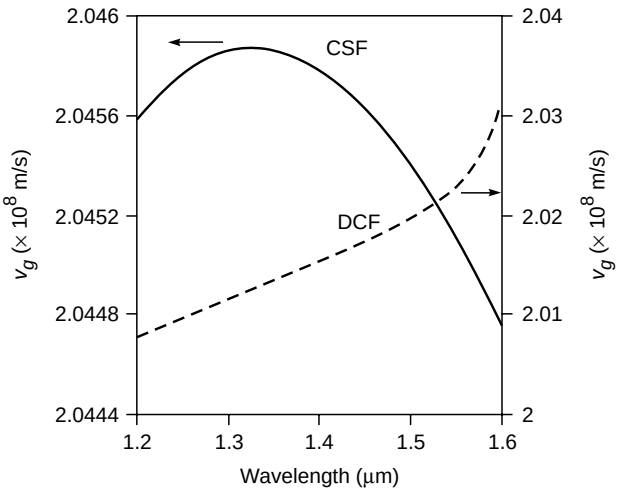


Fig. 29.9 The wavelength variation of group velocity for a typical dispersion compensating fiber and a typical conventional single-mode fiber.

The latest trend in optical communication has been to use DWDM (dense wavelength division multiplexed) systems in which many closely spaced wavelengths (in the wavelength region 1530 to 1565 nm) are simultaneously propagated and amplified by erbium-doped fiber amplifiers. Now, if the fiber is operated at the zero dispersion wavelength, then all nearby wavelengths will travel with the same group velocity because of which they interact with each other to create new frequencies—this is known as four wave mixing usually abbreviated as FWM. To overcome this difficulty, the use of small dispersion fiber has been suggested, where the dispersion is typically in the range of 2–8 ps km^{-1} nm^{-1}. Because of this, different wavelengths travel with different velocities, and the unwanted frequencies are not generated. In the inset of Fig. 29.10, we have given typical refractive index variations of a small dispersion fiber named the *small residual dispersion fiber* (SRDF). The figure also shows the corresponding total dispersion D_N as a function of wavelength; the tolerance of the dispersion characteristics on the refractive index profile is shown by dotted lines. However, if one wants repeaterless transmission over very large distances, the residual dispersion (2–8 ps km^{-1} nm^{-1}) in these fibers will go on accumulating and will limit the number of bits that can be sent at each wavelength. To overcome this difficulty, one has to use a DCF which will compensate the accumulated dispersion at all wavelengths simultaneously. The design of the DCF, therefore, has to be compatible with the small residual dispersion fibers. In the inset of Fig. 29.11, we have given typical refractive index variations of the corresponding DCF; the corresponding wavelength dependence of the

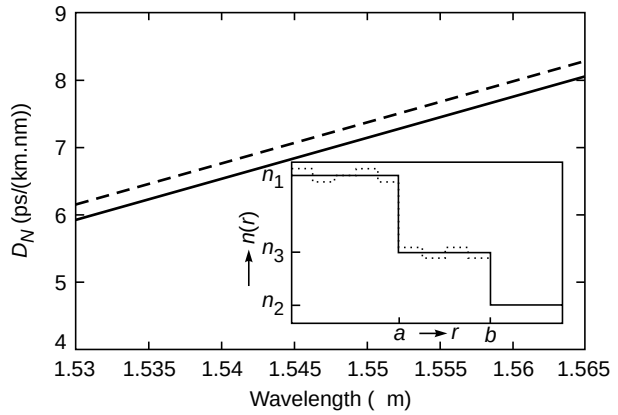


Fig. 29.10 Variation of the total dispersion D_N of the SRDF as a function wavelength. The solid and dashed curves correspond to the proposed and the perturbed refractive index profiles (shown schematically in inset), respectively [Figure adapted from Ref. 18].

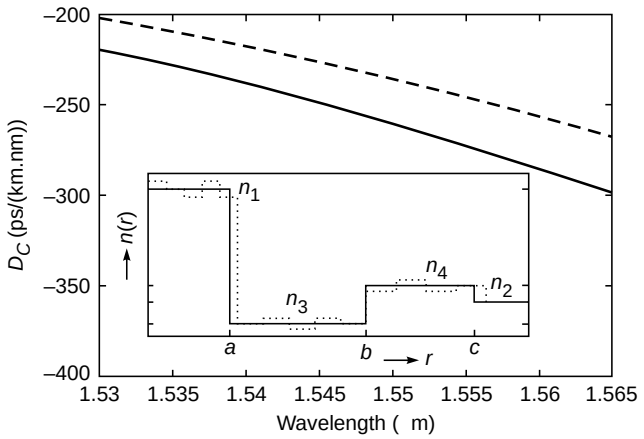


Fig. 29.11 Variation of the total dispersion D_C of the DCF as a function of wavelength. The solid and dashed curves correspond to the proposed and the perturbed refractive index profiles (shown schematically in inset), respectively [Figure adapted from Ref. 18].

total dispersion D_C has also been shown. The dispersion slopes are so adjusted that a small length of the DCF will approximately compensate the accumulated dispersion in SRDF simultaneously at all wavelengths. In Fig. 29.12 we have plotted

$$D_E = \frac{L_1 D_N + L_2 D_C}{L_1 + L_2}$$

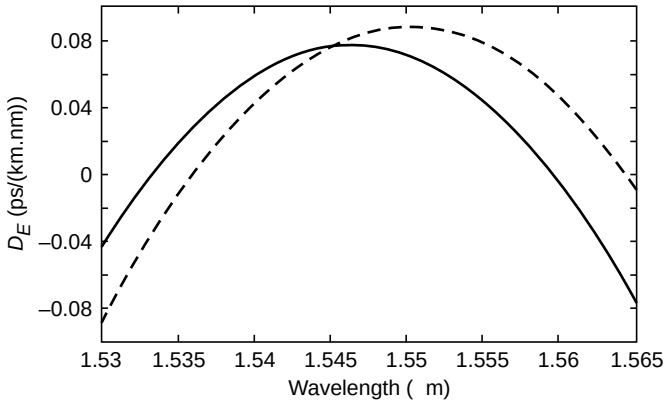


Fig. 29.12 Variation of the effective dispersion D_E of the system as a function of wavelength [Figure adapted from Ref. 18].

for $L_1 = 36.74L_2$, where D_N and D_C represent the dispersions associated with the SRDF and DCF, respectively. Note that the maximum value of the effective dispersion is less than 0.08 ps km^{-1} nm^{-1}.

Problems

29.1 Consider a step index fiber with $n_1 = 1.474$ and $n_2 = 1.470$ and having a core radius $a = 4.5$ μm. Determine the cutoff wavelength.

[**Ans:** $\lambda_c = 1.28$ μm]

29.2 Consider a step index fiber with $n_1 = 1.5$ and $n_2 = 1.48$ and having a core radius $a = 6.0$ μm. Determine the operating wavelength λ_0 for which $V = 8$.

[**Ans:** $\lambda_0 = 1.15$ μm]

29.3 In continuation of Prob. 29.2, **(a)** calculate the total number of modes for $V = 8$ and **(b)** compare with the approximate value given by Eq. (21) of Chapter (27).

[**Ans: (a)** 34; **(b)** 32]

29.4 Consider a single-mode fiber with $a = 3$ μm operating with a V number of 2.3. Calculate the spot size of the fundamental mode.

[**Ans:** 3.4 μm]

29.5 Assume the single-mode fiber to have a Gaussian spot size $w = 4.5$ μm. Calculate the splice loss at a joint between two such identical fibers with a transverse misalignment of 1, 2, and 3 μm.

[**Ans:** 0.21, 0.86, and 1.93 dB]

29.6 The modal field is said to be normalized if

$$\iint |\psi(x, y)|^2 dx\, dy = 1$$

Show that the normalized Gaussian field is given by

$$\psi(x, y) = \sqrt{\frac{2}{\pi}} \frac{1}{w} e^{-x^2 + y^2 / w^2} = \sqrt{\frac{2}{\pi}} \frac{1}{w} e^{-r^2 / w^2}$$

29.7 Consider two identical single-mode fibers joined together with a transverse misalignment of u (along the x axis). The fractional power that is coupled to the fundamental mode of the second fiber is given by the overlap integral

$$T = \left| \iint \psi_1(x, y)\, \psi_2(x, y)\, dx\, dy \right|^2$$

Show that

$$T = \exp\left(-\frac{u^2}{w^2}\right)$$

Thus

$$\text{Loss in dB} = 10 \log T = 4.34 \left(\frac{u}{w}\right)^2$$

29.8 Using the results derived in Sec. 28.5 (and using the method of separation of variables), solve the scalar wave equation for an infinitely extended parabolic index fiber characterized by the following refractive index variation

$$n^2(r) = n_1^2\left[1 - 2\Delta\left(\frac{r}{a}\right)^2\right] = n_1^2\left[1 - 2\Delta\frac{x^2 + y^2}{a^2}\right]$$

Derive the expression for propagation constants and show that (for low order modes and for $\Delta << 1$) the group velocity is approximately independent of the mode number. Also, using a method similar to that discussed in Sec. 28.5, calculate approximately the number of modes for a given value of V and compare with the one obtained by using Eq. (21) of Chapter (27).

[**Ans:** $\beta^2 = \beta_{mn}^2 \approx k_0^2 n_1^2 - 2(m + n + 1)\gamma k_0$;

$m, n = 0, 1, 2, 3, \ldots$. where $\gamma = \dfrac{n_1\sqrt{2\Delta}}{a}$]

REFERENCES AND SUGGESTED READINGS

1. A. W. Snyder and J. D. Love, *Optical Waveguide Theory*, Chapman & Hall, London, 1983.
2. D. Gloge, "Weakly Guiding Fibers," *Appl. Opt.*, Vol. 10, p. 2252, 1971.
3. A. Ghatak and K. Thyagarajan, *Introduction to Fiber Optics*, Cambridge University Press, Cambridge, 1998.
4. J. Irving and N. Mullineux, *Mathematics in Physics and Engineering*, Academic Press, New York, 1959.
5. G. Arfken, *Mathematical Methods for Physicists*, 2d ed., Academic Press, New York, 1970.
6. A. K. Ghatak, I. C. Goyal, and S. J. Chua, *Mathematical Physics*, Macmillan India, New Delhi, 1985.
7. A. Ghatak, A. Sharma, and R. Tewari, *Fiber Optics on a PC*, Viva Books, New Delhi, 1994.
8. A. Ghatak, I. C. Goyal, and R. Varshney, *FIBER OPTICA: A Software for Characterizing Fiber and Integrated-Optic Waveguides*, Viva Books, New Delhi, 1999.
9. M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, 1965.
10. T. I. Lukowski and F. P. Kapron, "Parabolic Fiber Cutoffs: A Comparison of Theories," *J. Opt. Soc. Am.*, Vol. 67, p. 1185, 1977.
11. D. K. Mynbaev and L. L. Scheiner, *Fiber-Optic Communications Technology*, Prentice-Hall, Englewood Cliffs, N.J., 2001.
12. D. Gloge and E. A. J. Marcatili, "Multimode Theory of Graded-Core Fibers," *Bell. Syst. Tech. J.*, Vol. 52, p. 1563, 1973.
13. D. Marcuse, "Gaussian Approximation of the Fundamental Modes of a Graded Index Fiber," *J. Opt. Soc. Am.*, Vol. 68, p. 103, 1978.
14. Arun Kumar, "Polarization Effects in Single Mode Optical Fibers," in *Guided Wave Optics* (Ed. Anurag Sharma), Viva Books, New Delhi, 2005.
15. D. Marcuse, "Interdependence of Waveguide and Material Dispersion," *Appl. Opt.*, Vol. 18, pp. 2930–2932, 1979.
16. J. L. Auguste et al., *Electron. Lett.*, Vol. 36, p. 1689, 2000.
17. K. Thyagarajan, R. Varshney, P. Palai, A. Ghatak, and I. C. Goyal, "A Novel Design of a Dispersion Compensating Fiber," *Photon. Tech. Letts.*,Vol. 8, p. 1510, 1996.
18. I. C. Goyal, R. K. Varshney, and A. K. Ghatak, "Design of a Small Residual Dispersion Fiber and a Corresponding Dispersion Compensating Fiber for DWDM Systems," *Optical Engineering*, 42, pp. 977–980, 2003.

PART 8
Special Theory of Relativity

This part consists of two short chapters, Chaps. 30 and 31, discussing the postulates and applications of the special theory of relativity.

Chapter Thirty

SPECIAL THEORY OF RELATIVITY I: TIME DILATION AND LENGTH CONTRACTION

Towards the end of the 19th century scientists believed they were close to a complete description of the universe. They imagined that space was filled everywhere by a continuous medium called the ether. Light rays and radio signals were waves in this ether just as sound is pressure waves in air. All that was needed to complete the theory was careful measurements of the elastic properties of the ether; once they have those nailed down, everything else would fall into place. Soon, however, discrepancies with the idea of an all pervading ether begin to appear. You would expect light to travel at a fixed speed through the ether. So if you were traveling in the same direction as the light, you would expect that its speed would appear to be lower, and if you were traveling in the opposite direction to the light, that its speed would appear to be higher. Yet a series of experiments failed to find any evidence for differences in speed due to motion through the ether.

—Stephen Hawking, in *A Brief History of Relativity* in the December 31, 1999, issue of *Time Magazine*

30.1 INTRODUCTION

A train is moving past a platform. For a person on the platform, the train is moving at a speed of say 50 km h^{-1}. I am inside the train, and if I throw a tennis ball horizontally (along the length of the train) at a speed of 10 km h^{-1}, then for a person on the platform the ball will move away at a speed of about 60 km h^{-1}. I next hold a laser pointer (I am still inside the moving train); for the person on the platform, the laser pointer is also moving with the speed of the moving train. Now, when I switch on the laser pointer, the light emitted by the laser pointer travels with the same speed with respect to me as well as for the observer on the platform. Thus the speed of light in vacuum (which is denoted by c) does not depend on the speed of the source of light. This was the remarkable statement that was made by Albert Einstein in his famous 1905 paper (Ref. 1). To quote from the English translation of this paper:

> Light always propagates in empty space with a definite velocity V that is independent of the state of motion of the emitting body. . . .

(The quantity V in Einstein's paper is now usually denoted by c.) That the speed of light does not depend on the speed of the source of light has since been verified in many experiments. The most important experiment was carried out in 1964 by Alvager (and his colleagues) (Ref. 2). In this experiment, neutral pi mesons traveling with speeds very close to that of light were produced; neutral pi mesons (denoted by π^0) have a mass of about 264 times the mass of an electron and decay (with a mean lifetime of about 8×10^{-17} s) to two gamma ray photons:

$$\pi^0 \rightarrow \gamma + \gamma$$

The photons (from the decay of very fast moving neutral pi mesons) were found to travel at a speed c. The measurement of the speed of the gamma ray photons was difficult, but it was unambiguously established that their speed was equal to c. Why do we require a fast-moving π^0 meson as a source of photon? Because it would be extremely difficult (and would require an enormous amount of energy) to make an object such as an ordinary light source travel with a speed close to that of light—see Example 31.1. There are other experiments which also show that the speed of light in vacuum does not depend on the speed of the source of light (see, e.g., Ref. 3).

In the year 1905 Einstein, while working at the Swiss Patent Office, published five outstanding papers; the year 1905 is therefore referred to as Einstein's *year of miracles*. All five papers appeared in the journal *Annalen der Physik* published from Germany, and English translations of the original papers appear in the book (Ref. 4) with the very appropriate title *Einstein's Miraculous Year: Five Papers That Changed the Face of Physics.*

Before we state the postulates of the special theory of relativity, it is necessary to define an inertial system:

> An inertial system is one in which Newton's first law holds.

That raises the question, "What is Newton's first law?" Newton wrote his famous laws in his incredible book entitled *Principia*.[1] The book was in Latin, and according to the English translation of this book, the first law is (quoted from Ref. 5)

> Every body perseveres in its state of rest, or of uniform motion in a straight line, unless it is compelled to change that state by forces impressed thereon.

Feynman writes Newton's first law as follows (Ref. 6):

> If something is moving, with nothing touching it and completely undisturbed, it will go on forever, coasting at a uniform speed in a straight line. (*Why* does it keep on coasting? We do not know, but that is the way it is.)

Feynman further writes, "Newton modified this idea, saying that the only way to change the motion of a body is to use force. If the body speeds up, a force has been applied in the direction of motion."

Further, any system moving with constant velocity with respect to an inertial system is also an inertial system. And Newton had written that the laws of mechanics (which determine the motion of bodies) are the same in all inertial systems. This implies, for example, that (to quote Feynman) "if a space ship is drifting along at a uniform speed, all experiments performed in the space ship will appear the same as if the ship was not moving, provided of course, that one does not look outside. That is the meaning of the principle of relativity." Einstein found that for the laws of electricity and magnetism (described by Maxwell's equations) to remain the same in a moving spaceship, the speed of light in vacuum should not depend on the speed of the source of light. This led Einstein to put forward (in 1905) the following two postulates of the special theory of relativity:

1. The first postulate states that the laws of physics are the same in all inertial systems.
2. The second postulate states that the speed of light in vacuum (which is denoted by c) does not depend on the speed of the source of light.

The first postulate was known much before Einstein. Isaac Newton, in one of his corollaries to the laws of motion, had written:[2]

> The motions of bodies included in a given space are the same among themselves, whether that space is at rest or moves uniformly forward in a straight line.

The first postulate is also known as the principle of relativity, and in 1904 the famous French mathematician Henri Poincaré stated this very precisely.[3]

> According to the principle of relativity, the laws of physical phenomena must be the same for a fixed observer as for an observer who has a uniform motion of translation relative to him, so that we have not, nor can we possibly have, any means of discerning whether or not we are carried along in such a motion.

30.2 SPEED OF LIGHT FOR A MOVING SOURCE

Let us consider two coordinate systems S and S' which are in uniform relative motion along the x axis as shown in Fig. 30.1. We have two persons A and B; A is at rest in the

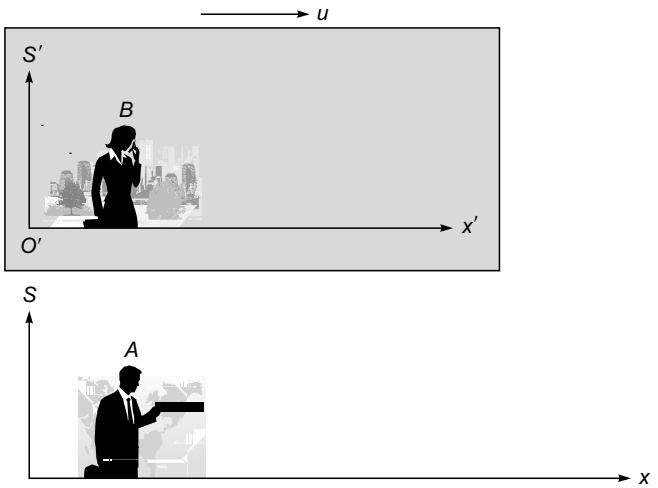


Fig. 30.1 Person A is on the platform, and person B is inside a train moving with velocity u in the $+x$ direction. According to A, person B is moving in the x direction with a constant velocity u. On the other hand, according to B, person A is moving in the $-x$ direction with the same speed u. Person A is holding a light source (such as a laser), and both A and B measure the same speed of light.

[1] The full title of Newton's book (in Latin) is *Philosophiæ Naturalis Principia Mathematica*.
[2] The author found this in Ref. 6; see also Ref. 5.
[3] The author found this in Ref. 6. Poincaré was the first to present the Lorentz transformations in their modern symmetric form.

coordinate system S and B is at rest in the coordinate system S'; thus according to A, person B is moving in the $+x$ direction with a constant velocity u. On the other hand, according to B, person A is moving in the $-x$ direction with the same speed u. Figure 30.1 shows A holding a light source (such as a laser), and of course, according to A the speed of light is c. Now, according to observer B, the laser pointer is moving in the $-x$ direction with speed u, and therefore according to the second postulate of Einstein, B must also measure the same speed of light. Thus we infer that

> A person moving with respect to a light source measures the same speed of light as the person who is stationary with respect to the light source.

30.3 TIME DILATION

Consider an observer B inside a train moving with speed u on a railway track. Inside the train (which is our reference frame S'), B produces a light pulse (by switching on a bulb and very quickly switching it off), allows the light beam to get reflected by a mirror M (which is right above the bulb), and detects the reflected light by a detector D (see Fig. 30.2). We have therefore two events: the first event is the switching on of the bulb, producing a light pulse, and the second event is its subsequent detection by the detector. Person B measures the time interval $\Delta t'$ between the two events; this time is obviously given by

$$\Delta t' = \frac{2H}{c} \qquad \textbf{(1)}$$

where H is the distance between the floor and the mirror as shown in Fig. 30.2. For an observer A on the platform (which is our reference frame S), the whole train is moving with speed u, and therefore the light beam will take a diagonal path which is longer than observed by B (see Fig. 30.3). Since the velocity of light is always the same, the time interval between the two events (as observed by A) will take a longer time by his clock. If Δt represents the time interval measured by A, then

$$\Delta t = \frac{PM + MD}{c} \qquad \textbf{(2)}$$

where we have used the fact that the speed of light for the observer outside the train will be the same as observed by the person inside the train. The time taken by the light ray to traverse the path PM (or the path MD) is $\Delta t/2$, and therefore

$$(PM)^2 = (MD)^2 = H^2 + \left(\frac{u\Delta t}{2}\right)^2$$

or

$$PM = MD = \sqrt{H^2 + \left(\frac{u\Delta t}{2}\right)^2}$$

Substituting in Eq. (2), we get

$$\Delta t = \frac{2}{c}\sqrt{H^2 + \left(\frac{u\Delta t}{2}\right)^2} \qquad \textbf{(3)}$$

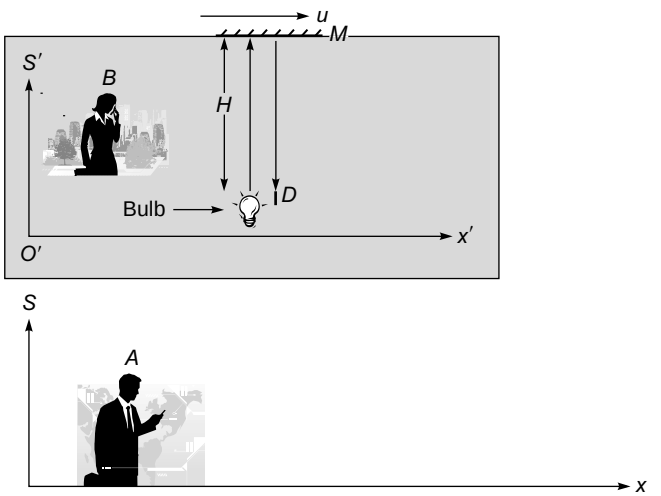


Fig. 30.2 An observer B is inside a train which is moving with speed u on a railway track. Inside the train (which is our reference frame S'), B switches on a bulb, allows the light beam to get reflected by a mirror M (which is right above the bulb), and detects the reflected beam by a detector D.

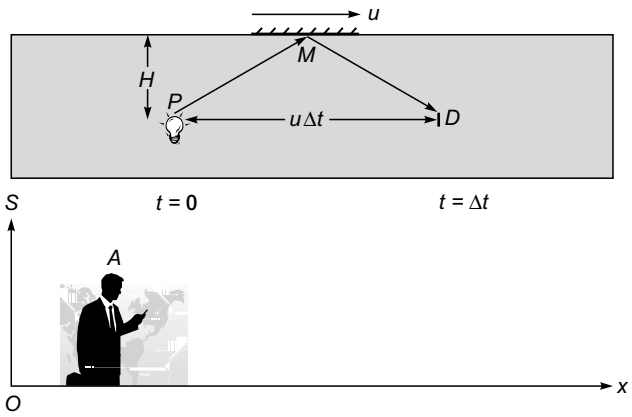


Fig. 30.3 According to A (who is on the platform), when the light reaches the detector D (via the mirror M), it has moved through a distance $u\,\Delta t$.

or $$(\Delta t)^2 = \frac{4H^2}{c^2} + \left(\frac{u\Delta t}{c}\right)^2 = (\Delta t')^2 + (\Delta t)^2 \frac{u^2}{c^2} \quad \textbf{(4)}$$

or $$\Delta t' = \sqrt{1-\frac{u^2}{c^2}}\,\Delta t \quad \textbf{(5)}$$

For the observer B on the train, the lightbulb and the mirror on the roof are stationary so that the two events (switching on the bulb and its subsequent detection by the detector) occur at the same place. The time interval between two events occurring at the same position is known as the *proper time*; thus $\Delta t'$ represents the *proper time* between the two events. To quote from Ref. 8,

> The *proper time* interval between two events is the time interval measured in the reference frame in which the two events occur at the same position. Time intervals that occur at different positions are called *improper*.

On the other hand, for the observer A (outside the train) both the lightbulb and the mirror are moving with velocity u, and the two events occur at different places.

Thus Eq. (5) represents this important result:

The time interval between two events occurring *at the same place* in a particular reference frame S' (referred to as the *proper time*)

$$= \sqrt{1-\frac{u^2}{c^2}} \times \begin{array}{l}\text{time interval between two events}\\ \text{in any reference frame } S \text{ moving}\\ \text{with relative speed } u\end{array} \quad \textbf{(6)}$$

Equation (5) is often written in the form

$$\Delta t = \gamma\,\Delta t' \quad \textbf{(7)}$$

where $$\gamma = \frac{1}{\sqrt{1-u^2/c^2}} \quad \textbf{(8)}$$

is known as the Lorentz factor.

30.4 THE MU MESON EXPERIMENT

A mu meson (also known as muon) is a negatively charged elementary particle which has exactly the same charge as the electron but has a mass about 207 times the mass of the electron. In 1937, mu mesons were first detected in cosmic rays by S. H. Neddermeyer and C. D. Anderson. These particles are created at the top of our atmosphere. It is believed that when high-energy protons (from outer space) collide with molecules in the outer region of the atmosphere, many particles (including the mu mesons) are created. Mu mesons have also been created in the laboratories. Mu mesons are radioactive and undergo the following decay

$$\text{Mu meson} \rightarrow \text{electron} + \text{neutrino} + \text{antineutrino}$$

The mean lifetime of the above process is about 2.2 μs. Thus if there are N_0 muons at $t = 0$ (at rest in the laboratory), then at a later time t, the number of muons which would not have undergone decay is given by

$$N(t) = N_0 \exp\left(-\frac{t}{\tau}\right) \quad \textbf{(9)}$$

where τ (≈ 2.2 μs) represents the mean lifetime of the muon. For example, to start, if we have 1000 muons, then after about 2.2 μs we will have about N_0/e (≈ 368) muons which did not undergo decay. The quantity $(\ln 2)\tau$ ($\approx 0.693\tau \approx 1.525$ μs) represents the half-life of the muon. In this time, one-half the number of muons will not have undergone decay:

$$1000 \quad \exp\left(-\frac{1.525\ \mu s}{2.2\ \mu s}\right) \approx 500$$

In 1941 Rossi and Hall carried out an experiment at the top of Mt. Washington which is about 6300 ft (or about 1920 m) above sea level. It was found that about 568 mu mesons were detected in about 1 h (the numbers taken from Ref. 10). The velocities of mu mesons were about $0.995c$, and therefore for an observer on the Earth, it would take about

$$\frac{1920\text{m}}{0.995\times 3\times 10^8\ \text{m/s}} \approx 6.4\ \mu s$$

to traverse the distance of 1920 m [see Fig. 30.4(a)]. In this time, the number of mu mesons that should reach the surface of the Earth is about

$$568 \exp\left(-\frac{6.4\ \mu s}{2.2\ \mu s}\right) \approx 568e^{-2.9} \approx 31 \quad \textbf{(10)}$$

Thus in traversing the distance of 1920 m, about 537 mu mesons should have undergone decay and only 31 of them should have reached the surface of the Earth. However, when the experiment was performed, it was found that about 412 muons were detected. This will correspond to a mean lifetime of τ' where

$$568 \exp\left(-\frac{6.4\ \mu s}{\tau'}\right) \approx 412 \qquad \Rightarrow \qquad \tau' \approx 20\ \mu s \quad \textbf{(11)}$$

Thus detection of 412 muons on the surface of the Earth would imply a muon mean lifetime of about 20 μs.

We may understand the above experiment by noting that in the muon reference frame [i.e., inside the spaceship where the muon is at rest; see Fig. 30.4(b)], the position of the

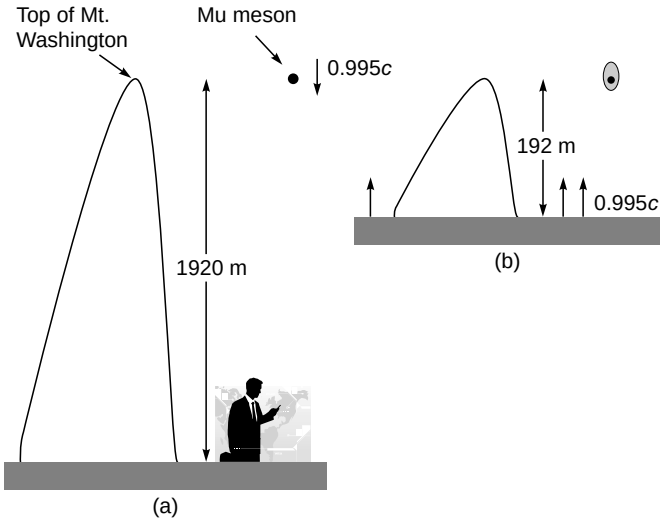


Fig. 30.4 (a) For an observer on the Earth, a mu meson (moving with a velocity of about 0.995c) would take about 6.4 μs to traverse the distance of 1920 m. (b) Inside the spaceship, the mu meson is at rest, and an observer inside the spaceship sees the Earth moving toward him with a speed 0.995c and a contracted distance of 192 m which is covered in 0.64 μs.

muon does not change and the events occur at the same place. Therefore if

and $\Delta t'$ = time interval in reference frame of muon
Δt = time interval in reference frame of Earth

then
$$\Delta t' = \sqrt{1-\frac{u^2}{c^2}}\,\Delta t$$

where u is the velocity of the mu meson as seen by an observer on the Earth. For $u \approx 0.995c$,

$$\sqrt{1-\frac{u^2}{c^2}} \approx 0.1$$

giving
$$\Delta t' \approx 0.1 \times 6.4\ \mu s = 0.64\ \mu s \qquad \textbf{(12)}$$

and therefore the number of muons that will undergo decay in 0.64 μs is

$$\approx 568 \exp\left(-\frac{0.64\ \mu s}{2.2\ \mu s}\right) \approx 425 \qquad \textbf{(13)}$$

which agrees very well with the observed value. Thus

> Whereas in the reference frame of the Earth, the time elapsed is 6.6 μs, in the reference frame of the muon (which is moving at a speed of 0.995c with respect to the Earth), the time elapsed is only 0.64 μs.

In an experiment at CERN (in Geneva) by Bailey and coworkers, muons of velocity 0.9994c were created by accelerating them in a circular path and were found to have a lifetime about 29 times the laboratory lifetime. This follows from the relation

$$\Delta t = \frac{\Delta t'}{\sqrt{1-u^2/c^2}} = \frac{\Delta t'}{\sqrt{1-(0.9994)^2}} \approx 29\Delta t'$$

Thus if we create two muon twins in the laboratory, one of them remains at rest and the other is accelerated to a speed of 0.9994c, then the muon (moving with a speed of 0.9994c) would come back to find its "twin" had undergone decay long, long time ago!

30.5 THE LENGTH CONTRACTION

We again consider two coordinate systems S and S' which are in uniform relative motion along the x axis as shown in Fig. 30.5. We have two persons A and B; person A is at rest in the coordinate system S, and person B (inside a moving train) is at rest in the coordinate system S'. Consider a rod RR' (of length L_0) at rest in the reference frame S. Now,

> The length of the rod L_0, measured in an inertial frame in which the rod is at rest, is known as the *proper length.*

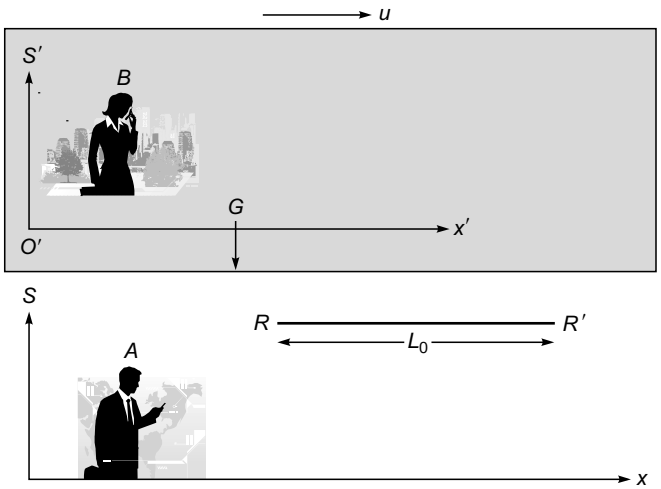


Fig. 30.5 Person A is on the platform and B is inside a train moving with velocity u in the $+x$ direction. A rod RR' (of length L_0) is at rest in the reference frame S. In the inertial frame S' (moving with velocity u with respect to the frame S), we have an observer B and an arrow G.

In the inertial frame S' (moving with velocity u with respect to the frame S), we have an observer B and an arrow G as shown in Fig. 30.5. We have two events: the first event is when the arrow G is in front of the end R of the rod, and the second event is when the arrow G is in front of the end R' of the rod.

Observer A in the inertial frame S sees the arrow move with velocity u, and if Δt is the time elapsed (as measured by A) for the arrow to go from the end R of the rod to the end R', then

$$L_0 = u\,\Delta t \qquad \textbf{(14)}$$

In the inertial frame S', observer B sees the rod moving with speed u in the $-x$ direction. Thus the length L of the rod as measured by B is given by

$$L = u\,\Delta t' \qquad \textbf{(15)}$$

where $\Delta t'$ is the time elapsed (as measured by B) as the ends R and R' of the rod cross the arrow. Now, $\Delta t'$ represents the time interval of the two events occurring at the same place G, and therefore it is the *proper time* and

$$\Delta t' = \sqrt{1-\frac{u^2}{c^2}}\;\Delta t$$

[see Eq. (6)]. Thus

$$L = \sqrt{1-\frac{u^2}{c^2}}\,L_0 \qquad \textbf{(16)}$$

Thus observer B measures a contracted length given by the above equation.

30.6 UNDERSTANDING THE MU MESON EXPERIMENT VIA LENGTH CONTRACTION

We revisit the mu meson experiment. For observer A (in the reference frame S at rest on Earth) the mu meson moves with velocity $0.995c$ and traverses the distance of 1920 m (the height of Mt. Washington) in about 6.6 μs [see Fig. 30.4(a)].

We next consider the mu meson inside a spaceship which is same velocity as the mu meson [see Fig. 30.4(b)]. Thus the mu meson is at rest inside the spaceship. For an observer B inside the spaceship, the Earth is moving toward it with velocity $u = 0.995c$. Because of length contraction, for the observer inside the spaceship, the distance between the top of Mt. Washington and the Earth is not 1920 m but the contracted distance of

$$\sqrt{1-\frac{u^2}{c^2}} \times 1920\ \text{m} \approx 0.1 \times 1920\ \text{m} = 192\ \text{m}$$

This distance is traversed (by the Earth) in only

$$\frac{192\ \text{m}}{0.995c} = \frac{192\ \text{m}}{0.995\times 3\times 10^8\ \text{m s}^{-1}} = 0.64\times 10^{-6}\ \text{s} = 0.64\ \mu\text{s}$$

And in this time, the number of muons that will undergo decay is

$$\approx 568\exp\left(-\frac{0.64\ \mu\text{s}}{2.2\ \mu\text{s}}\right) \approx 425$$

which is the same as Eq. (13) and agrees well with the observed value.

30.7 LENGTH CONTRACTION OF A MOVING TRAIN

Consider a mirror M placed inside a train (moving with speed u) as shown in Fig. 30.6. A pulse of light emitted from the light source P gets reflected by the mirror and is detected at D. Obviously, the time interval (as measured by observer B in the moving train) between the emission of light and its subsequent detection is given by

$$\Delta t' = \frac{2L_0}{c} \qquad \textbf{(17)}$$

where L_0 is the distance between the light source and the mirror as measured by observer B in the moving train.

We next consider the events as seen by a person on the platform; for him, the speed of light is the same, and we assume that the distance between point P and mirror M is L.

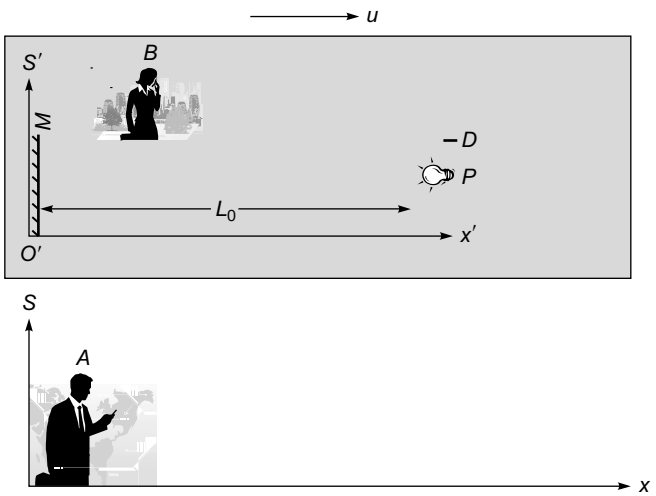


Fig. 30.6 A mirror M is placed inside a moving train. A pulse of light emitted from the light source P gets reflected by the mirror and is detected at D. For observer B, the mirror and the source of light are stationary and are at a distance L_0 from each other.

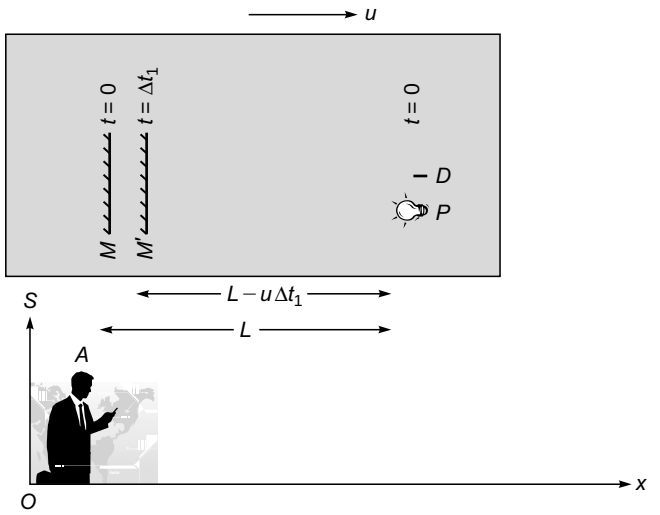


Fig. 30.7 Observer A on the platform sees a contracted distance L between P and the mirror (inside the moving train). The pulse of light emitted from the light source P reaches the mirror at $t = \Delta t_1$, and by then the mirror has moved through the distance $u\,\Delta t_1$.

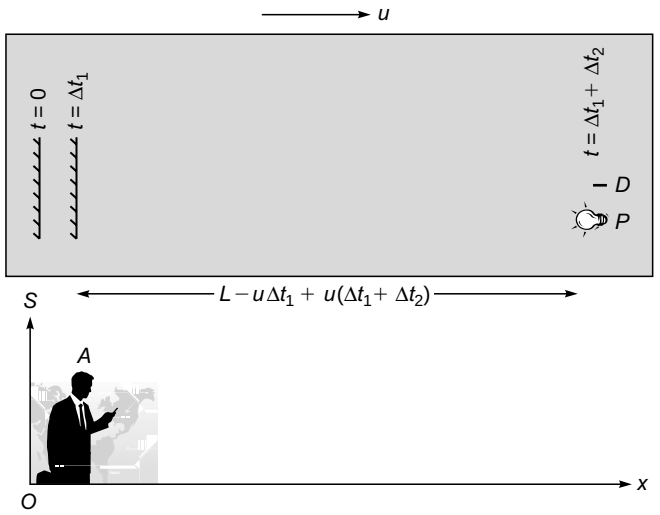


Fig. 30.8 For observer A on the platform, the light beam reflected from the mirror reaches the detector D at $t = \Delta t_1 + \Delta t_2$.

If Δt_1 represents the time interval (as observed by person A on the platform) for light to travel to the mirror, then

$$\Delta t_1 = \frac{L - u\,\Delta t_1}{c} \quad \Rightarrow \quad \Delta t_1 = \frac{L}{c(1+u/c)} \tag{18}$$

where we have taken into account the fact that in time Δt_1 the mirror has moved through a distance $u\,\Delta t_1$ (see Fig. 30.7). Similarly, if Δt_2 represents the time interval (as observed by person A on the platform) for light to travel from the mirror to the detector, then (see Fig. 30.8).

$$\Delta t_2 = \frac{L - u\,\Delta t_1 + u\left(\Delta t_1 + \Delta t_2\right)}{c} \quad \Rightarrow \quad \Delta t_2 = \frac{L}{c(1-u/c)} \tag{19}$$

Thus if Δt represents the time interval (as observed by the person on the platform) between the emission of light and its subsequent detection, then it is given by

$$\Delta t = \Delta t_1 + \Delta t_2 = \frac{2L}{c(1-u^2/c^2)} \tag{20}$$

Now, since

$$\Delta t' = \frac{2L_0}{c} \tag{21}$$

represents the time interval between two events occurring *at the same place* inside the moving train, $\Delta t'$ and Δt are related by

$$\Delta t' = \sqrt{1-\frac{u^2}{c^2}}\,\Delta t \tag{22}$$

Using Eqs. (20) to (22), we immediately get

$$L = \sqrt{1-\frac{u^2}{c^2}}\,L_0 \tag{23}$$

Thus because the speed of light is the same in all inertial frames, the observer on the platform will calculate a shorter length of the train.

30.8 SIMULTANEITY OF TWO EVENTS

We next consider an atom (at rest in the moving train) emitting simultaneously two photons. Two detectors D_1 and D_2 are at the same distance (= L_0) from the atom, therefore, for observer B in reference frame S', the photons are detected simultaneously (see Fig. 30.9). For observer A on the platform, let Δt_1 and Δt_2 represent the time taken by the two photons to reach the two detectors D_1 and D_2, respectively. Using the arguments given in the previous section

$$\Delta t_1 = \frac{L}{c(1+u/c)} \tag{24}$$

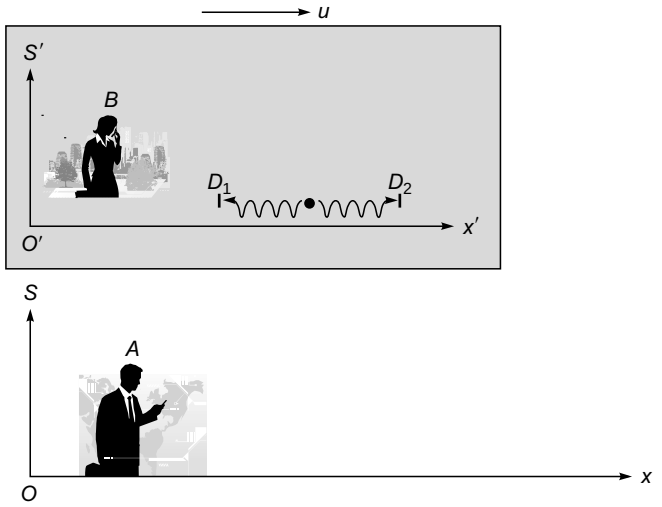


Fig. 30.9 For observer *B* on the moving train, the atom is at rest and emits two photons simultaneously in opposite directions. The two detectors D_1 and D_2 are equidistant from the atom and detect the photons simultaneously; however, the two events are not simultaneous for observer *A*.

and
$$\Delta t_2 = \frac{L}{c(1-u/c)} \qquad \textbf{(25)}$$

where L is the contracted distance between the atom and either of the detectors. Thus in the reference frame *S*, the time difference between the two events will be

$$\Delta t = \Delta t_2 - \Delta t_1 = \frac{L}{c\left(1-\frac{u}{c}\right)} - \frac{L}{c\left(1+\frac{u}{c}\right)}$$

$$= \frac{2u\gamma}{c^2} L_0$$

where we have used Eq. (23). Thus whereas the two events are simultaneous in the reference frame *S′*, they are not simultaneous in the reference frame *S*. We will re-derive this result again in Example 31.6.

30.9 THE TWIN PARADOX

The star closest to us is Proxima Centauri, and it is about 4.2 light-years away; i.e., a light beam will take about 4.2 years to travel from Earth to the star. Now

$$1 \text{ yr} = 365 \times 24 \times 60 \times 60 \text{ s} \approx 3.15 \times 10^7 \text{ s}$$

Thus the distance of the star is

$$D \approx 4.2 \times 3.15 \times 10^7 \times 3 \times 10^8 \text{ m}$$
$$\approx 4 \times 10^{16} \text{ m} = 40 \text{ trillion km}$$

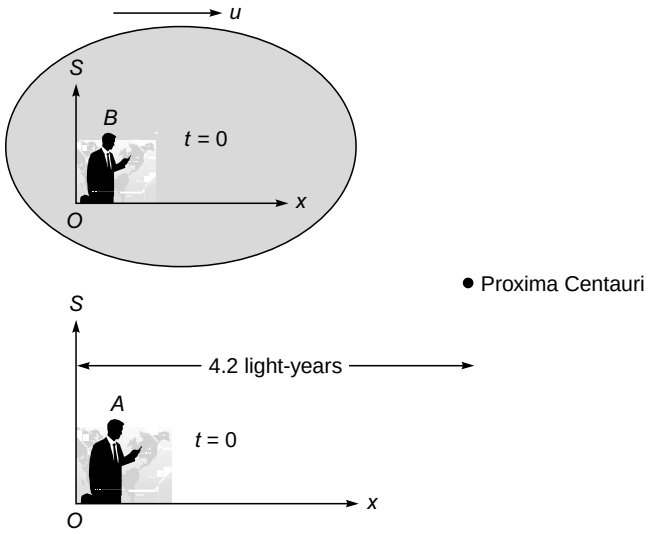


Fig. 30.10 Person *A* is on the Earth and *B* is inside a spaceship; when *O* coincides with *O′*, the clocks are synchronized so that $t' = t = 0$. Person *B* quickly accelerates, attains the velocity *u*, and moves toward the star Proxima Centauri.

We consider the following experiment:

A (say Arjun) and *B* (say Bob) are twins. Bob enters a spacecraft (see Fig. 30.10) and synchronizes his watch with Arjun ($t = t' = 0$). The spacecraft closes and quickly accelerates to a velocity 0.995*c*. Bob is at rest inside the aircraft. According to Arjun (who is on Earth), Bob will take

$$\frac{4 \times 10^{16}}{0.995 \times 3 \times 10^8} \approx 1.3 \times 10^8 \text{ s} \approx 4.2 \text{ years}$$

to reach the Proxima Centauri star. According to Bob, the star is moving toward him with a velocity 0.995*c*, and he will see the contracted distance given by

$$\sqrt{1-\frac{u^2}{c^2}} \times D = \sqrt{1-(0.995)^2}\ D \approx 0.1D \approx 0.42 \text{ light-year}$$

Thus according to Bob, he will reach the star in 0.42 yr which is one-tenth of the time recorded by Arjun (see Fig. 30.11). Bob returns to Earth at the same speed. On his return journey, he finds that the Earth is moving towards him with a velocity 0.995*c* and the contracted distance is be ≈0.42 light-year. When Bob's spaceship stops, he finds that his clock shows only 0.84 yr whereas Arjun's clock would show 8.4 yr (see Fig. 30.12).

We can understand the above situation from another point of view. The first event corresponds to when the spaceship starts moving (with velocity *u*) from the Earth, and the second event corresponds to when the spaceship reaches the star. In the moving frame (i.e., inside the

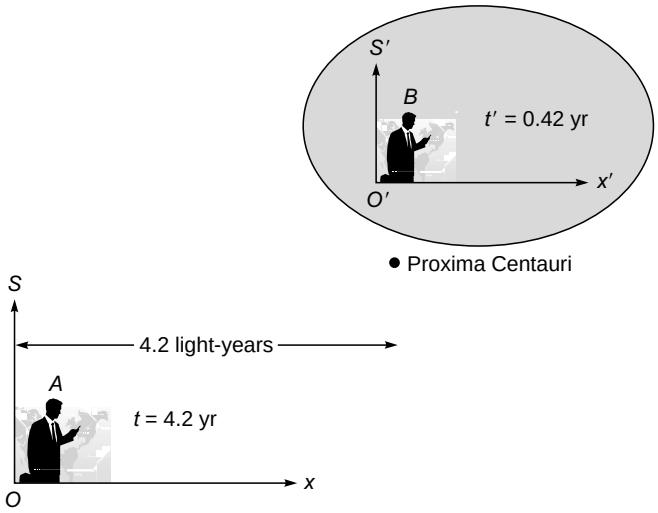


Fig. 30.11 When *B* reaches Proxima Centauri, to him it has been only 0.42 yr, but for the observer *A* it would be 4.2 yr.

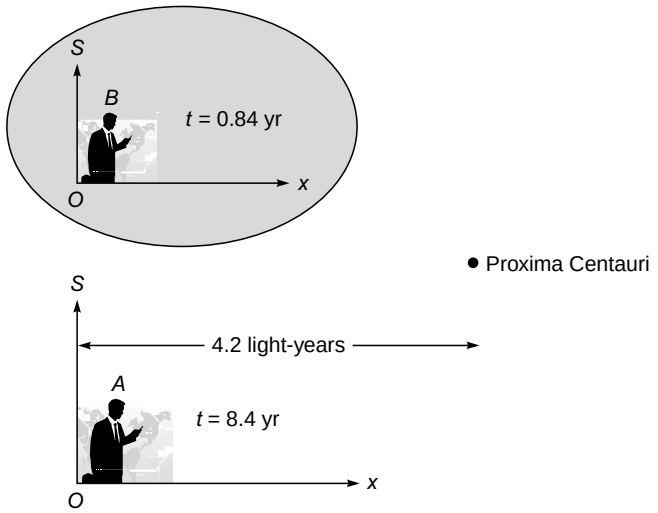


Fig. 30.12 Person *B* returns to Earth at the same speed. When his spaceship stops, he finds that his clock shows only 0.84 yr. On the other hand, *A*'s clock shows a lapse of 8.4 yr.

spaceship) both events occur at the same space point. Thus the time interval $\Delta t'$ measured by Bob is the proper time and will be less than the time measured by Arjun by a factor $\sqrt{1-u^2/c^2}$ (see the discussion in Sec. 30.2).

We next consider the case when the star was 42 light-years away and the spaceship was traveling with the same speed. Let us assume that, to start, both Arjun and Bob were 20 years old. When Bob returns to Earth, Arjun will be about 104 years old and Bob will be only 28.4 years old. Thus Arjun would have aged significantly!

The above experiment has led to a lot of controversy—scientists have argued that according to Arjun, Bob was moving with a velocity 0.995*c* and according to Bob, Arjun was moving with a velocity 0.995*c* (in the opposite direction). But there is really no controversy when we consider that we must always be careful to define the "proper time," and when Bob returns from his space journey, he *will* be younger to Arjun. Also it is Bob who undergoes acceleration (and deceleration) and because of this the motions of Arjun and Bob are not symmetrical.

30.10 THE MICHELSON–MORLEY EXPERIMENT

In the beginning of the nineteenth century, a few very beautiful experiments were carried out which demonstrated the interference and diffraction phenomena of light. Both interference and diffraction phenomena could only be explained by assuming a wave model of light. However, it was believed that a wave would always require a medium and since light could propagate through vacuum, the presence of an "all pervasive" medium called the ether was assumed.

If we assume the existence of this "all pervasive" ether, then the observed velocity of light would change if we move with respect to the ether. We know that the Earth moves around the Sun in an approximately circular orbit with a speed of about 30 km/s (see Fig. 30.13). Thus we should expect that, whatever may be the motion of the solar system, during a certain period of time in a year, the Earth will be moving with respect to the ether with a speed of at least 30 km/s and experience what is often referred as "the ether wind."

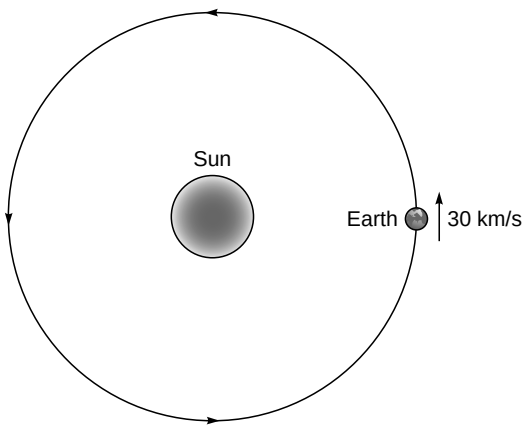


Fig. 30.13 The Earth rotates around the Sun in an approximately circular orbit with a speed of about 30 km/s.

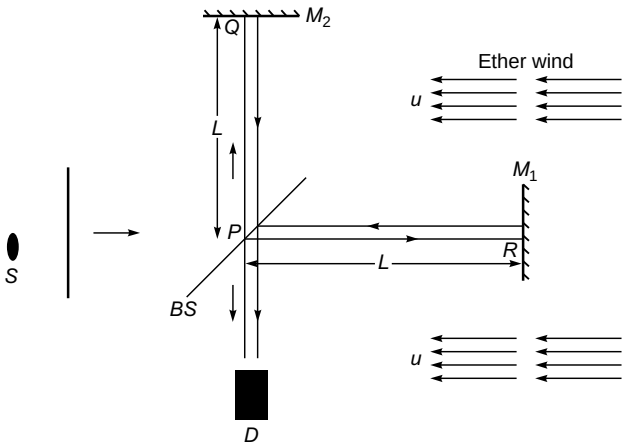


Fig. 30.14 The Michelson interferometer arrangement. An observer at rest with the interferometer, experiences the ether wind as shown above.

The experiment involved the famous Michelson interferometer shown in Fig. 30.14 (see Sec. 15.11). The beam splitter (shown as *BS* in the figure) splits the light beam into two beams traveling at right angles to each other. Subsequently the beams get reflected by M_1 and M_2 and the beam reflected from M_1 gets reflected by the beam splitter and superposes with the beam reflected from M_2 to form an interference pattern. We assume the positions of the two mirrors to be such that the distance of the beam splitter to each of the mirrors is exactly the same and equal to *L*.

If the whole interferometer is at rest with respect to the ether then the light would travel with the same velocity in all directions and therefore the light reflected by the mirrors M_1 and M_2 would reach the detector *D* at the same time.

We next assume that the apparatus moving through the ether so that with respect to the interferometer, the ether is moving to the left with velocity *u* as shown in Fig. 30.14. Thus as the light beam travels from *P* to *R*, it is opposed by the ether wind and its velocity is $c - u$. On the other hand, when the light beam travels from to *R* to *P*, it is carried by the ether wind[4] and its velocity is $c + u$. Thus if t_{PR} and t_{RP} are the time taken for the outward and return trips then

$$t_{PR} = \frac{L}{c-u} \qquad \text{and} \qquad t_{RP} = \frac{L}{c+u}$$

Therefore the total time for the light beam to travel from *P* to *R* and back will be given by

$$t_1 = t_{PR} + t_{RP} = \frac{2L}{c}\frac{1}{1-\frac{u^2}{c^2}} \qquad \textbf{(26)}$$

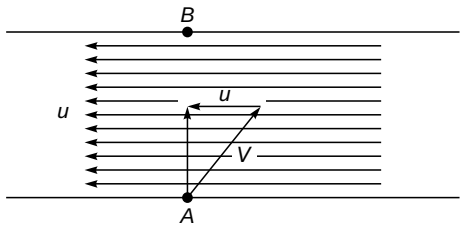


Fig. 30.15 There is a current in the river. A swimmer is trying to cross from the point A to the point B. The swimmer must try to swim slightly to the right so that his trajectory is straight.

We next consider the light beam reflected by the mirror M_2. This case is similar to a ship trying to cross from the point A to the point B when there is current in the river (see Fig. 30. 15). Obviously the ship must point slightly to the right so that its trajectory is straight. In the absence of the current if the speed was *V* then the actual speed will be $\sqrt{V^2 - u^2}$. Thus the effective speed of the light beam for the path *PQ* will be $\sqrt{c^2 - u^2}$. Similarly, the effective speed of the light beam for the return path *QP* will also be $\sqrt{c^2 - u^2}$. Thus if t_{PQ} and t_{RP} are the time taken for the outward and return trips then

$$t_{PQ} = \frac{L}{\sqrt{c^2 - u^2}} \quad \text{and} \quad t_{QR} = \frac{L}{\sqrt{c^2 - u^2}}$$

Therefore the total time for the light beam to travel from *P* to *Q* and back will be given by

$$t_2 = t_{PQ} + t_{QR} = \frac{2L}{\sqrt{c^2 - u^2}} = \frac{2L}{c}\frac{1}{\sqrt{1-\frac{u^2}{c^2}}} \qquad \textbf{(27)}$$

Thus

$$t_2 = \sqrt{1-\frac{u^2}{c^2}}\, t_1 \qquad \textbf{(28)}$$

and t_2 will always be less than t_1. The time difference $t_1 - t_2$ will be given by

$$\Delta t = t_1 - t_2 = \frac{2L}{c}\left[\frac{1}{1-\frac{u^2}{c^2}} - \frac{1}{\sqrt{1-\frac{u^2}{c^2}}}\right] \qquad \textbf{(29)}$$

[4] The easiest way to understand this is to first consider a ship which travels with a certain velocity *V* in still water. We next assume that there is a current in the water moving with speed *u*. If the ship travels along the current, its velocity will increase to $V + u$. On the other hand, if the ship travels against the current, its velocity will decrease to $V - u$.

When $\frac{u}{c} << 1$, the quantity within the square brackets will be

$$\frac{1}{1-\frac{u^2}{c^2}} - \frac{1}{\sqrt{1-\frac{u^2}{c^2}}} = \left(1-\frac{u^2}{c^2}\right)^{-1} - \left(1-\frac{u^2}{c^2}\right)^{-\frac{1}{2}}$$

$$\approx \left(1+\frac{u^2}{c^2}\right) - \left(1+\frac{1}{2}\frac{u^2}{c^2}\right) = \frac{1}{2}\frac{u^2}{c^2}$$

Thus, when $\frac{u}{c} << 1$,

$$\Delta t \approx \frac{L u^2}{c^3} \qquad \textbf{(30)}$$

This will correspond to a path difference of

$$c\,\Delta t \approx \frac{L u^2}{c^2} \qquad \textbf{(31)}$$

If we now rotate the interferometer by exactly 90°, the two beams will exchange their time of traversals and therefore the fringe shift will correspond to twice the path difference given above. Thus the effective path difference will be

$$2c\,\Delta t \approx \frac{2L u^2}{c^2} \qquad \textbf{(32)}$$

Now a path difference of λ results in shift of one fringe; thus the fractional fringe shift will be

$$\frac{2c\,\Delta t}{\lambda} \approx \frac{2L u^2}{\lambda c^2}$$

In one of the experiments carried by Michelson and Morley $L \approx 11\,\text{m}$, $\lambda \approx 6\times10^{-7}\ \text{m}$ and if we assume that the relative velocity of the ether is at least the velocity of the Earth (i.e., $u \approx 3\times10^4\ \text{km/s}$), we get

$$\frac{2L u^2}{\lambda c^2} \approx \frac{2\times(11\,\text{m})\times\left(3\times10^4\,\text{m/s}\right)^2}{\left(6\times10^{-7}\ \text{m}\right)\times\left(3\times10^8\,\text{m/s}\right)^2} \approx 0.4$$

Thus a shift of about 0.4 fringe should have been observed. During 1881–1887, Professor Michelson (along with his colleague Edward Morley) carried out a series of very careful measurements for different orientations of the interferometer and they always got a null result; their apparatus was capable of detecting 0.01 fringe shift. These came to be known as the famous Michelson–Morley experiments which proved that the ether did not exist. David Park (in his book *The Fire within the Eye: A Historical Essay on the Nature and Meaning of Light*) has written *"He* (Michelson) *was 34 when he established that ether cannot be found; he made delicate optical measurements for 44 more years and to the end of his days did not believe there could be a wave without some material substance to do the waving."*

30.11 BRIEF HISTORICAL REMARKS

When Einstein wrote the famous 5 papers in 1905, he was working in Swiss Patent office and did not have much discussion with other physicists—he studied on his own and it appears he was not aware of the Michelson–Morley experiments. Casper and Noer (Ref. 8) have made a careful study of the history and they write:

> Einstein was then an unknown physicist, largely self-taught in this area of physics, and in his patent office job somewhat cutoff from the discussions and ideas current in the physics community. He was apparently only vaguely aware of the ether experiments . . . and he did not know of the later papers of Lorentz and Poincaré. . . .

Einstein was of course aware of Maxwell's equations and laws of electricity and magnetism and the fact that Maxwell's equations were not invariant under Galilean transformation. Physically this implies that if Galilean transformation was correct then (to quote from Ref. 6)

> in a moving space ship the electrical and optical phenomena should be different from those in a stationary ship. Thus one could use these optical phenomena to determine the speed of the ship; in particular one could determine the absolute speed of the ship by making suitable optical and electrical measurements

Thus, for Einstein, the fundamental question was (to quote from Ref. 8)

> "Why should the laws of electromagnetism and light, alone among the laws of physics, allow the possibility of detecting the motion of an inertial reference frame?"

Einstein started his 1905 paper by writing

> It is well known that Maxwell's electrodynamics—as usually understood at present—when applied to moving bodies, lead to asymmetries that do not seem to be inherent in the phenomena.

He further wrote (in the same paper)

> The same laws of electrodynamics and optics will be valid for all coordinate systems . . . We shall raise this conjecture to the status of a postulate and shall also introduce another postulate, namely that light always propagates in free space with a definite velocity V that is independent of the state of motion of the emitting body.

The null result of the Michelson–Morley experiment is consistent with Einstein's postulates. Indeed Einstein wrote

> The introduction of a 'light ether' will prove to be superfluous inasmuch as the view to be developed here will not require a 'space at absolute rest' endowed with special properties.

Problems

30.1 At a height of about 3 km above sea level, about 1000 mu mesons were detected in 1 h. Calculate the number that will decay before they reach sea level. Assume that the mean lifetime of the mu meson is about 2.2 μs and that its velocity is about 0.9c.

30.2 A spacecraft of "proper length" 10 meters is passing by with a speed of 30 km s^{-1}. (a) Calculate the contracted length observed by a person on the Earth. (b) Calculate the contracted length if the velocity of the spacecraft was 0.99 c.

[**Ans:** (a) 9.99999995 m (b) 1.41 m]

30.3 A and B are twins. Person B enters a spacecraft (see Fig. 30.10) and synchronizes his watch with A ($t = t' = 0$). The spacecraft closes and quickly accelerates to a velocity given by

$$u = \sqrt{\frac{15}{16}}\, c \approx 0.9682c$$

The spacecraft goes to a nearby star 10 light-years away and promptly returns to Earth with the same speed. What will be the age difference between A and B?

30.4 A and B are twins. Person B enters a spacecraft (see Fig. 30.10) and synchronizes his watch with A ($t = t' = 0$). The spacecraft closes and quickly accelerates to the same velocity as in Prob. 30.3 ($\approx 0.9682c$). The spacecraft goes to the Moon (which is about 384,000 km away) and promptly returns to Earth with the same speed. What will be the age difference between A and B?

30.5 Consider an atom (at rest inside a spaceship moving with velocity 3 km s^{-1}) emitting simultaneously two photons; the two detectors D_1 and D_2 are at the same distance (= 10 m) from the atom, and therefore for an observer B inside the spaceship, the photons are detected simultaneously (see Fig. 30.7). For the observer A on Earth, calculate the time difference between the two events and compare with the result obtained by using the formulas developed in Chap. 31.

REFERENCES AND SUGGESTED READINGS

1. A. Einstein, "On the Electrodynamics of Moving Bodies," *Annalen der Physik*, Vol. 17, pp. 891–921, 1905.
2. T. Alvager et al., *Physics Letters*, Vol. 12, p. 260, October 1, 1964.
3. K. Brecher, "Is the Speed of Light Independent of the Velocity of the Source?" *Phys. Rev. Letts.* Vol. 39, pp. 1051–1054, 1977.
4. J. Stachel (Ed.), *Einstein's Miraculous Year: Five Papers That Changed the Face of Physics*, Princeton University Press, 1998. Reprinted by Srishti Publishers, New Delhi.
5. http://members.tripod.com/~gravitee/axioms.htm.
6. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley Publishing Co., Reading, Mass., 1963.
7. R. Baierlein, *Newton to Einstein: The Trail of Light*, Cambridge University Press, United Kingdom, 1992.
8. B. M. Casper and R. J. Noer, *Revolutions in Physics*, W.W. Norton & Co., New York, 1972.
9. B. Rossi and D. B. Hall, "On Muon Time Dilation," *Physical Review*, Vol. 59, p. 223, 1941. In 1963 David Frisch and James Smith repeated the Mt. Washington experiment and reported their measurements in the paper entitled "Measurement of the Relativistic Time Dilation Using Mesons," *Am. J. Phys.*, Vol. 31, p. 342, 1963.
10. http://www.egglescliffe.org.uk/physics/relativity/muons1_.htm.

Chapter Thirty-one

SPECIAL THEORY OF RELATIVITY II: MASS-ENERGY RELATION AND LORENTZ TRANSFORMATIONS

Politics is for the moment, ... while an equation is for eternity.
—Albert Einstein

31.1 INTRODUCTION

In this chapter we will continue our discussions on the consequences of the two postulates of the special theory of relativity and derive the famous equation describing the mass-energy relationship

$$E = mc^2 \tag{1}$$

We will also derive expressions for Doppler shift and what are known as Lorentz transformations. Using the equations describing Lorentz transformations, we will rederive some of the results obtained in Chap. 30.

31.2 MASS-ENERGY RELATION

In this section we will carry out a very simple and straightforward derivation of the mass-energy relation. The analysis is somewhat similar to that given in Ref. 1 (see also Refs. 2 to 4).

A_1 and A_2 are in reference frame S, and B is in reference frame S' which is moving with respect to reference frame S with velocity u. An atom is at rest in reference frame S'. The atom emits two photons (of the same frequency ν_0) in opposite directions as shown in Fig. 31.1. Thus, in reference frame S', the total momentum of the two photons will be zero, and from the law of conservation of momentum, the atom will remain at rest (in reference frame S') . The change in the energy of the atom (as observed by B in reference frame S') will be given by

$$(\Delta E)_{S'} = 2h\nu_0 \tag{2}$$

A_1 and A_2 are both in reference frame S; for A_1 the source is moving away from the observer, and for A_2 the source is moving toward the observer. As such, A_1 and A_2 will observe different Doppler shifted frequencies ν_1 and ν_2 given by (see Sec. 31.3)

$$\nu_1 = \nu_0 \sqrt{\frac{1-u/c}{1+u/c}} \tag{3}$$

and

$$\nu_2 = \nu_0 \sqrt{\frac{1+u/c}{1-u/c}} \tag{4}$$

Thus, the change in the energy of the atom (as observed in reference frame S) will be given by

$$(\Delta E)_S = h\nu_1 + h\nu_2 = \frac{2h\nu_0}{\sqrt{1-u^2/c^2}} \tag{5}$$

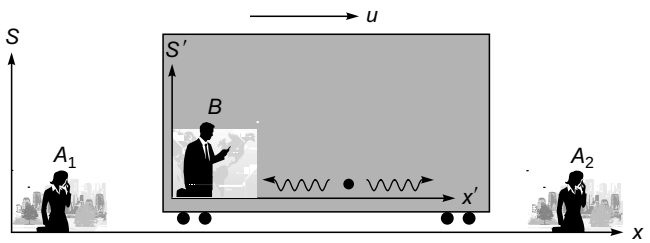


Fig. 31.1 A_1 and A_2 are at rest in reference frame S. B is at rest in reference frame S' which is moving with speed u with respect to frame S. In reference frame S', an atom is at rest and emits two photons of the same frequency ν_0.

As mentioned earlier, in reference frame S', the atom (after the emission of two photons) will remain at rest. Thus, for an observer in reference frame S, the atom will be moving with velocity u before and after the emission of photons. In reference frame S, since $\nu_2 > \nu_1$, the momentum of the two photons is different. Therefore if we use the law of conservation of momentum in reference frame S, the atom (which is moving with the same velocity u) must have a slightly lesser mass given by

$$(\Delta m)_S u = \frac{h\nu_2}{c} - \frac{h\nu_1}{c} = \frac{2h\nu_0}{\sqrt{1-u^2/c^2}}\frac{u}{c^2}$$

$$= \frac{(\Delta E)_S u}{c^2} \qquad \textbf{(6)}$$

Thus we get the relation

$$(\Delta E)_S = (\Delta m)_S c^2 \qquad \textbf{(7)}$$

In the above equation, ΔE and Δm need not be infinitesimal amounts, and therefore we get the mass-energy relation (see Ref. 2)

$$E = mc^2$$

In his 1905 paper entitled "Does the Inertia of a Body Depend on Its Energy Content?" Einstein wrote (Ref. 4):

> If a body emits the energy L in the form of radiation, its mass diminishes by L/c^2.... The mass of a body is a measure of its energy content.

Thus when a hydrogen atom makes a transition from an excited state to the ground state with the emission of a photon, the mass of the hydrogen atom (which still consists of one proton and one electron) will decrease by a small amount. In general, whenever a loosely bound system goes over to a tightly bound system, a small amount of mass gets converted to energy.

Further, if we write

$$(\Delta E)_{S'} = (\Delta m)_{S'} c^2 \qquad \textbf{(8)}$$

then from Eqs. (2) and (5) we get

$$(\Delta m)_S = \frac{(\Delta m)_{S'}}{\sqrt{1-u^2/c^2}} \qquad \textbf{(9)}$$

Thus the mass varies with velocity according to the following equation:

$$m = \gamma m_0 = \frac{m_0}{\sqrt{1-u^2/c^2}} \qquad \textbf{(10)}$$

where m_0 is the mass of the body when it is at rest and is usually referred to as the *rest mass* and

$$\gamma = \frac{1}{\sqrt{1-u^2/c^2}} \qquad \textbf{(11)}$$

is the Lorentz factor. About Eq. (10), Feynman has written

> For those who want to learn just enough about it so that they can solve problems, that is all there to the theory of relativity—it just changes Newton's laws by introducing a correction factor to the mass.

The momentum of a body of rest mass m_0 moving with velocity u will be given by

$$p = mu = \frac{m_0 u}{\sqrt{1-u^2/c^2}} \qquad \textbf{(12)}$$

Further the kinetic energy of a particle of rest mass m_0 will be given by

$$T = mc^2 - m_0c^2$$

$$= m_0c^2\left(\frac{1}{\sqrt{1-u^2/c^2}} - 1\right) \qquad \textbf{(13)}$$

When $\frac{u}{c} << 1$,

$$\gamma = \frac{1}{\sqrt{1-u^2/c^2}} \approx 1 + \frac{u^2}{2c^2} \qquad \textbf{(14)}$$

and

$$K \approx m_0c^2\left(1 + \frac{u^2}{2c^2} - 1\right) = \frac{1}{2}m_0u^2 \qquad \textbf{(15)}$$

which is the nonrelativistic expression for the kinetic energy of a particle.

Example 31.1 Consider a body with rest mass 50 kg. If we have to make it move with a velocity $0.9c$, the Lorentz factor will be ≈ 2.3 and therefore the kinetic energy will be

$$T \approx (50 \text{ kg}) \times (3 \times 10^8 \text{ m s}^{-1})^2 \times 1.3 \approx 6 \times 10^{18} \text{ J}$$

This is an enormous amount of energy; for example, a 100 MW power station will generate 6×10^{18} J of energy in about 2000 years.[1] Even if we have to make the mass move with a

[1] $(100 \times 10^6 \text{ W}) \times (2000 \times 3.1 \times 10^7 \text{ s}) \approx 6 \times 10^{18}$ Joules where we have assumed 1 year $\approx 3.1 \times 10^7$ s

velocity 0.5c, the factor 1.3 will be replaced by 0.15 and the kinetic energy will be

$$T \approx 0.7 \times 10^{18} \text{ J}$$

which is also an enormous amount of energy! Thus it would require a tremendously large amount of energy to make a spacecraft (which would have a much larger rest mass) move close to the speed of light.

Example 31.2 The rest mass of the proton is $m_p \approx 1.67 \times 10^{-27}$ kg, thus the rest mass energy of the proton is given by

$$m_p c^2 \approx (1.67 \times 10^{-27}\text{kg}) \times (3 \times 10^8 \text{ m/s})^2 \approx 1.5 \times 10^{-10} \text{ J} \approx 938 \text{ MeV}$$

In the large hadron collider (LHC), the protons are accelerated to about 99.9999991% of the speed of light (see, e.g., Ref. 5). Thus

$$\frac{u}{c} \approx 0.999999991$$

and the Lorentz factor will be given by

$$\gamma = \frac{1}{\sqrt{1 - u^2/c^2}} \approx 7500$$

Thus the kinetic energy of the proton will be

$$T \approx 7000 \text{ GeV}$$

Example 31.3 **Energy from the Sun** The outer periphery of Earth's atmosphere receives, from the Sun, about 1.4 kW m^{-2} of energy; this would imply that about 1400 J of radiation is received (per second) on 1 m^2 of area placed perpendicularly to the light beam coming from the Sun. The distance between the Earth and the Sun is about 1.5×10^{11} m (it takes about 8.5 min for the light to travel from Sun to Earth). If we assume that the solar energy spreads out uniformly in all directions, then the total energy liberated from the Sun is about[2] $1400 \times \pi \times (1.5 \times 10^{11})^2 \cong 4 \times 10^{26}$ J s^{-1}. If we now use Einstein's mass-energy relation $E = mc^2$, we get the result

$$m = \frac{E}{c^2} = \frac{4 \times 10^{26} \text{ J s}^{-1}}{(3 \times 10^8 \text{ m s}^{-1})^2} \approx 4 \times 10^9 \text{ kg s}^{-1} \quad \textbf{(16)}$$

Thus every second about 4 billion kg of mass is continuously getting converted to energy.

31.3 THE DOPPLER SHIFT

In astronomy, we can determine how fast the stars or galaxies are moving (either directly away or directly toward us) by measuring the Doppler shift of spectral lines. When the star is moving away from the observer, the measured frequency is slightly less than the actual value leading to the well-known *red shift* of spectral lines. It is this Doppler shift that we calculate in this section.

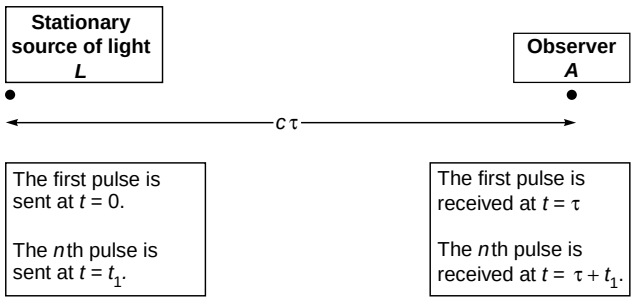


Fig. 31.2 Here L is a stationary light source. Light takes time τ to reach observer A.

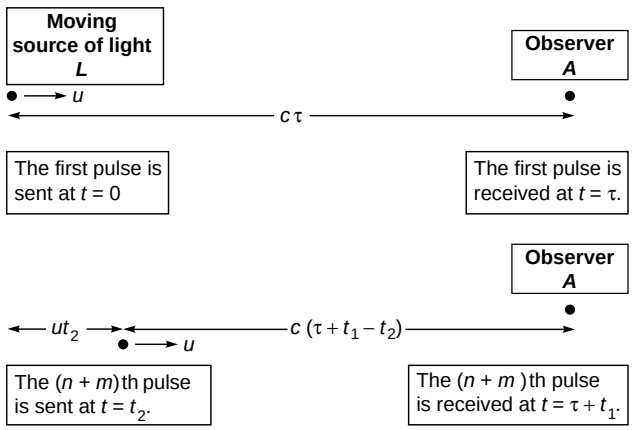


Fig. 31.3 The light source L is moving toward observer A with velocity u. The first pulse is sent at $t = 0$ which is received by A at $t = \tau$. According to A, the $(n + m)$th pulse is sent at $t = t_2$ which is received by A at $t = \tau + t_1$.

We will follow the method put forward by Feynman (see Sec. 34-6 of Ref. 5). Let us consider a light source which is at rest with respect to observer A. Instead of a continuous wave, we assume that the light source emits ν_0 pulses in 1 s. We assume that the first pulse is sent at $t = 0$ and the nth pulse is sent at $t = t_1$; thus $n = \nu_0 t_1$. The time taken by each pulse to reach A is assumed to be τ; thus the first pulse is received at $t = \tau$, and the nth pulse is received at $t = \tau + t_1$; obviously, the distance between the source and observer is $c\tau$ (see Fig. 31.2).

We next assume the source to be moving with velocity u toward the observer as shown in Fig. 31.3. In time t_1, observer A will now receive a larger number of pulses because as the source moves toward the observer, the light pulse takes less time to

[2] (This is indeed an enormous amount of energy and is equivalent to the detonation of about 100 billion megatons of TNT every second; one ton of TNT is a unit of energy equal to 1 billion (= 10^9) calories equal to about 4.2×10^9 J. One of the largest power plants on earth produces about 6000 MW (= 6×10^9 J s^{-1}) of energy—thus a total of 2×10^{17} J of energy is given out every year (1 yr is about 31 million s); this would imply that about 2 billion such power plants would produce about the same energy in *one year* that the Sun produces in 1 s!

reach the observer. If in time t_1 the observer receives $n + m$ pulses, and if the source has moved through the distance ut_2 when the source emits the $(n + m)$th pulse at $t = t_2$ (see Fig. 31.3), then

$$c\tau = ut_2 + c(\tau + t_1 - t_2) \qquad \Rightarrow \qquad \frac{t_1}{t_2} = 1 - \frac{u}{c} \tag{17}$$

Both t_1 and t_2 are as measured by observer *A*. Now an observer *B* (who is moving with the atom) will observe the duration t_2 as

$$t_2' = t_2\sqrt{1-\frac{u^2}{c^2}} \tag{18}$$

Thus the number of pulses received by observer *A* (during the time interval $t = \tau$ and $t = \tau + t_1$) is $\nu_0 t_2'$, and therefore the observed frequency will be

$$\nu_1 = \frac{\nu_0 t'_2}{t_1} = \nu_0 \frac{t_2\sqrt{1-u^2/c^2}}{t_1}$$

Using Eq. (17), we get

$$\nu_1 = \nu_0 \frac{\sqrt{1-u^2/c^2}}{1-u/c} = \nu_0\sqrt{\frac{1+u/c}{1-u/c}} \tag{19}$$

which is the Doppler shifted frequency observed by *A*. Thus we may write for the Doppler shifted (DS) frequency

$$\nu_{DS} = \nu_0\sqrt{\frac{1\pm\frac{u}{c}}{1\mp\frac{u}{c}}} \qquad \text{(Longitudinal Doppler Effect)} \tag{20}$$

where the upper sign corresponds to the source moving toward the observer and the lower sign corresponds to the source moving away from the observer. The corresponding Doppler shifted wavelengths are given by

$$\lambda_{DS} = \lambda_0\sqrt{\frac{1\mp\frac{u}{c}}{1\pm\frac{u}{c}}} \tag{21}$$

Equations (20) and (21) correspond to what is known as the *longitudinal Doppler effect* because the source is moving along the line joining the source and the observer. When

$$\frac{u}{c} << 1$$

we get the nonrelativistic expressions for the Doppler shift:

$$\lambda_{DS} \approx \lambda_0\left(1\pm\frac{u}{c}\right) \tag{22}$$

The fractional change in the wavelength is approximately given by

$$\frac{\Delta\lambda}{\lambda_0} \approx \pm\frac{u}{c} \tag{23}$$

If a star is moving away from us, we must take the lower sign and the wavelength increases—this is known as the *red shift* of spectral lines.

Example 31.4 A distant galaxy is moving away from us at a speed of about 60,000 km s^{-1}. Thus

$$\frac{u}{c} \approx 0.2$$

and the Doppler wavelength will be red shifted by about 22%.

Example 31.5 According to Hubble's law, the greater the distance of the galaxy, the greater the velocity of the galaxy moving away from us. Thus if u represents the velocity of the galaxy, then

$$u \approx HD \tag{24}$$

where D is the distance from the galaxy; the parameter H is known as the Hubble parameter and

$$H \approx 15\text{–}30 \text{ km s}^{-1} \text{ per million light-years}$$

There is a lot of controversy on the validity of Hubble's law. However, if we assume $H \approx 20$ km s^{-1} per million light-years, the above equation implies that if a galaxy is about 150 million light-years away, then it will be moving away from us with a speed of about 3000 km s^{-1} and

$$\frac{u}{c} \approx 0.01$$

Thus the fractional increase in the Doppler shifted wavelength is about 1%.

As mentioned earlier, Eqs. (20) and (21) correspond to what is known as the *longitudinal Doppler effect* because the source is moving along the line joining the source and the observer. On the other hand, if the source is moving in a direction perpendicular to the line joining the source and the observer, we have what is known as the *transverse Doppler effect*. There is then only time dilation, and we have for the Doppler shifted frequency

$$\nu_{TDS} = \nu_0\sqrt{1-\frac{u^2}{c^2}} \qquad \text{transverse Doppler effect} \tag{25}$$

31.4 THE LORENTZ TRANSFORMATION

Observer *A* is on the platform, and observer *B* is inside a train which is moving with velocity u in the $+x$ direction with respect to *A*. Let t and t' be the times measured by *A* and *B*, respectively, and we assume that the two clocks are synchronized such that at $t = t' = 0$, the origin *O* coincides with the origin *O′* (see Fig. 31.4).

A certain event occurs at point *P*; the event could be the switching on of a lightbulb. For *A* the event occurs at time t at a distance of x from the origin (see Fig. 31.5). In this time (according to *A*), *O′* has moved through a distance ut. Now,

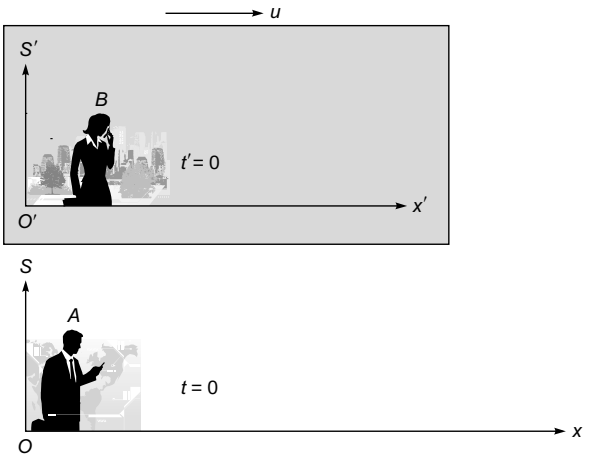


Fig. 31.4 Observer A is on the platform, and B is inside a train moving with velocity u in the + direction. When O coincides with O' the clocks are synchronized so that $t' = t = 0$.

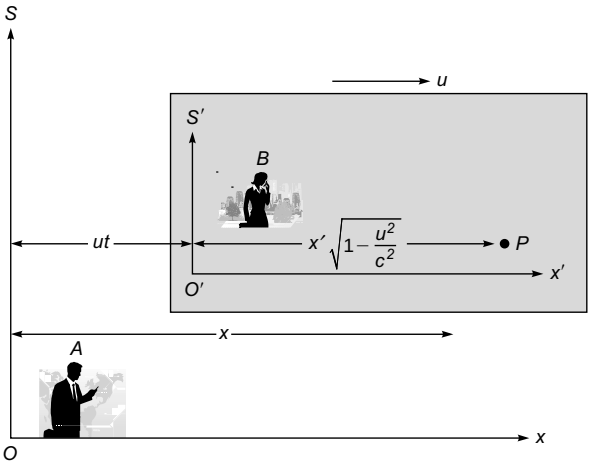


Fig. 31.5 An event occurs at point P; for A, the event occurs at time t at a distance x from O. For B, the event occurs at a distance x' from O' which A sees as a contracted distance, as shown in the figure.

according to B, the distance of point P from her origin is x', which A observes as the contracted distance $x' \sqrt{1-u^2/c^2}$. Thus according to A,

$$x = ut + x' \sqrt{1 - \frac{u^2}{c^2}} \tag{26}$$

or
$$x' = \gamma(x - ut) \tag{27}$$

where γ is the Lorentz factor [see Eq. (11)]. For B, observer A is moving with speed u in the $-x$ direction and the event occurs at time at a distance x' from O'. In this time (according to B), the origin O has moved through a distance $-ut'$. Since the distance x is measured by A, observer B will measure the contracted distance $x\sqrt{1-u^2/c^2}$. Thus (see Fig. 31.6)

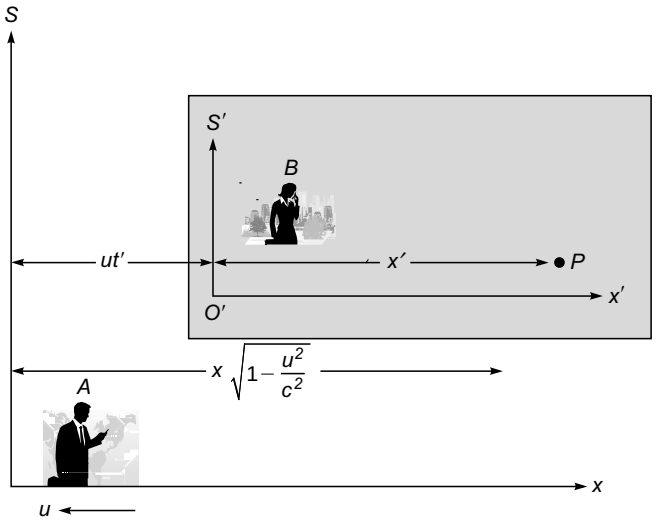


Fig. 31.6 For B, observer A is moving with speed u in the $-x$ direction. For B the event occurs at time t' at a distance x' from O. For A, the event occurs at a distance x from O which B sees as a contracted distance, as shown in the figure.

$$x' = x \sqrt{1 - \frac{u^2}{c^2}} - ut' \tag{28}$$

In the above equation, we substitute for x' from Eq. (27) and simplify to obtain

$$t' = \gamma \left(t - \frac{ux}{c^2} \right) \tag{29}$$

Equation (28) can also be written as

$$x = \gamma(x' + ut') \tag{30}$$

If we substitute for x from Eq. (30) in Eq. (27), we obtain

$$t = \gamma \left(t' + \frac{ux'}{c^2} \right) \tag{31}$$

The following equations

$$x' = \gamma(x - ut) \tag{32}$$

$$t' = \gamma \left(t - \frac{ux}{c^2} \right) \tag{33}$$

$$x = \gamma(x' + ut') \tag{34}$$

$$t = \gamma \left(t' + \frac{ux'}{c^2} \right) \tag{35}$$

along with the equations

$$y' = y \quad \text{and} \quad z' = z \tag{36}$$

describe what are known as Lorentz transformations. All the results derived in Chap. 30 can be derived using the above equations, as illustrated below.

Example 31.6 We consider two events. In the reference frame S the two events occur at (x_1, t_1) and (x_2, t_2), and in the reference frame S′ the two events occur at (x'_1, t'_1) and (x'_2, t'_2). Thus

$$x'_1 = \gamma(x_1 - ut_1) \quad \text{and} \quad x'_2 = \gamma(x_2 - ut_2)$$

and therefore

$$\Delta x' = x'_2 - x'_1 = \gamma(\Delta x - u\Delta t) \tag{37}$$

(a) If the two events take place at the same place in the S' frame, then $\Delta x' = 0$ (see Fig. 30.2) and therefore $\Delta x = u\,\Delta t$ (see Fig. 30.3).

Now, if we use Eq. (35) for the two events, we readily get

$$\Delta t = \gamma\left(\Delta t' + \frac{u\Delta x'}{c^2}\right) \tag{38}$$

If we now assume that the two events take place at the same place in the S' frame, then $\Delta x' = 0$ and we obtain

$$\Delta t' = \frac{1}{\gamma}\Delta t = \sqrt{1 - \frac{u^2}{c^2}}\,\Delta t \tag{39}$$

which tells us the following:

The time interval between two events occurring *at the same place* in a particular reference frame S' (referred to as the *proper time*)

$$= \sqrt{1 - \frac{u^2}{c^2}} \times \begin{array}{l}\text{time interval between two events in any}\\ \text{referenc frame } S \text{ moving with relative}\\ \text{speed } u\end{array} \tag{40}$$

which is the same as Eq. (5) of Chap. 30.

(b) We next consider two events which occur at the same time in the reference frame S' but at points separated by $2L_0$ (see Fig. 30.9). Thus in Eq. (38) we must substitute $\Delta t' = 0$ and $\Delta x' = 2L_0$ to obtain

$$\Delta t = \gamma L_0 \frac{2u}{c^2} \tag{41}$$

Thus whereas the two events are simultaneous in reference frame S', they are not simultaneous in reference frame S. We will get the same result if we use Eqs. (24) and (25) of Chap. 30.

Equations describing Lorentz transformations were derived by Lorentz much before Einstein. Lorentz (and later Poincaré) had shown that Maxwell's equations are invariant under Lorentz transformations; in App. F we show the invariance of the wave equation under Lorentz transformations. In his 1905 paper, Einstein (using his two postulates) derived Eqs. (32) and (33), but he made no mention of Lorentz' work. Many feel that Einstein was probably not aware of the work of Lorentz. Even the length contraction (discussed in Sec. 30.4) was first suggested by Fitzgerald in 1889 and shortly later by Lorentz independently; that is why length contraction is often called the Fitzgerald–Lorentz contraction or Lorentz–Fitzgerald contraction.

31.5 ADDITION OF VELOCITIES

Once again we consider the situation when observer A is on the platform and observer B is inside a train which is moving with velocity u in the $+x$ direction with respect to A. Let t and t' be the times measured by A and B, respectively, and we assume that the two clocks are synchronized such that at $t = t' = 0$, the origin O coincides with the origin O' (see Fig. 31.4). Inside the train a tennis ball (which is initially at the origin) is moving with velocity v in the $+x$ direction. The displacement of the tennis ball in reference frame S' is given by

$$x' = vt' \tag{42}$$

Substituting in Eq. (37), we get

$$x = \gamma(v + u)t' \tag{43}$$

We substitute Eq. (42) in Eq. (35), to get

$$t = \gamma\left(t' + \frac{ux'}{c^2}\right) = \gamma\left(1 + \frac{uv}{c^2}\right)t' \tag{44}$$

We divide Eq. (43) by Eq. (44) to obtain the following expression for the velocity of the tennis ball as seen by the observer in reference frame S:

$$V = \frac{x}{t} = \frac{v + u}{1 + uv/c^2} \tag{45}$$

This is the rule for *addition of velocities*. If $v = \frac{1}{3}c$ and $u = \frac{1}{3}c$, we get $V = \frac{3}{5}c$; thus $\frac{1}{3} + \frac{1}{3}$ is $\frac{3}{5}$. On the other hand, if $v = c$ and $u = \frac{1}{2}c$, we get $V = c$, showing that the speed of light remains constant.

REFERENCES AND SUGGESTED READINGS

1. F. Rohrlich, "An Elementary Derivation of $E = mc^2$," *Am. J. Phys.*, Vol. 58, No. 4, pp. 348–349, April 1990.
2. R. Baierlein, *Newton to Einstein: The Trail of Light*, Cambridge University Press, United kingdom, 1992.
3. R. Baierlein, "Does Nature Convert Mass into Energy?" *Am. J. Phys.*, Vol. 75, No. 4, pp. 320–325, April 2007.
4. A. Einstein, "Does the Inertia of a Body Depend on Its Energy Content?" *Annalen der Physik*, Vol. 18, pp. 639–641, 1905..
5. http://en.wikipedia.org/wiki/Large_Hadron_Collider.
6. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley Publishing Co., Reading, Mass., 1963.
7. B. M. Casper and R. J. Noer, *Revolutions in Physics*, W. W. Norton & Co., New York, 1972.

Appendix A

GAMMA FUNCTIONS AND INTEGRALS INVOLVING GAUSSIAN FUNCTIONS

We will first show that

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2+\beta x}\,dx = \sqrt{\frac{\pi}{\alpha}}\exp\left(\frac{\beta^2}{4\alpha}\right) \quad \text{Re}\,\alpha > 0 \qquad \textbf{(1)}$$

We consider the integral

$$I = \int_{-\infty}^{+\infty} e^{-x^2}\,dx \qquad \textbf{(2)}$$

Thus

$$I^2 = \int_{-\infty}^{+\infty} e^{-x^2}\,dx \int_{-\infty}^{+\infty} e^{-y^2}\,dy$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(x^2+y^2)}\,dx\,dy$$

Transforming to polar coordinates, we get

$$I^2 = \int_0^{+\infty} e^{-r^2} r\,dr \int_0^{\pi} d\theta$$

$$= \left[-\frac{1}{2}e^{-r^2}\right]_0^{\infty} \times 2\pi$$

$$= \pi$$

Thus

$$I = \int_{-\infty}^{+\infty} e^{-x^2}\,dx = \sqrt{\pi} \qquad \textbf{(3)}$$

Now

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2+\beta x}\,dx = \exp\left(\frac{\beta^2}{4\alpha}\right)\int_{-\infty}^{+\infty}\exp\left[-\alpha\left(x-\frac{\beta}{2\alpha}\right)^2\right]dx$$

$$= \exp\left(\frac{\beta^2}{4\alpha}\right)\int_{-\infty}^{+\infty} e^{-\alpha z^2}\,dz$$

where $z = x - \beta/2\alpha$. Using Eq. (3), we get

$$\int_{-\infty}^{+\infty} e^{-\alpha z^2}\,dz = \sqrt{\frac{\pi}{\alpha}} \qquad \textbf{(4)}$$

using which we obtain Eq. (1). We also get

$$\sqrt{\pi} = 2\int_0^{+\infty} e^{-x^2}\,dx$$

$$= \int_0^{+\infty} y^{-1/2} e^{-y}\,dy$$

Thus

$$\Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi} \qquad \textbf{(5)}$$

where $\Gamma(z)$ is defined through the equation

$$\boldsymbol{\Gamma(z)} = \int_0^{+\infty} x^{z-1}e^{-x}\,dx \qquad \text{Re}\,z > 0 \qquad \textbf{(6)}$$

For Re z >1, if we integrate by parts, we obtain

$$\Gamma(z) = (z-1)\Gamma(z-1) \qquad \textbf{(7)}$$

Since

$$\Gamma(1) = \int_0^{+\infty} e^{-x}\,dx = 1$$

we obtain

$$\Gamma(n+1) = n! \qquad n = 0,1,2,\ldots \qquad \textbf{(8)}$$

Further since $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, we obtain

$$\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{1}{2}\sqrt{\pi}$$
$$\Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\Gamma\left(\frac{3}{2}\right) = \frac{3}{2}\cdot\frac{1}{2}\sqrt{\pi} \qquad \textbf{(9)}$$
$$\Gamma\left(\frac{7}{2}\right) = \frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}\sqrt{\pi}$$

etc. Finally for $n = 0, 1, 2, \ldots$.

$$\int_{-\infty}^{+\infty} x^{2n} e^{-x^2}\,dx = \Gamma\left(n+\frac{1}{2}\right) = \frac{1\cdot 3\cdot 5\cdots(2n-1)\sqrt{\pi}}{2^n} \qquad \textbf{(10)}$$

and

$$\int_{-\infty}^{+\infty} x^{2n+1} e^{-x^2}\,dx = 0 \qquad \textbf{(11)}$$

Appendix B

EVALUATION OF THE INTEGRAL $\int_{-\infty}^{\infty} \frac{\sin gx}{x}\,dx$

The Laplace transform of a function $f(x)$ is defined by the equation

$$F(p) = L[f(x)] = \int_0^\infty e^{-px} f(x)\,dx \qquad \textbf{(1)}$$

Now,

$$\int_s^\infty F(p)\,dp = \int_s^\infty \int_0^\infty e^{-px} f(x)\,dx\,dp$$

$$= \int_0^\infty dx\, f(x) \left[\int_s^\infty e^{-px}\,dp \right]$$

Carrying out the integration over p, we get

$$\int_s^\infty F(p)\,dp = \int_0^\infty \frac{f(x)}{x} e^{-sx}\,dx \qquad \textbf{(2)}$$

In the limit of $s \to 0$, we obtain

$$\int_0^\infty F(p)\,dp = \int_0^\infty \frac{f(x)}{x}\,dx \qquad \textbf{(3)}$$

We next assume

$$f(x) = \sin gx \qquad g > 0 \qquad \textbf{(4)}$$

then

$$F(p) = \int_0^\infty \sin gx\, e^{-px}\,dx$$

$$= \frac{1}{2i} \int_0^\infty \left(e^{-(p-ig)x} - e^{-(p+ig)x} \right) dx$$

$$= \frac{1}{2i} \left(\frac{1}{p-ig} - \frac{1}{p+ig} \right)$$

$$= \frac{g}{p^2 + g^2}$$

Thus,

$$\int_0^\infty F(p)\,dp = g \int_0^\infty \frac{dp}{p^2 + g^2}$$

$$= \int_0^\infty \frac{dx}{1 + x^2} \qquad x = \frac{p}{g}$$

$$= \tan^{-1} x \Big|_0^\infty = \frac{\pi}{2} \qquad \text{for } g > 0 \qquad \textbf{(5)}$$

Obviously, for $g < 0$, the above integral is $-\pi/2$. Thus, using Eq. 3, we get

$$\int_0^\infty \frac{\sin gx}{x}\,dx = \begin{cases} \dfrac{\pi}{2} & \text{for } g > 0 \\ 0 & \text{for } g = 0 \\ -\dfrac{\pi}{2} & \text{for } g < 0 \end{cases} \qquad \textbf{(6)}$$

The integrand is an even function of x; thus

$$\int_{-\infty}^{+\infty} \frac{\sin gx}{\pi x}\,dx = \frac{2}{\pi} \int_0^\infty \frac{\sin gx}{x}\,dx = 1 \qquad g > 0 \qquad \textbf{(7)}$$

Appendix C THE REFLECTIVITY OF A FIBER BRAGG GRATING

In the core of a single-mode optical fiber, we assume a small z-dependent periodic variation of the refractive index; thus the complete refractive index variation is assumed to be given by

$$n(r, z) = n_0 + \Delta n \sin Kz \tag{1}$$

where $\Delta n << n_0$,

$$K = \frac{2\pi}{\Lambda} \tag{2}$$

and Λ represents the period of the z-dependent variation (see Fig. 15.12). The complete expression for the reflectivity is given by

$$R \approx \frac{\kappa^2 \sinh^2 \alpha L}{\kappa^2 \cosh^2 \alpha L - \Gamma^2/4} \tag{3}$$

where L is the length of the FBG,

$$\Gamma = 4\pi n_0 \left(\frac{1}{\lambda_0} - \frac{1}{\lambda_B} \right) \tag{4}$$

$$\alpha = \sqrt{\kappa^2 - \frac{\Gamma^2}{4}} \tag{5}$$

$$\lambda_B = 2\Lambda n_0 \tag{6}$$

is the Bragg wavelength and

$$\kappa = \frac{\pi \Delta n}{\lambda_0} \tag{7}$$

is known as the coupling coefficient. The maximum reflectivity occurs when $\lambda_0 = \lambda_B$ (thus $\Gamma = 0$), and one obtains the following expression for the peak reflectivity [see Eq. (47) of Chap. 15].

$$R = \tanh^2 \frac{\pi \Delta n L}{\lambda_B} \tag{8}$$

When $\Gamma > 2\kappa$, α becomes imaginary and Eq. (3) takes the form

$$R = \frac{\kappa^2 \sin^2 \gamma L}{-\kappa^2 \cos^2 \gamma L + \Gamma^2/4} \tag{9}$$

where

$$\gamma = \sqrt{\frac{\Gamma^2}{4} - \kappa^2} \tag{10}$$

Thus $R = 0$ when $\gamma L = n\pi$ ($n = 1, 2, 3, \ldots$). The wavelengths at which $R = 0$ [see Fig. 15.13(c)] are given by

$$\lambda_0 \approx \lambda_B \pm \frac{\lambda_B^2}{2\pi n_0 L} \left(\kappa^2 L^2 + n^2 \pi^2 \right)^{1/2} \tag{11}$$

We define the bandwidth of the reflection spectrum as half of the wavelength difference between the first minima on either side of the central peak, then it would be given by

$$\Delta\lambda_0 \approx \frac{\lambda_B^2}{2n_0 L} \sqrt{1 + \frac{\kappa^2 L^2}{\pi^2}} = \frac{\lambda_B^2}{2n_0 L} \sqrt{1 + \left(\frac{\Delta n L}{\lambda_B} \right)^2} \tag{12}$$

or

$$\frac{\Delta\lambda_0}{\lambda_B} \approx \frac{\lambda_B}{2n_0 L} \sqrt{1 + \left(\frac{\Delta n L}{\lambda_B} \right)^2} \tag{13}$$

Appendix D

DIFFRACTION OF A GAUSSIAN BEAM

If the amplitude and phase distribution on the plane $z = 0$ is given by $A(\xi, \eta)$, then the diffraction pattern is given by [see Eq. (23) of Chap. 20]

$$u(x, y, z) \approx -\frac{i}{\lambda z} \exp(ikz) \int\int A(\xi, \eta) \times \exp\left\{+\frac{ik}{2z}[(x-\xi)^2 + (y-\eta)^2]\right\} d\xi\, d\eta \quad \textbf{(1)}$$

We consider a Gaussian beam propagating along the z direction whose amplitude distribution on the plane $z = 0$ is given by

$$A(\xi, \eta) = a \exp\left(-\frac{\xi^2 + \eta^2}{w_0^2}\right) \quad \textbf{(2)}$$

implying that the phase front is plane at $z = 0$. Thus at a distance w_0 from the z axis, the amplitude falls by a factor $1/e$ (i.e., the intensity reduces by a factor $1/e^2$). This quantity w_0 is called the *spot size* of the beam. Substituting Eq. (2) in Eq. (1), we obtain

$$u(x, y, z) \approx -\frac{ia}{\lambda z} e^{ikz} \int_{-\infty}^{\infty} \exp\left[\frac{ik}{2z}(x-\xi)^2 - \frac{\xi^2}{w_0^2}\right] d\xi \times \int_{-\infty}^{\infty} \exp\left[\frac{ik}{2z}(y-\eta)^2 - \frac{\eta^2}{w_0^2}\right] d\eta$$

or

$$u(x, y, z) \approx -\frac{iae^{ikz}}{\lambda z} e^{\frac{ik}{2z}(x^2+y^2)} \int_{-\infty}^{+\infty} e^{-\alpha\xi^2 + \beta_1 \xi} d\xi \times \int_{-\infty}^{+\infty} e^{-\alpha\eta^2 + \beta_2 \eta} d\eta \quad \textbf{(3)}$$

where

$$\alpha = \frac{1}{w_0^2} - \frac{ik}{2z} = -\frac{ik}{2z}(1 + i\gamma) \quad \textbf{(4)}$$

$$\gamma = \frac{\lambda z}{\pi w_0^2} \quad \textbf{(5)}$$

$$\beta_1 = -\frac{ikx}{z} \qquad \beta_2 = -\frac{iky}{z}$$

If we now use the integral

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2 + \beta x} dx = \sqrt{\frac{\pi}{\alpha}} \exp\left(\frac{\beta^2}{4\alpha}\right) \quad \textbf{(6)}$$

we get

$$u(x, y, z) \approx \frac{a}{1 + i\gamma} \exp\left[-\frac{x^2 + y^2}{w^2(z)}\right] e^{i\Phi} \quad \textbf{(7)}$$

where

$$w(z) = w_0(1 + \gamma^2)^{1/2} = w_0\left(1 + \frac{\lambda^2 z^2}{\pi^2 w_0^4}\right)^{1/2} \quad \textbf{(8)}$$

and

$$\Phi = kz + \frac{k}{2z}(x^2 + y^2) - \frac{k(x^2 + y^2)}{2z(1 + \gamma^2)} = kz + \frac{k}{2R(z)}(x^2 + y^2) \quad \textbf{(9)}$$

where

$$R(z) \equiv z\left(1 + \frac{1}{\gamma^2}\right) = z\left(1 + \frac{\pi^2 w_0^4}{\lambda^2 z^2}\right) \quad \textbf{(10)}$$

Appendix E

TE AND TM MODES IN PLANAR WAVEGUIDES

In this appendix, we will derive the equations that are starting points for modal analysis. We start with Maxwell's equations, which for an isotropic, linear, nonconducting, and nonmagnetic medium take the form

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -i\mu_0 \frac{\partial \mathbf{H}}{\partial t} \quad (1)$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} = \varepsilon_0 n^2 \frac{\partial \mathbf{E}}{\partial t} \quad (2)$$

$$\nabla \cdot \mathbf{D} = 0 \quad (3)$$

$$\nabla \cdot \mathbf{B} = 0 \quad (4)$$

where we have used the constitutive relations

$$\mathbf{B} = \mu_0 \mathbf{H} \quad (5)$$

$$\mathbf{D} = \varepsilon \mathbf{E} = \varepsilon_0 n^2 \mathbf{E} \quad (6)$$

in which **E**, **D**, **B**, and **H** represent the electric field, electric displacement, magnetic induction, and magnetic intensity, respectively; μ_0 ($= 4\pi \times 10^{-7}$ N s^2 C^{-2}) represents the free space magnetic permeability, ε ($= \varepsilon_0 n^2$) represents the dielectric permittivity of the medium, n is the refractive index, and ε_0 ($= 8.854 \times 10^{-12}$ C^2 N^{-1} m^{-2}) is the permittivity of free space. If the refractive index varies only in the x direction, that is,

$$n^2 = n^2(x) \quad (7)$$

then we can always choose the z axis along the direction of propagation of the wave and we may, without any loss of generality, write the solutions of Eqs. (1) and (2) in the form

$$\mathcal{E} = \mathbf{E}(x) e^{i(\omega t - \beta z)} \quad (8)$$

$$\mathcal{H} = \mathbf{H}(x) e^{i(\omega t - \beta z)} \quad (9)$$

where β is known as the propagation constant. Equations (8) and (9) define the modes of the system. Thus modes represent transverse field distributions that suffer a phase change only as they propagate through the waveguide along z; the transverse field distributions described by **E**(x) and **H**(x) do not change as the field propagates through the waveguide. The quantity β represents the propagation constant of the mode. We rewrite the components of Eqs. (8) and (9):

$$\mathcal{E}_j = E_j(x) e^{i(\omega t - \beta z)} \qquad j = x, y, z \quad (10)$$

$$\mathcal{H}_j = H_j(x) e^{i(\omega t - \beta z)} \qquad j = x, y, z \quad (11)$$

Substituting the above expressions for the electric and magnetic field in Eqs. (1) and (2) and taking their x, y, and z components, we obtain

$$i\beta E_y = -i\omega\mu_0 H_x \quad (12)$$

$$\frac{dE_y}{dx} = -i\omega\mu_0 H_z \quad (13)$$

$$-i\beta H_x - \frac{dH_z}{dx} = i\omega\varepsilon_0 n^2(x) E_y \quad (14)$$

$$i\beta H_y = i\omega\varepsilon_0 n^2(x) E_x \quad (15)$$

$$\frac{dH_y}{dx} = i\omega\varepsilon_0 n^2(x) E_z \quad (16)$$

$$-i\beta E_x - \frac{dE_z}{dx} = -i\omega\mu_0 H_y \quad (17)$$

As can be seen, the first three equations involve only E_y, H_x, and H_z, and the last three equations involve only E_x, E_z, and H_y. Thus, for such a waveguide configuration, Maxwell's equations reduce to two independent sets of equations. The first set corresponds to nonvanishing values of E_y, H_x, and H_z with E_x, E_z, and H_y vanishing, giving rise to what are known as TE modes because the electric field has only a transverse component. The second set corresponds to nonvanishing values of E_x, E_z, and H_y with E_y, H_x, and H_z vanishing, giving rise to what are known as TM modes because the magnetic field now has only a transverse component. The propagation of waves in such planar waveguides may thus be described in terms of TE and TM modes.

TE Modes

We first consider TE modes: we substitute for H_x and H_z from Eqs. (12) and (13) in Eq. (14) to obtain

$$\frac{d^2 E_y}{dx^2} + [k_0^2 n^2(x) - \beta^2] E_y = 0 \quad (18)$$

where

$$k_0 = \omega\sqrt{\varepsilon_0\mu_0} = \frac{\omega}{c} \quad \textbf{(19)}$$

is the free space wave number and $c\,(=1/\sqrt{\varepsilon_0\mu_0})$ is the speed of light in free space. For a given refractive index profile $n^2(x)$, the solution of Eq. (18) (subject to appropriate boundary and continuity conditions) gives the field profile corresponding to the TE modes of the waveguide. Since $E_y(x)$ is a tangential component, it should be continuous at any discontinuity; further since dE_y/dx is proportional to $H_z(x)$ (which is a tangential component), it should also be continuous at any discontinuity. Once $E_y(x)$ is known, $H_x(x)$ and $H_z(x)$ can be determined from Eqs. (12) and (13), respectively. In Secs. 28.2 and 28.4 we solved Eq. (18) for a symmetric step index waveguide and for a parabolic index waveguide, respectively.

TM Modes

The TM modes are characterized by field components E_x, E_z, and H_y [see Eqs. (15) and (16)]. If we substitute for E_x and E_z from Eqs. (15) and (16) in Eq. (17), we get

$$n^2(x)\,\frac{d}{dx}\left[\frac{1}{n^2(x)}\frac{dH_y}{dx}\right] + [k_0^2\,n^2(x) - \beta^2]H_y(x) = 0 \quad \textbf{(20)}$$

The above equation is of a form that is somewhat different from the equation satisfied by E_y for TE modes [see Eq. (18)]. However, for the step index waveguide shown in Fig. 28.1, the refractive index is constant in each region, and we have

$$\frac{d^2H_y}{dx^2} + (k_0^2 n_i^2 - \beta^2)\,H_y(x) = 0 \quad \textbf{(21)}$$

At at each discontinuity

$$H_y \quad \text{and} \quad \frac{1}{n^2}\frac{dH_y}{dx} \quad \textbf{(22)}$$

should be continuous. This follows from the fact that since $H_y(x)$ is a tangential component, it should be continuous at any discontinuity; further since

$$\frac{1}{n^2}\frac{dH_y}{dx}$$

is proportional to $E_z(x)$ (which is a tangential component), it should also be continuous at any discontinuity.

Appendix F

SOLUTION FOR THE PARABOLIC INDEX WAVEGUIDE

In this appendix we show that for the solution of the equation

$$\frac{d^2\psi}{d\xi^2} + (\Lambda - \xi^2)\psi(\xi) = 0 \tag{1}$$

to be well behaved, we must have Λ = 1, 3, 5, 7, . . . ; i.e., Λ must be an odd integer. These are the eigenvalues of Eq. (1). We introduce the variable

$$\eta = \xi^2 \tag{2}$$

Thus

$$\frac{d\psi}{d\xi} = \frac{d\psi}{d\eta}\frac{d\eta}{d\xi} = \frac{d\psi}{d\eta}2\xi \tag{3}$$

and

$$\frac{d^2\psi}{d\xi^2} = 4\eta\frac{d^2\psi}{d\eta^2} + 2\frac{d\psi}{d\eta} \tag{4}$$

Substituting in Eq. (1), we obtain

$$\frac{d^2\psi}{d\eta^2} + \frac{1}{2\eta}\frac{d\psi}{d\eta} + \left(\frac{\Lambda}{4\eta} - \frac{1}{4}\right)\psi(\eta) = 0 \tag{5}$$

To determine the asymptotic form, we let $\eta \to \infty$ so that the above equation takes the form

$$\frac{d^2\psi}{d\eta^2} - \frac{1}{4}\psi(\eta) = 0$$

the solution of which is $e^{\pm\frac{1}{2}\eta}$. This suggests that we try out the following solution:

$$\psi(\eta) = y(\eta)e^{-\frac{1}{2}\eta} \tag{6}$$

Thus

$$\frac{d\psi}{d\eta} = \left(\frac{dy}{d\eta} - \frac{1}{2}y\right)e^{-\frac{1}{2}\eta} \tag{7}$$

and

$$\frac{d^2\psi}{d\eta^2} = \left(\frac{d^2y}{d\eta^2} - \frac{dy}{d\eta} + \frac{1}{4}y(\eta)\right)e^{-\frac{1}{2}\eta} \tag{8}$$

Substituting Eqs. (7) and (8) in Eq. (5), we get

$$\eta\frac{d^2y}{d\eta^2} + \left(\frac{1}{2} - \eta\right)\frac{dy}{d\eta} + \frac{\Lambda - 1}{4}y(\eta) = 0 \tag{9}$$

Now the confluent hypergeometric equation is given by (see, e.g., Refs. 1 and 2)

$$x\frac{d^2y}{dx^2} + (c - x)\frac{dy}{dx} - ay(x) = 0 \tag{10}$$

where a and c are constants. For $c \neq 0, \pm1, \pm2, \pm3, \pm4, \ldots$ the two independent solutions of the above equation are

$$y_1(x) = {}_1F_1(a, c, x) \tag{11}$$

and

$$y_2(x) = x^{1-c}\,{}_1F_1(a - c + 1, 2 - c, x) \tag{12}$$

where ${}_1F_1(a, c, x)$ is known as the confluent hypergeometric function and is defined by

$${}_1F_1(a, c, x) = 1 + \frac{a}{c}\frac{x}{1!} + \frac{a(a+1)}{c(c+1)}\frac{x^2}{2!} + \frac{a(a+1)(a+2)}{c(c+1)(c+2)}\frac{x^3}{3!} + \ldots \tag{13}$$

Obviously, for $a = c$ we will have

$${}_1F_1(a, a, x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots = e^x \tag{14}$$

Thus although the series given by Eqs. (13) and (14) are convergent for all values of x, they would blow up at infinity.

Indeed the asymptotic form of ${}_1F_1(a, c, x)$ is given by

$$ {}_1F_1(a, c, x) \underset{x\to\infty}{\rightarrow} \frac{\Gamma(c)}{\Gamma(a)} x^{a-c} e^{x} \qquad \textbf{(15)} $$

The confluent hypergeometric series ${}_1F_1(a, c, x)$ is very easy to remember, and its asymptotic form is easy to understand. Returning to Eq. (9), we find that $y(\eta)$ satisfies the confluent hypergeometric equation with

$$ a = \frac{1-\Lambda}{4} \quad \text{and} \quad c = \frac{1}{2} \qquad \textbf{(16)} $$

Thus the two independent solutions of Eq. (1) are

$$ \psi_1(\eta) = F_1\left(\frac{1-\Lambda}{4}, \frac{1}{2}, \eta\right) e^{-\frac{1}{2}\eta} \qquad \textbf{(17)} $$

and

$$ \psi_2(\eta) = \sqrt{\eta}\ {}_1F_1\left(\frac{3-\Lambda}{4}, \frac{3}{2}, \eta\right) e^{-\frac{1}{2}\eta} \qquad \textbf{(18)} $$

We must remember that $\eta = \xi^2$. Using the asymptotic form of the confluent hypergeometric function [Eq. (15)], we can readily see that if the series does not become a polynomial, then as $\eta\to\infty$, $\psi(\eta)$ will blow up as $e^{\frac{1}{2}\eta}$. To avoid this, the series must become a polynomial. Now $\psi_1(\eta)$ becomes a polynomial for $\Lambda = 1, 5, 9, 13, \ldots$ and $\psi_2(\eta)$ becomes a polynomial for $\Lambda = 3, 7, 11, 15$. Thus only when

$$ \Lambda = 1, 3, 5, 7, 9, \ldots \qquad \textbf{(19)} $$

we will have a well-behaved solution of Eq. (1) — these are the eigenvalues of Eq. (1). The corresponding wave functions are the Hermite-Gauss functions

$$ \psi_n(\xi) = N_n H_n(\xi) \exp\left(-\frac{1}{2}\xi^2\right) \qquad 5n = 0, 1, 2, 3, \ldots \qquad \textbf{(20)} $$

where

$$ N_n = \left[\frac{1}{2^n\, n!\sqrt{\pi}}\right]^{1/2} $$

is the normalization constant so that

$$ \int_{-0}^{+\infty} |\psi_n(\xi)|^2\, d\xi = 1. $$

Further,

$$ H_n(\xi) = (-1)^{n/2} \frac{n!}{(n/2)!}\, {}_1F_1\left(-\frac{n}{2}, \frac{1}{2}, \xi^2\right) \text{ for } n = 0, 2, 4, \ldots \qquad \textbf{(21)} $$

and

$$ H_n(\xi) = (-1)^{(n-1)/2} \frac{n!}{[(n-1)/2]!}\, 2\xi\, {}_1F_1\left(-\frac{n-1}{2}, \frac{3}{2}, \xi^2\right) \text{ for } n = 1, 3, 5, \ldots \qquad \textbf{(22)} $$

REFERENCES AND SUGGESTED READINGS

1. J. Irving and N. Mullineux, *Mathematics in Physics and Engineering*, Academic Press, New York, 1959.
2. A. K. Ghatak, I. C. Goyal, and S. J. Chua, *Mathematical Physics*, Macmillan India, New Delhi, 1985.

Appendix G

INVARIANCE OF THE WAVE EQUATION UNDER LORENTZ TRANSFORMATION

In this appendix we show that the scalar wave equation

$$\nabla^2\psi = \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} \tag{1}$$

is invariant under Lorentz transformation. In Cartesian coordinates

$$\nabla^2\psi = \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} \tag{2}$$

The equations describing the Lorentz transformations are given by (see Sec. 31.3)

$$x' = \gamma(x - ut) \tag{3}$$

$$t' = \gamma\left(t - \frac{ux}{c^2}\right) \tag{4}$$

$$y' = y \tag{5}$$

and

$$z' = z \tag{6}$$

where

$$\gamma = \frac{1}{\sqrt{1 - u^2/c^2}} \tag{7}$$

is the Lorentz factor. Since $y' = y$ and $z' = z$,

$$\frac{\partial^2\psi}{\partial y^2} = \frac{\partial^2\psi}{\partial y'^2} \quad \text{and} \quad \frac{\partial^2\psi}{\partial z^2} = \frac{\partial^2\psi}{\partial z'^2} \tag{8}$$

From Eqs. (3) and (4)

$$\frac{\partial x'}{\partial x} = \gamma \qquad \frac{\partial t'}{\partial x} = -\frac{\gamma u}{c^2} \tag{9}$$

Now

$$\frac{\partial\psi}{\partial x} = \frac{\partial\psi}{\partial x'}\frac{\partial x'}{\partial x} + \frac{\partial\psi}{\partial y'}\frac{\partial y'}{\partial x} + \frac{\partial\psi}{\partial z'}\frac{\partial z'}{\partial x} + \frac{\partial\psi}{\partial t'}\frac{\partial t'}{\partial x} \tag{10}$$

Using Eq. gives (9)

$$\frac{\partial\psi}{\partial x} = \gamma\frac{\partial\psi}{\partial x'} - \frac{\gamma u}{c^2}\frac{\partial\psi}{\partial t'} \tag{11}$$

and

$$\frac{\partial^2\psi}{\partial x^2} = \gamma\left(\frac{\partial^2\psi}{\partial x'^2}\frac{\partial x'}{\partial x} + \frac{\partial^2\psi}{\partial x'\partial t'}\frac{\partial t'}{\partial x}\right) - \frac{\gamma u}{c^2}\left(\frac{\partial^2\psi}{\partial x'\partial t'}\frac{\partial x'}{\partial x} + \frac{\partial^2\psi}{\partial t'^2}\frac{\partial t'}{\partial x}\right)$$

$$= \gamma^2\frac{\partial^2\psi}{\partial x'^2} - \frac{2\gamma^2 u}{c^2}\frac{\partial^2\psi}{\partial x'\partial t'} + \frac{\gamma^2 u^2}{c^4}\frac{\partial^2\psi}{\partial t'^2} \tag{12}$$

From Eqs. (3) and (4)

$$\frac{\partial x'}{\partial t} = -\gamma u \qquad \text{and} \qquad \frac{\partial t'}{\partial t} = \gamma \tag{13}$$

Thus

$$\frac{\partial\psi}{\partial t} = \frac{\partial\psi}{\partial x'}\frac{\partial x'}{\partial t} + \frac{\partial\psi}{\partial y'}\frac{\partial y'}{\partial t} + \frac{\partial\psi}{\partial z'}\frac{\partial z'}{\partial t} + \frac{\partial\psi}{\partial t'}\frac{\partial t'}{\partial t}$$

$$= -\gamma u\frac{\partial\psi}{\partial x'} + \gamma\frac{\partial\psi}{\partial t'} \tag{14}$$

$$\frac{\partial^2\psi}{\partial t^2} = -\gamma u\left(\frac{\partial^2\psi}{\partial x'^2}\frac{\partial x'}{\partial t} + \frac{\partial^2\psi}{\partial x'\partial t'}\frac{\partial t'}{\partial t}\right) + \gamma\left(\frac{\partial^2\psi}{\partial x'\partial t'}\frac{\partial x'}{\partial t} + \frac{\partial^2\psi}{\partial t'^2}\frac{\partial t'}{\partial t}\right)$$

$$= +\gamma^2 u^2\frac{\partial^2\psi}{\partial x'^2} - 2\gamma^2 u\frac{\partial^2\psi}{\partial x'\partial t'} + \gamma^2\frac{\partial^2\psi}{\partial t'^2}$$

$$\frac{\partial^2\psi}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} = \gamma^2\frac{\partial^2\psi}{\partial x'^2} - \frac{2\gamma^2 u}{c^2}\frac{\partial^2\psi}{\partial x'\partial t'} + \frac{\gamma^2 u^2}{c^4}\frac{\partial^2\psi}{\partial t'^2}$$

$$-\frac{\gamma^2 u^2}{c^2}\frac{\partial^2\psi}{\partial x'^2} + \frac{2\gamma^2 u}{c^2}\frac{\partial^2\psi}{\partial x'\partial t'} - \frac{\gamma^2}{c^2}\frac{\partial^2\psi}{\partial t'^2}$$

$$= \frac{\partial^2\psi}{\partial x'^2} - \frac{1}{c^2}\frac{\partial^2\psi}{\partial t'^2}$$

where we have used Eq. (7). Thus

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} = \frac{\partial^2\psi}{\partial x'^2} + \frac{\partial^2\psi}{\partial y'^2} + \frac{\partial^2\psi}{\partial z'^2} - \frac{1}{c^2}\frac{\partial^2\psi}{\partial t'^2} = 0 \tag{15}$$

which proves the invariance of the wave equation under Lorentz transformation.

SUBJECT INDEX

NAME INDEX

PHOTOGRAPHS

(a)

(b)

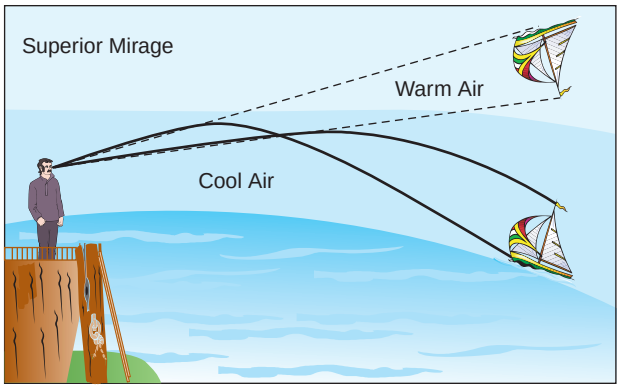
Chapter 2—Fig. 2.3 Photographs on the Moon. Because the Moon does not have any atmosphere, the sky and shadows are very dark. In (a) we can also see the Earth.

Chapter 3—Fig. 3.7 A paraboloidal satellite dish.

Chapter 3—Fig. 3.8 Fully steerable 45 m paraboloidal dishes of the Giant Metrewave Radio Telescope (GMRT) in Pune, India. The GMRT consists of 30 dishes of 45 m diameter with 14 antennas in the central array. Photograph courtesy: Professor Govind Swarup, GMRT, Pune.

Chapter 3—Fig. 3.16 A typical mirage as seen on a hot road on a warm day. The photograph was taken by Professor Piotr Pieranski of Poznan University of Technology in Poland; used with permission from Professor Pieranski.

Chapter 3—Fig. 3.17 This is actually *not* a reflection in the ocean, but the miraged (inverted) image of the Sun's lower edge. A few seconds later (notice the motion of the bird to the left of the Sun!), the reflection fuses with the erect image. The photographs were taken by Dr. George Kaplan of the U.S. Naval Observatory and are on the website http://mintaka.sdsu.edu/GF/explain/simulations/infmir/Kaplan_photos.html created by Dr. A Young; photographs used with permissions from Dr. Kaplan and Dr. A Young.

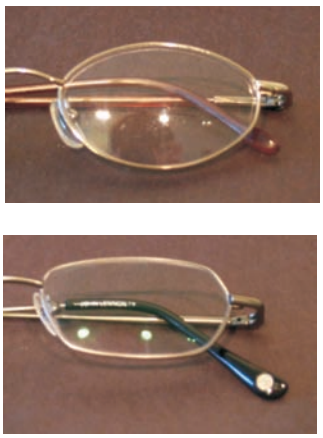


Chapter 3—Fig. 3.19 If we are looking at the ocean on a cold day, then the air near the surface of the water is cold and gets warmer as we go up. Thus, as we go up, the refractive index decreases continuously; and because of curved ray paths, one will observe an inverted image of the ship as shown in the figure above.

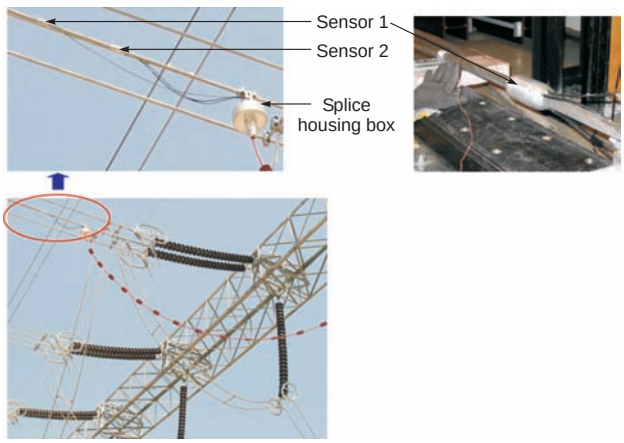
Chapter 3—Fig. 3.20 A house in the archipelago with a superior mirage. Figure adapted from http:/virtual.finland.fi/netcomm/news showarticle.asp?intNWSAID=25722. The photograph was taken by Dr. Pekka Parviainen in Turku, Finland; used with permission from Dr. Parviainen.

Chapter 3—Fig. 3.21 The noncircular shape of the Sun at sunset.

Chapter 7 Full moon over landscape at dusk. Notice the blue sky and the red glow of the setting sun. Both phenomena are due to Rayleigh scattering. [(c) Alamy Images RF]

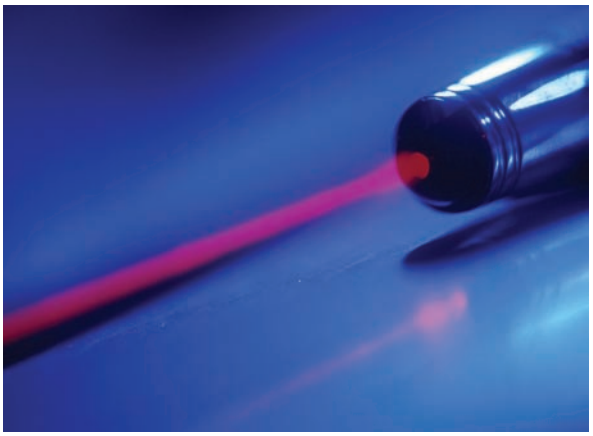


Chapter 15—Fig. 15.8 Comparison between an eyeglass lens without antireflective coating (top) and a lens with anti-reflective coating (bottam). Note the reflection of the photographer in the top lens and the tinted reflection in the bottam. The photograph was taken by Justin Lebar; used with permission from Mr. Lebar.

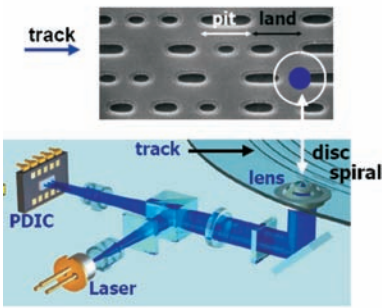


Chapter 15—Fig. 15.14 FBG-based temperature sensor system on 400 kV power conductor at Subhashgram substation (near Kolkata) of Powergrid Corporation of India. Photographs courtesy of Dr. Kamal Dasgupta and Dr. Tarun Gangopadhyay, CGCRI, Kolkata.

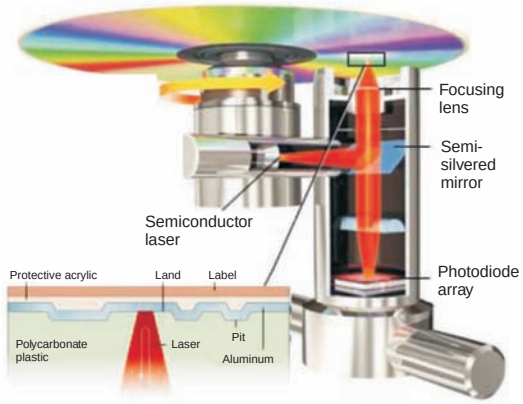
Chapter 15—Fig. 15.15 The substation of Powergrid Corporation of India (near Kolkata, India) where the FBG temperature sensors have been installed. In the photograph, the author is with Dr. Tarun Gangopadhyay and Dr. Kamal Dasgupta of CGCRI, Kolkata. Photo courtesy of Dr. Kamal Dasgupta and Dr. Tarun Gangopadhyay, CGCRI, Kolkata.



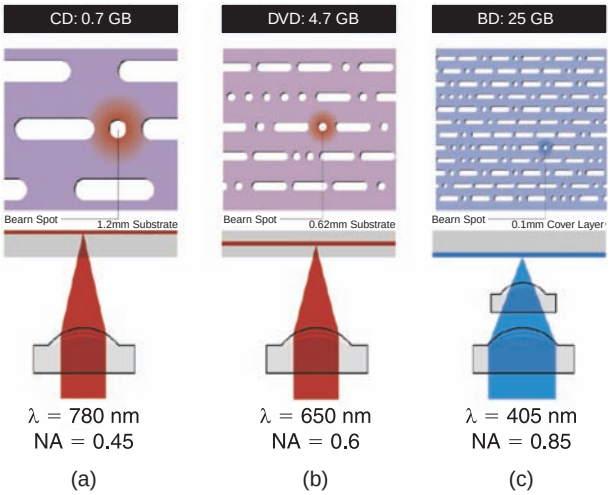
Chapter 18—Fig. 18.15 A laser beam. Notice the nondivergence of the beam.



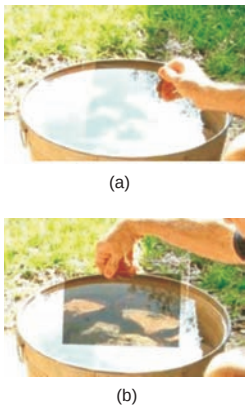
Chapter 18—Fig. 18.57 The pits and lands are essentially physical features (protrusions) on the disc surface, which are put there through injection molding. The heights of the pits from the surface are not arbitrary; rather they are fixed, being equal to $\lambda/4$ where λ is the wavelength of the laser used. Figure kindly provided by Dr. Rajeev Jindal and Mr. Giriraj Nyati of Moser Bear India in Greater Noida, India.



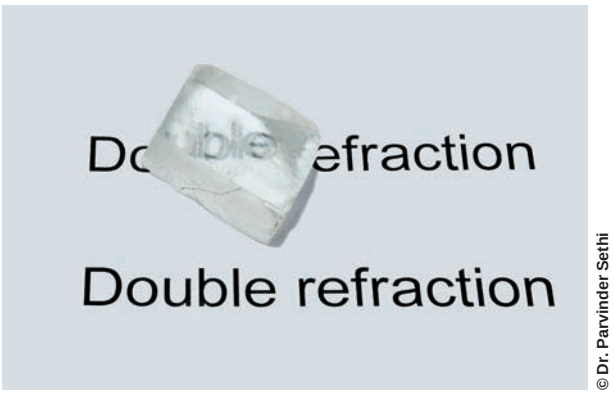
Chapter 18—Fig. 18.58 A CD-ROM substrate is made of optically clear polycarbonate over which the data marks are made through injection molding. The inner hole has a diameter of 15 mm while the overall diameter of the disc is 120 mm and the thickness is 1.2 mm. The top of the disc is covered with a very thin layer of silver or gold to form a reflective layer that reflects the laser beam so as to be read back. The reflected light is incident on a quadrant photodetector, which converts the light to suitable electrical pulses, which are subsequently processed to extract relevant data. Details are given in the text. Figure kindly provided by Dr. Rajeev Jindal and Mr. Giriraj Nyati of Moser Bear India in Greater Noida, India.



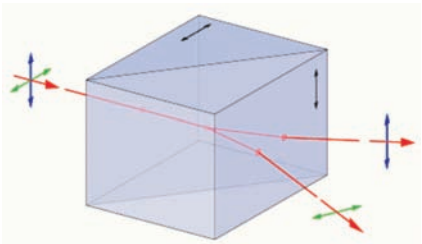
Chapter 18—Fig. 18.61 (a) Infrared diode laser (λ = 780 nm) with a simple objective lens with NA = 0.45. (b) Red laser (λ = 650 nm) with increase aperture objective with NA = 0.60. (c) Blue laser (λ = 405 nm) with further increase in NA = 0.85. Figure kindly provided by Dr. Rajeev Jindal and Mr. Giriraj Nyati of Moser Bear India in Greater Noida, India.



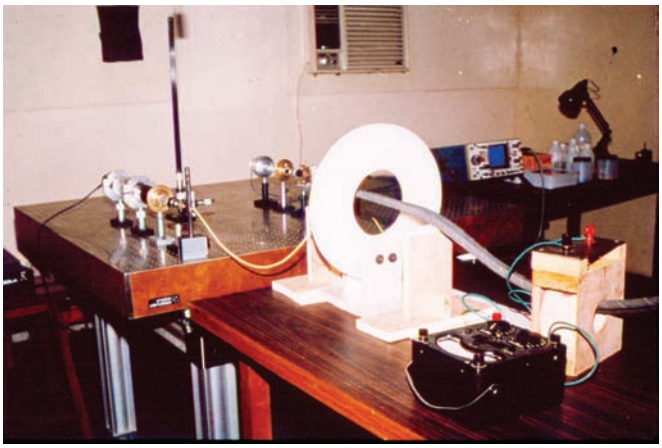
Chapter 22—Fig. 22.11 If the sunlight is incident on the water surface at an angle close to the polarizing angle, then the reflected light will be almost polarized. (a) If the Polaroid allows the (almost polarized) reflected beam to pass through, we see the glare from the water surface. (b) The glare can be blocked by using a vertical polarizer, and one can see inside the water. Figure adapted from the website www.polarization.com/water/water.html©J. Alcoz, 2001; used with permission of Dr. Alcoz.



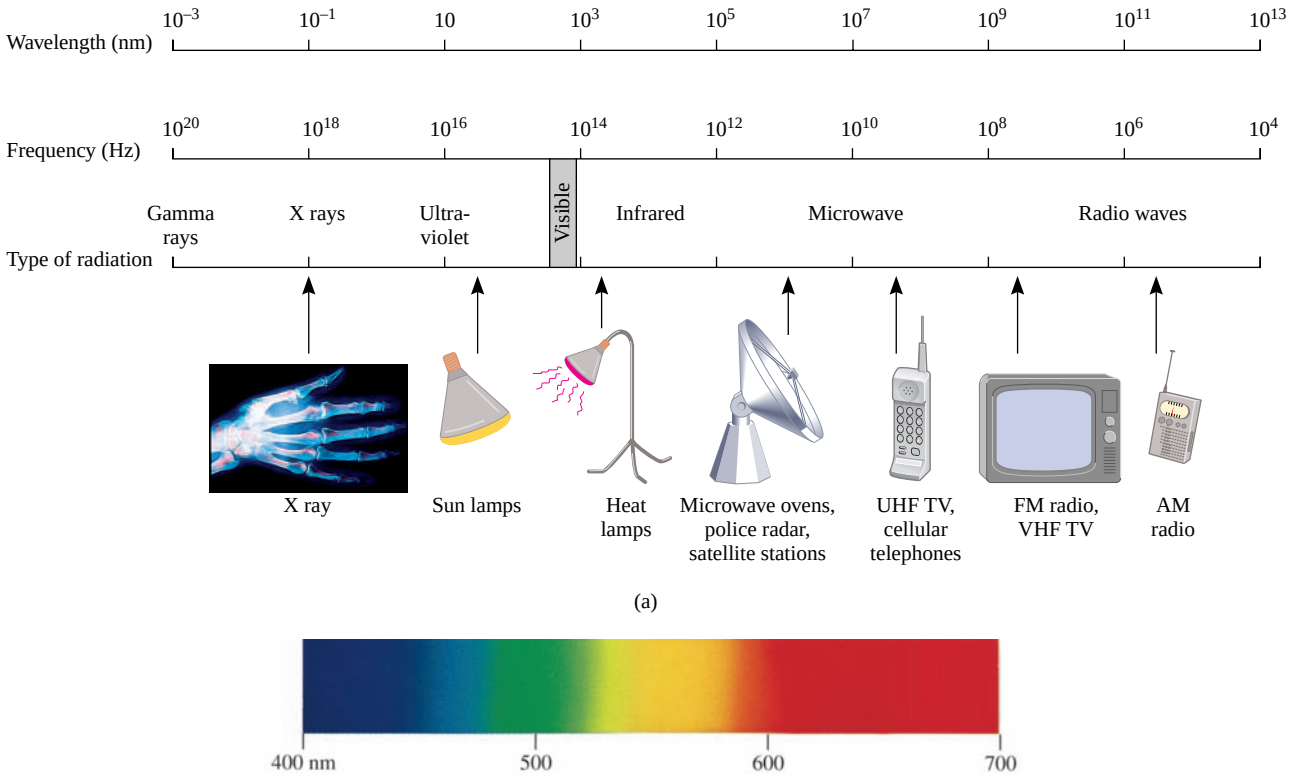
Chapter 22—Fig. 22.19 A calcite crystal showing double refraction.



Chapter 22—Fig. 22.30 Schematic of an actual Wollaston prism. The prism separates an unpolarized light beam into two linearly polarized beams. It typically consists of two properly oriented calcite prisms (so that the optic axes are perpendicular to each other), cemented together typically with Canada balsam. A commercially available Wollaston prism has divergence angles from 15° to about 45°.

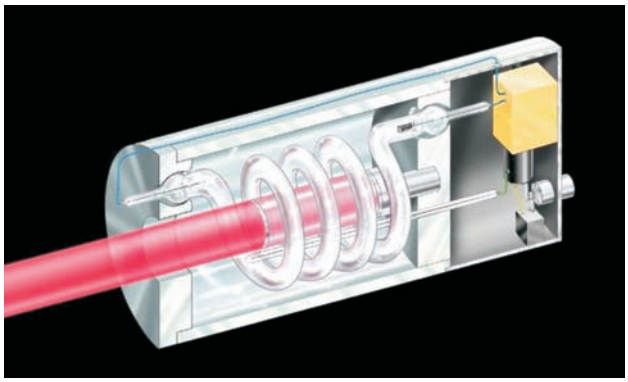


Chapter 22 As experimental setup to measure Faraday rotation in optical fibers because of large current passing through a conductor. Photograph courtesy: Professor Chandra Sekhar and Professor K Thyagarajan, IIT Delhi.

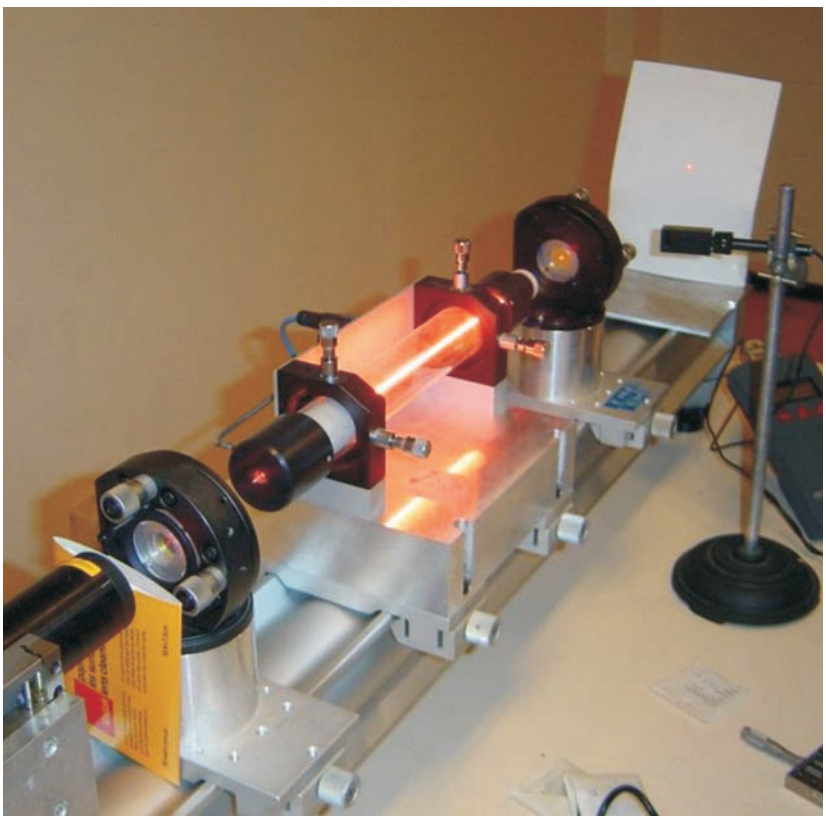


Chapter 23—Fig. 23.4 The electromagnetic spectrum; visible light occupies a very small portion of the spectrum.

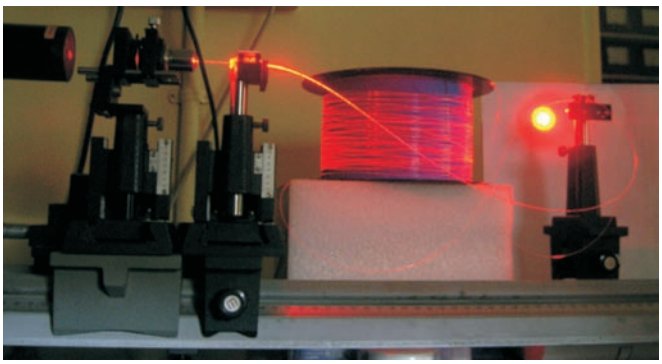
Chapter 26 A 2-kW fiber laser-mounted to a robotic system cutting mild steel. [Photograph (c) Getty Images RF]



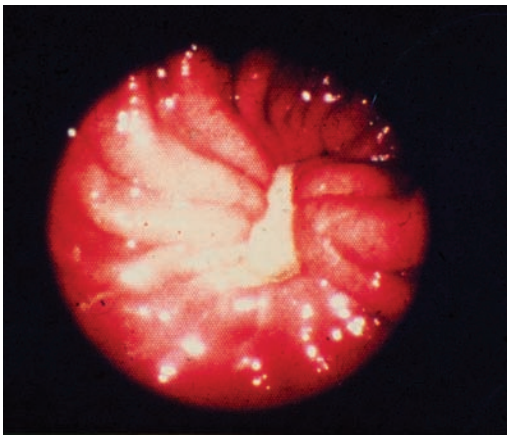
Chapter 26—Fig. 26.17 The first ruby laser.



Chapter 26—Fig. 26.21 A helium-neon laser demonstration at the Kastler-Brossel Laboratory at Univ. Paris 6. The glowing ray in the middle is an electric discharge producing light in much the same way as a neon light. It is the gain medium through which the laser passes, *not* the laser beam itself, which is visible there. The laser beam crosses the air and marks a red point on the screen to the right. Photograph by Dr. David Monniaux; used with kind permission of Dr. Monniaux.



Chapter 27—Fig. 27.3 A step index multimode fiber illuminated by He-Ne laser with bright output light spot. The light coming out of the optical fiber is primarily due to Rayleigh scattering. The fiber was produced at the fiber drawing facility at CGCRI, Kolkata; Photograph courtesy Dr. Shyamal Bhadra and Ms. Atasi Pal.



Chapter 27—Fig. 27.10(b) A stomach ulcer as seen through an endoscope. (Photograph courtesy United States Information Service, New Delhi)

(a)

(b)

Chapter 27—Fig. 27.2a (a) Guidance of light beam through optical fibers; the light scattered out of the fiber is due to Rayleigh scattering. (b) Optical fibers held by a hand.

Praise from Reviewers

This is an excellent text, based on what I have reviewed, and I would definitely recommend it for any physics or engineering program. It is also an excellent reference source for practicing engineers wanting to obtain a greater understanding of optics. It provides Maxwell's Equations for easy reference and presents excellent developments of the required equations for understanding and solving optical problems.

—Alfonso D'Alessio, New Jersey Institute of Technology

Seems quite readable with adequate detail but not overpowering math. The problems are challenging but not impossible given the excellent examples and careful developments in the text. They are also quite educational in many cases. The writing is clear and very readable…the examples used to illustrate points are practical and current.

—Dr. Thomas Plant, Oregon State University